Mixture of Ordered Scoring Experts for Cross-prompt Essay Trait Scoring

Po-Kai Chen[3], Bo-Wei Tsai[3], Kuan-Wei Shao[1], Chien-Yao Wang[2], Jia-Ching Wang[3], and Yi-Ting Huang[1*]

[1]National Taiwan University of Science and Technology, [2]Institute of Information Science, Academia Sinica, [3]National Central University

ACL 2025 VIENNA JULY 27 - AUGUST 1

Abstract

We propose the Mixture of Ordered Scoring Experts (**MOOSE**), a framework for essay trait scoring. It imitates the scoring process of professional human raters by integrating three specialized experts to evaluate: (1) the overall quality of an essay, (2) its relative quality compared to other essays, and (3) its relevance to the given prompt. Furthermore, by reformulating essay trait scoring as a scoring cue retrieval problem and using the essay as the query, MOOSE achieves state-of-the-art performance in cross-prompt essay trait scoring on the ASAP++ dataset. It offers stable and trait-consistent results, surpassing previous models including LLM-based methods.

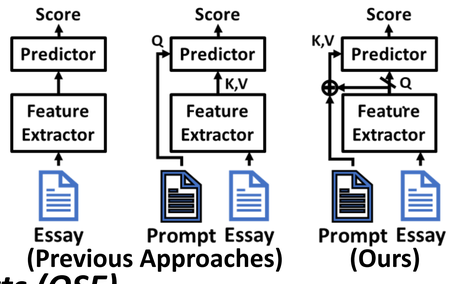
Methodology

Feature Extraction

- *Capturing hierarchical features of essay and prompt by Multi-Chunk BERT [1] and Trait Attention [2].*
- *Extract 86 linguistic features (POS, readability, syntax, …).*

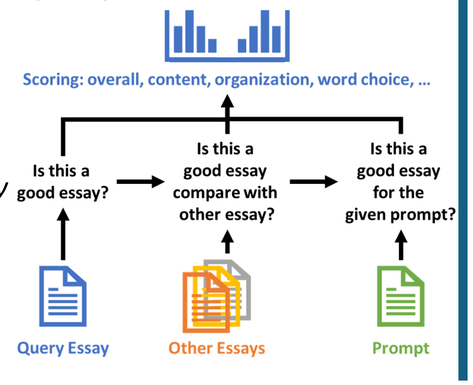
Essay as Query

- *Prevents overfitting to seen prompts.*
- *Enables generalized scoring cue retrieval.*

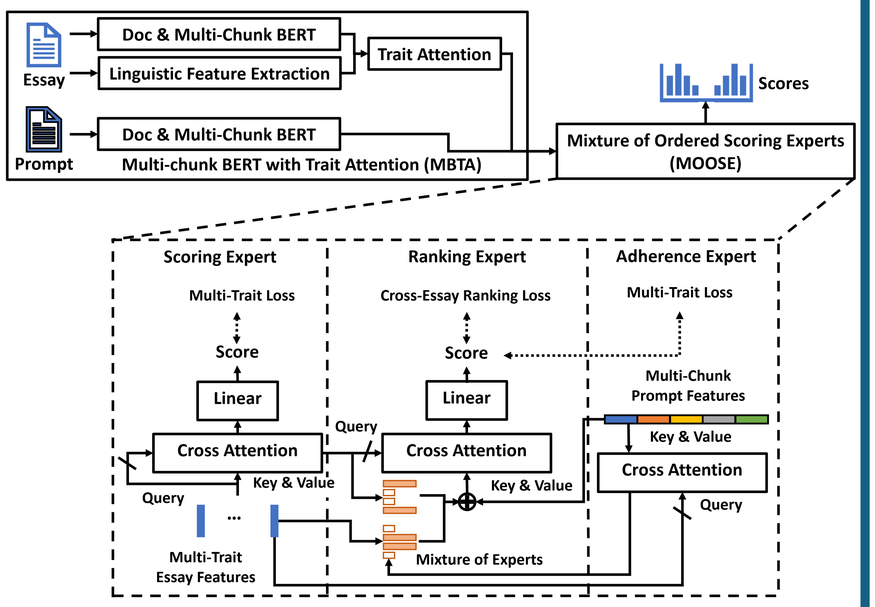


Ordered Scoring Experts (OSE)

- **Scoring Expert**: *Learn essay inherent scoring cues.*
- **Ranking Expert**: *Compare relative quality across different essays.*
- **Adherence Expert**: *Estimate the degree of prompt adherence.*



Mixture of Ordered Scoring Experts (MOOSE)



Performance

Cross-Prompt QWK (Avg. over 8 prompts)

Model	QWK	STD
RDCTS [3]	0.570	0.085
ProTACT [2]	0.592	0.067
EPCTS [4] (LLM-based)	0.632	0.038
OSE (Ours)	0.638	0.037
MOOSE (Ours)	**0.642**	**0.036**

Multi-Trait QWK (Avg. over 9 traits)

Model	QWK	STD
RDCTS [3]	0.568	0.065
ProTACT [2]	0.586	0.058
EPCTS [4] (LLM-based)	0.623	0.035
OSE (Ours)	0.634	0.023
MOOSE (Ours)	**0.641**	**0.018**

Improvements

- *Outperforms all SoTAs on cross-prompt essay trait scoring.*
- *Achieves exceptionally stable performance across different prompts and traits.*
- *Makes the prediction of the model be interpretable.*

Cross-Prompt QWK of Different Query Type

Model	QWK	STD
Prompt as query	0.591	0.091
Essay as query	**0.624**	**0.057**

Cross-Prompt QWK of Different Learning Goal

Model	QWK	STD
Learning to scoring	0.589	0.058
Learning to retrieve scoring cues	**0.596**	**0.056**

Cross-Prompt QWK of Different Scoring Experts

Model	QWK	STD
Scoring experts	0.597	0.059
Ranking experts	0.607	**0.054**
Ordered experts	**0.624**	0.058

Insights

- *Using **essay as query** strongly improves the performance via estimating distribution of essay over prompt & essay.*
- *Reformulating learning goal to **scoring cue retrieval** makes the model more robust on the unseen prompt.*
- *By imitating scoring process of human raters, **ordered experts** get outstanding performance on essay scoring.*

Reference

[1] Yongjie Wang et al. "On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation." NAACL, 2022.

[2] Heejin Do et al. "Prompt- and trait relation-aware cross-prompt essay trait scoring." ACL Findings, 2023.

[3] Jingbo Sun et al. "Enhanced cross-prompt trait scoring via syntactic feature fusion and contrastive learning." The Journal of Supercomputing, 2024.

[4] Jiangsong Xu et al. "EPCTS: Enhanced prompt-aware cross-prompt essay trait scoring." Neurocomputing, 2025.

Code & Demo

The code and the demo of the paper are publicly available at https://antslabtw.github.io/MOOSE