ESS-MOTIFS: Discovering Rubric-Aligned Motifs for Cohort-Level Essay Assessment

Ngoc Thi Nguyen [1,2] **N. Duane Loh** [1,2]

[1]*Department of Physics, National University of Singapore, Singapore 117551, Singapore* [2]*Centre for Bio-imaging Sciences, National University of Singapore, Singapore 117557, Singapore*. Correspondence to: ngoc.nguyen@nus.edu.sg.

Assessing large volumes of open-ended essays one-by-one is cognitively demanding and time-consuming, often resulting in subjective, inconsistent, and difficult-to-audit grading. While automated essay scoring approaches provide quantitative consistency, black box models risk replacing human subjectivity with outputs that lack traceability. We introduce the ESS-MOTIFS (i.e., motifs in essays) framework that shifts essay grading from isolated, instance-level scoring to motif-based analysis for scalable, interpretable, and auditable cohort-level assessment. ESS-MOTIFS is motivated by the observation that, despite superficial variations, student essays on a given topic often share a limited set of conceptual and argumentative patterns. The framework quantitatively identify prevalent similarity and variance across essays through three stages: *rubric-augmented tokenization*, *motif discovery*, and *retrieval-augmented synthesis*. The process converts essays into semantically coherent tokens aligned with rubric criteria, clusters them to identify prevalent motifs, and generates motif- and essay-topic level summaries using retrieval-augmented large language models grounded in student essays and task context. By summarizing shared patterns and variations across the cohort, ESS-MOTIFS provides an objective scaffold that supports consistent, interpretable, auditable, and pedagogically meaningful assessment.

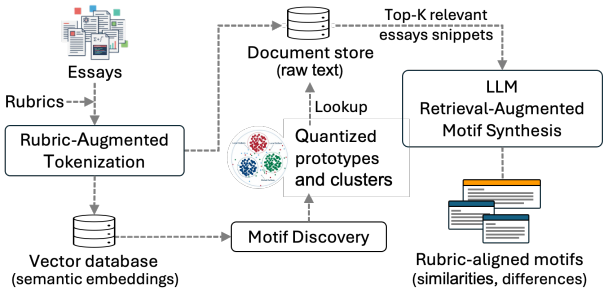

Fig. 1: ESS-MOTIFS framework transforms essays into semantic embeddings through Rubric-Augmented Tokenization. Motif Discovery clusters these embeddings and identifies semantic motifs, followed by Retrieval-Augmented Motif Synthesis, where an LLM synthesizes representative essay excerpts into cohort-level insights.

1. Introduction

Human judgment is essential for nuanced essay evaluations, however, grading large volumes in limited time introduces significant cognitive load, making evaluation vulnerable to inconsistency and bias [1] . While rubrics provide guidance for consistent assessment, grading essays in isolation hinders recognition of overlapping patterns and recurring forms of reasoning across students. As a result, essays of similar quality are prone to different scores depending on when and how they are evaluated. These challenges call for comparative, cohort-level evaluation to support more consistent scoring [2].

Automated essay scoring (AES) systems provide quantitative consistency and reduce instructor fatigue [3] but treat essays grading on individual basis, failing to capture shared patterns across essays. These systems typically utilize statistical models or machine learning to predict holistic score or coarse categories from surface linguistic and structural features aligned with rubrics [4, 5]. However, they optimizes for scoring efficiency rather than modeling student conceptual understanding or domain reasoning, resulting in scores that lack pedagogical justification. More recent approaches using Large Language Models (LLMs) enable semantically informed evaluations and feedback [6, 7] but using them as end-to-end graders risks replacing human subjectivity with opaque algorithm judgment, leaving evaluative logic hidden in 'black box' numerical models or flattened by 'grey box' evaluation that lack traceability.

Student essays on a given topic typically draw from a limited repertoire of conceptual understandings and argumentative structures, even though they differ in wording and style. These recurring structures can be viewed as patterns of reasoning that appear with varying levels of correctness, completeness, and coherence across students. From this perspective, writing assessment involves not only evaluating essays against rubric criteria, but also recognizing and comparing these shared patterns across a cohort. Making such patterns explicit can anchor evaluation in common reference structures, thereby supporting more interpretable, consistent, and fair assessment.

This work introduces ESS-MOTIFS framework that shifts essay grading from isolated, instance-level scoring to motif-based analysis across the cohort. The framework represents essays as collections of tokens, where each token is a vector embeddings corresponding to a semantically text unit associated with a particular rubric dimension. ESS-MOTIFS groups tokens into thematic clusters to identify prevalent, rubric-aligned motifs, then leverages a Retrieval Augmented Generation (RAG) approach to provide relevant context to an LLM. The LLM summarizes these motifs, producing syntheses that are both computationally scalable and pedagogically auditable, augmenting human grading with cohort-level insights.

2. ESS-MOTIFS Framework

Figure 1 outlines ESS-MOTIFS for identifying and summarizing rubric-aligned motifs across essays.

Rubric-Augmented Tokenization. ESS-MOTIFS segments essays into hierarchical *tokens*: paragraph and essay levels. Each token is associated with a rubric and encoded as vector embeddings capturing semantic meaning. This tokenization stage lays the foundation for consistent evaluation and scalable analysis while preserving rubric-relevant meaning for interpretable and auditable assessment.

Motif Discovery. ESS-MOTIFS organizes the semantic embedding space into clusters based on similarity within each rubric dimension. By vector quantizing these clusters, ESS-MOTIFS identifies prevalent patterns, or *motifs*, that represent typical student approaches. To ensure a nuanced, auditable assessment, the framework extracts both prototypes (centroids) to define central behaviors and boundary cases to delineate the outer limits of each motif.

Retrieval-Augmented Motif Synthesis. ESS-MOTIFS utilizes a retrieval-augmented generation approach to stimulate LLM generating human-readable summaries for rubric-aligned motifs. By providing the LLM with rubric criteria and representative paragraph-level tokens retrieved from all cluster members, ESS-MOTIFS reduces the LLM's hallucination and ensure that summaries are faithful to the underlying evidence.

3. Results and Discussions

Cluster-Level Insights. Applying ESS-MOTIFS framework to an essay corpus ($N = 535$) reveals their dominant patterns (i.e., prototype cases) and deviations (i.e., boundary cases) within each of rubric dimensions: content, reasoning, structure, and clarity. LLM-generated summaries allow instructors to quickly identify which essays shared common patterns, how these patterns diverge, and the sources of arguments (see Box 1), thereby providing traceable, rubric-aligned evidences.

Box 1: Cluster C1 (Q1) - Reasoning

Cluster members (11): S0002, S0039, S0056, S0093, S0095, S0042, S0047, S0010, S0104, S0105, S0009
Anchors: Prototype: S0002 | Boundary: S0009

Reasoning: The students in this cluster emphasize the importance of provenance in ensuring the credibility and trustworthiness of data used in AI projects. They believe that establishing a clear chain of data collection, processing, and transformation is crucial for mitigating risks and defining operational limitations. Student S0002 prioritizes documenting every step in a reproducible pipeline, while Student S0009 focuses on avoiding provenance drift and ensuring consistency in high-dimensional codes ...

Question-Level Synthesis. For each question and rubric dimension, ESS-MOTIFS framework aggregates syntheses across all clusters to characterize the dominant reasoning patterns as well as minority perspectives (see Box 2), providing interpretable insights while remaining auditable and traceable.

Box 2: Q1 - Reasoning

1. The majority of students employ causal reasoning to explain the impact of technology on various aspects of society, with some clusters emphasizing individual agency, structure, or complex interplay of factors.
2. A few clusters exhibit descriptive reasoning (Clusters 7, 17, 25, 29, and 35), inductive reasoning (Clusters 11, 15, 26, 31, and 37), or a comparison and contrast approach (Clusters 34 and 35) ...

Interpretability and Scalability. The results demonstrate that rubric-aligned clustering combined with nested retrieval and LLM synthesis can make large-scale essay assessment interpretable and scalable. By organizing essays into thematic clusters aligned with rubric dimensions, ESS-MOTIFS preserves dominant and minority reasoning patterns that are often smoothed out by end-to-end LLM scoring.

Implications for Essay Assessment in the LLM Era. As generative AI tools become prevalent, student writing increasingly resemble AI-generated text [8]. Even when essays are different in wording and style, they may share highly similar underlying reasoning structures. By operating in a semantic embedding space, the framework can reveal this latent convergence, allowing instructors to distinguish genuinely diverse lines of thoughts from merely paraphrased variants.

Room for Improvement. ESS-MOTIFS assumes that essays exhibit recurring patterns that can be meaningfully grouped by rubrics. The interpretability of the discovered motifs and their summaries therefore depends on rubric granularity and clarity. The framework provides motifs, prototypes, and boundaries but interpreting them still require instructor judgment.

4. Conclusion

We introduce ESS-MOTIFS, a rubric-aligned motif analysis framework that that shifts essay assessment from instance-level evaluation to motif-based analysis for scalable, consistent, and auditable cohort-level assessment. By identifying and synthesizing prevalent, rubric-aligned reasoning motifs using retrieval-augmented large language models, ESS-MOTIFS produces interpretable cluster- and question-level summaries grounded in student writing. This approach offers an objective scaffold that complements human judgment, supporting consistent, auditable, and pedagogically meaningful essay assessment.

Acknowledgments

The authors gratefully acknowledge funding support from the Department of Physics at the National University of Singapore (NUS), and the Institute for Digital Molecular Analytics and Science (IDMxS). We also thank HPC support from the NUS Centre for Bio-Imaging Sciences.

References

[1] Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2(1):319–330, 2003.

[2] Jeffrey T Steedle and Steve Ferrara. Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, 29(3):211–223, 2016.

[3] Jingbo Sun, Tianbao Song, Weiming Peng, and Jihua Song. A survey of automated essay scoring: Challenges, advances, and future. *Neurocomputing*, page 130916, 2025.

[4] Fei Dong, Yue Zhang, and Jie Yang. 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, 2017.

[5] Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 3416–3425, 2022.

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

[7] Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. Unleashing large language models' proficiency in zero-shot essay scoring. *arXiv preprint arXiv:2404.04941*, 2024.

[8] Petteri Nurmi, Musfira Khan, Zahra Safaei, Ngoc Thi Nguyen, Fatemeh Sarhaddi, Mika Seppo Tompuri, Henrik Nygren, Päivi Anneli Kinnunen, and Agustin Zuniga. Ai see what you did there: Assessing the rise of llm-generated responses in online learning. In *Technical Symposium on Computer Science Education*, 2026.