ACL 2025 VIENNA
JULY 27 - AUGUST 1

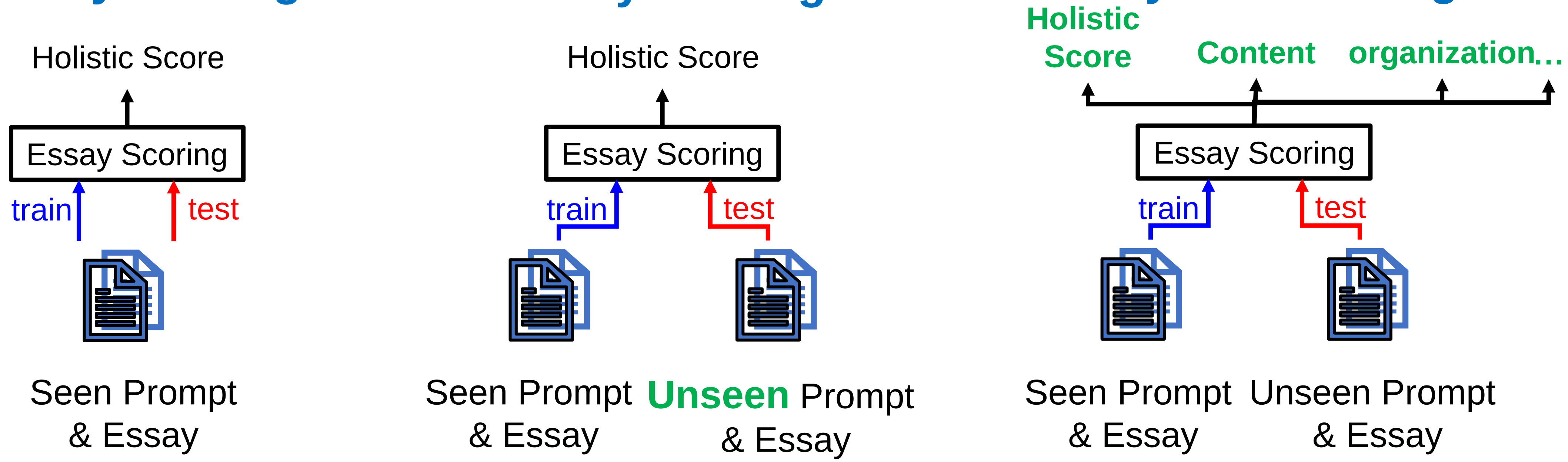
Mixture of Ordered Scoring Experts for Cross-prompt Essay Trait Scoring

Po-Kai Chen[+], Bo-Wei Tsai[+], Kuan-Wei Shao[★],
Chien-Yao Wang[♪], Jia-Ching Wang[+], and Yi-Ting Huang[★]

National Taiwan University of Science and Technology[★],
Academia Sinica[♪], National Central University[+]

Task definition

Essay scoring	Cross-prompt essay scoring	Cross-prompt essay trait scoring



(Taghipour and Ng, 2016;Dong and Zhang, 2016; Yang et al., 2020; Wanget al., 2022)

(Jin et al., 2018; Li et al., 2020; Ridley et al., 2020)

(Ridley et al., 2021; Chenand Li, 2023; Do et al., 2023; Xu et al., 2025)

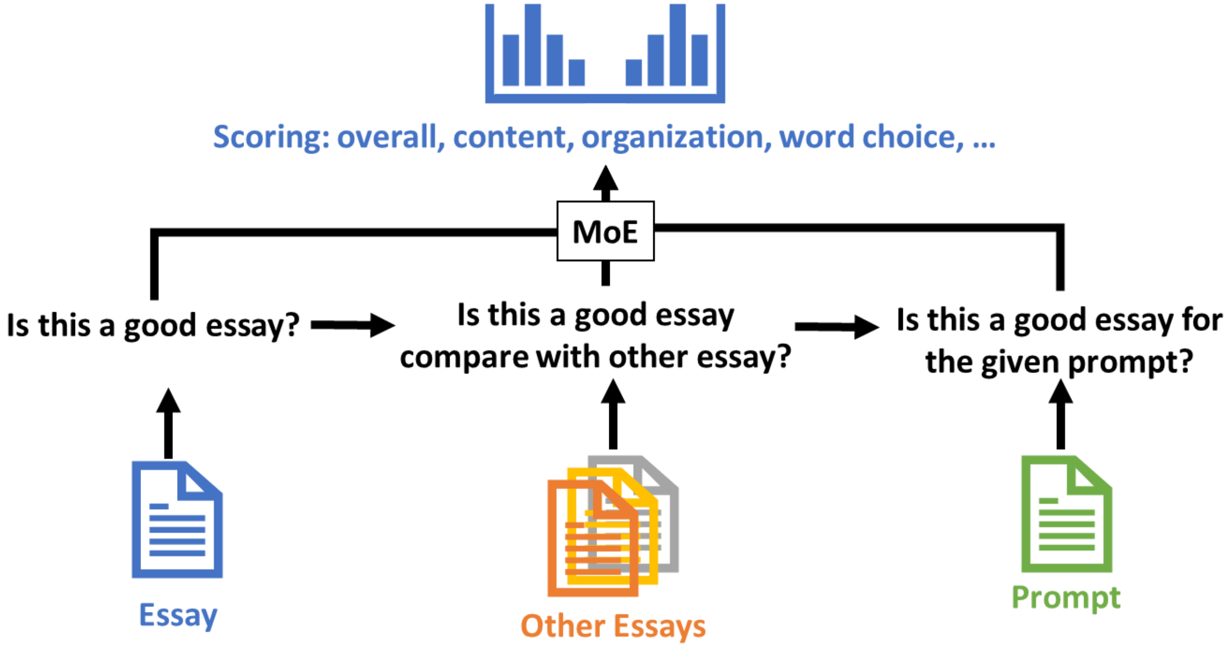
- Our work focuses on **cross-prompt essay trait scoring**.

Motivation

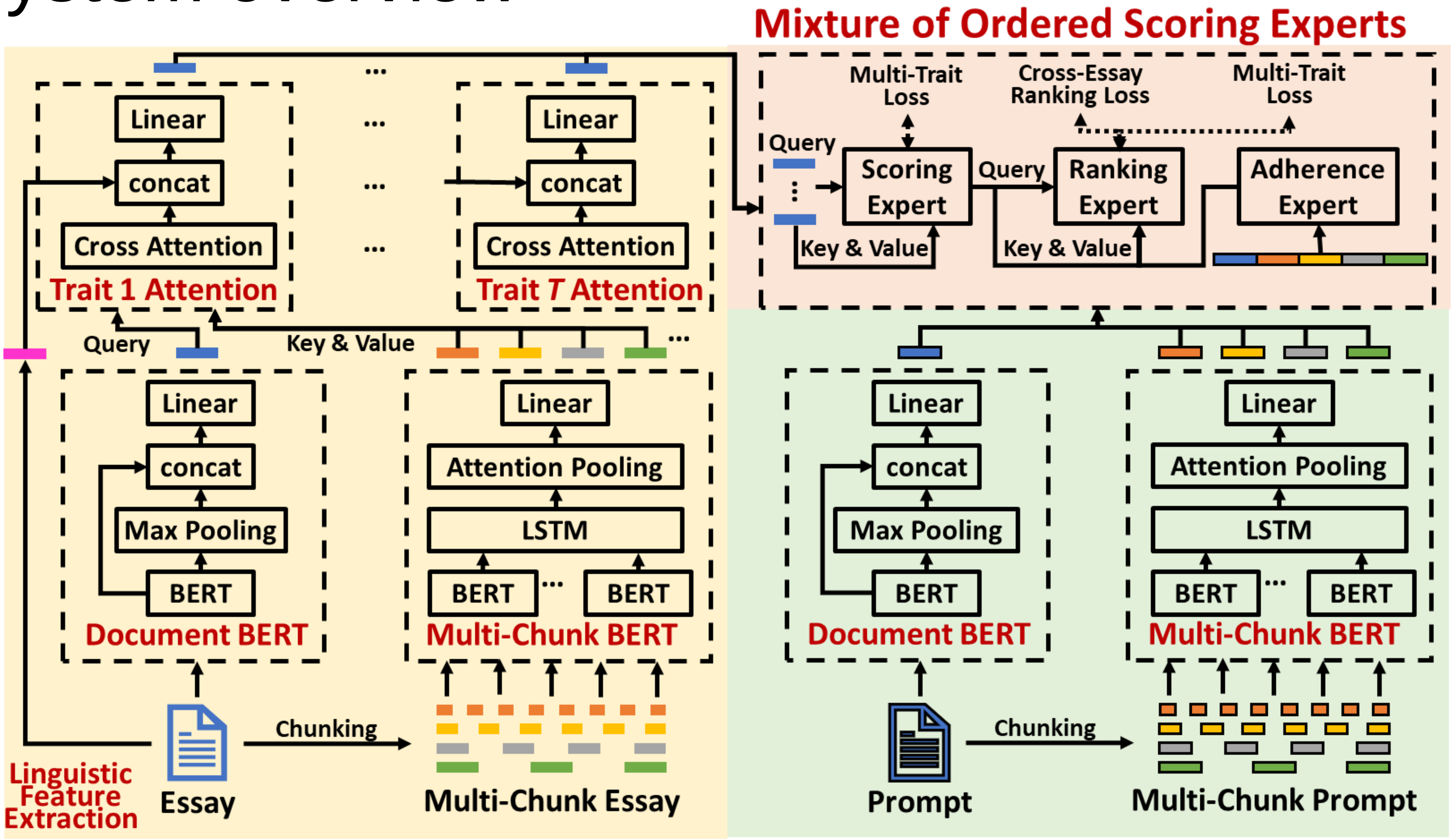
- Previous work like PAES (Ridleyet al., 2020)
  - Consider only essay as input.
  - Focus on the essay quality and ignoring prompt adherence.
- SOTA: ProTACT (Do et al., 2023)
  - Use LDA to extract essay-prompt correlation.
  - Rely solely on syntactic features for essay representation.
- They overlook content-level features in both prompts and essays, such as semantic and linguistic information.
- They develops **ONE single model** to evaluate multiple traits, failing to capture different perspectives specific to each trait.

Research purpose

- In this work, we propose MOOSE (Mixture of Ordered Scoring Experts) framework for cross-prompt essay trait scoring.
  - **Ordered Scorer Experts (OSE):** designs three experts to imitate the reasoning process of a human rater.
  - **Mixture of Experts (MoE):** dynamically selects different scoring cues that are specific to each trait.



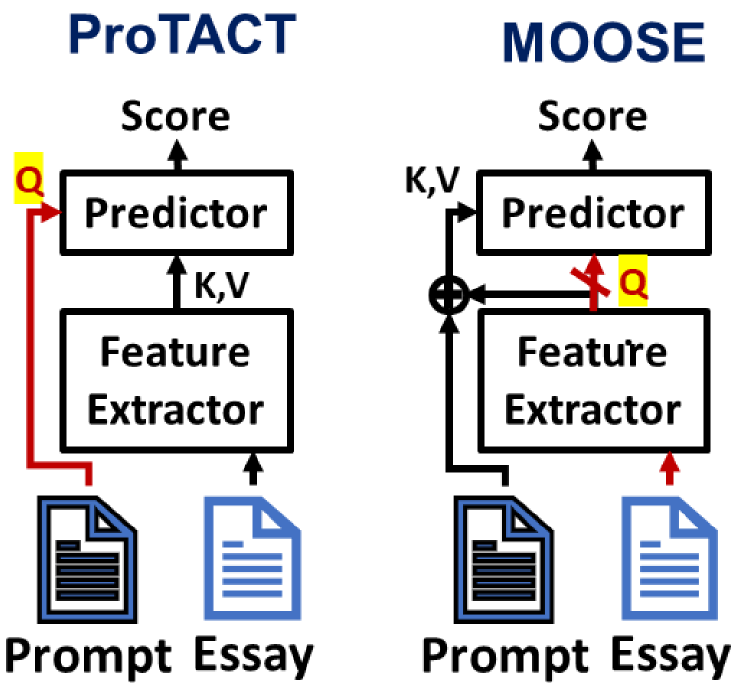
System overview



Novelty 1: Essay as Query

- ProTACT (SOTA):
  - Treat **the** **prompt** **as the query**
  - Evaluate essays from the **prompt's perspective** to determine whether a given essay is likely to receive a high score under the given prompt.

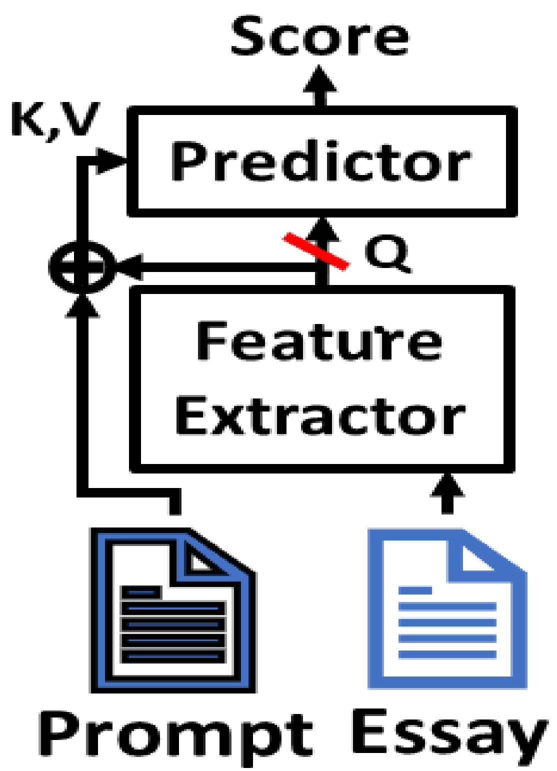
- MOOSE:
  - Uses **the** **essay** **as a query** to learn essay representation.
  - To estimate the distribution of the query (essay) over the values (prompt and essay).



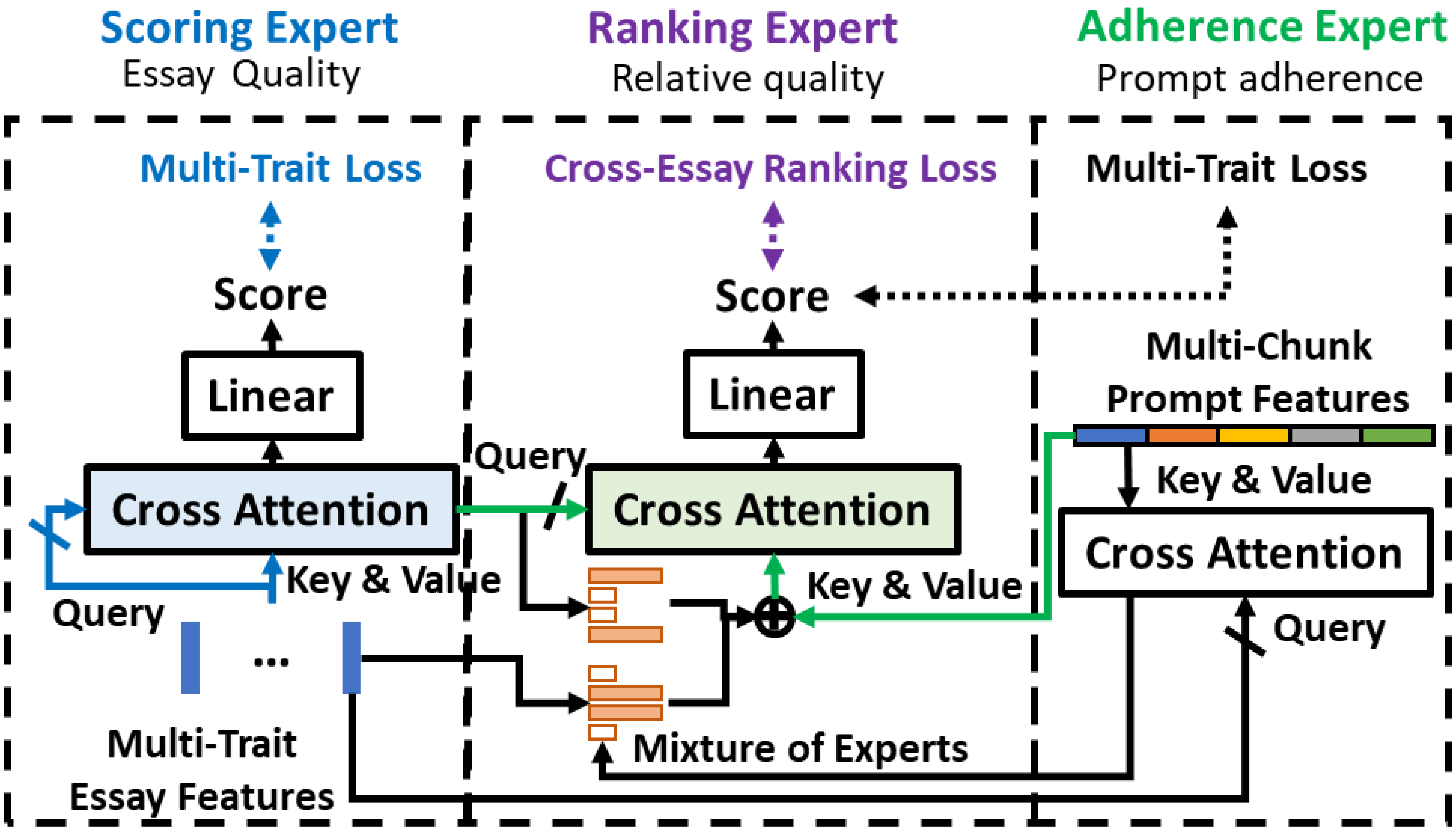
Novelty 2: From Scoring to **Scoring Cue Retrieval**

- Issue:
  - When training a cross-prompt model, the number of available **prompts** is severely **limited**.
  - Focusing on learning essay representation (query) may lead to **overfitting on seen prompts**.

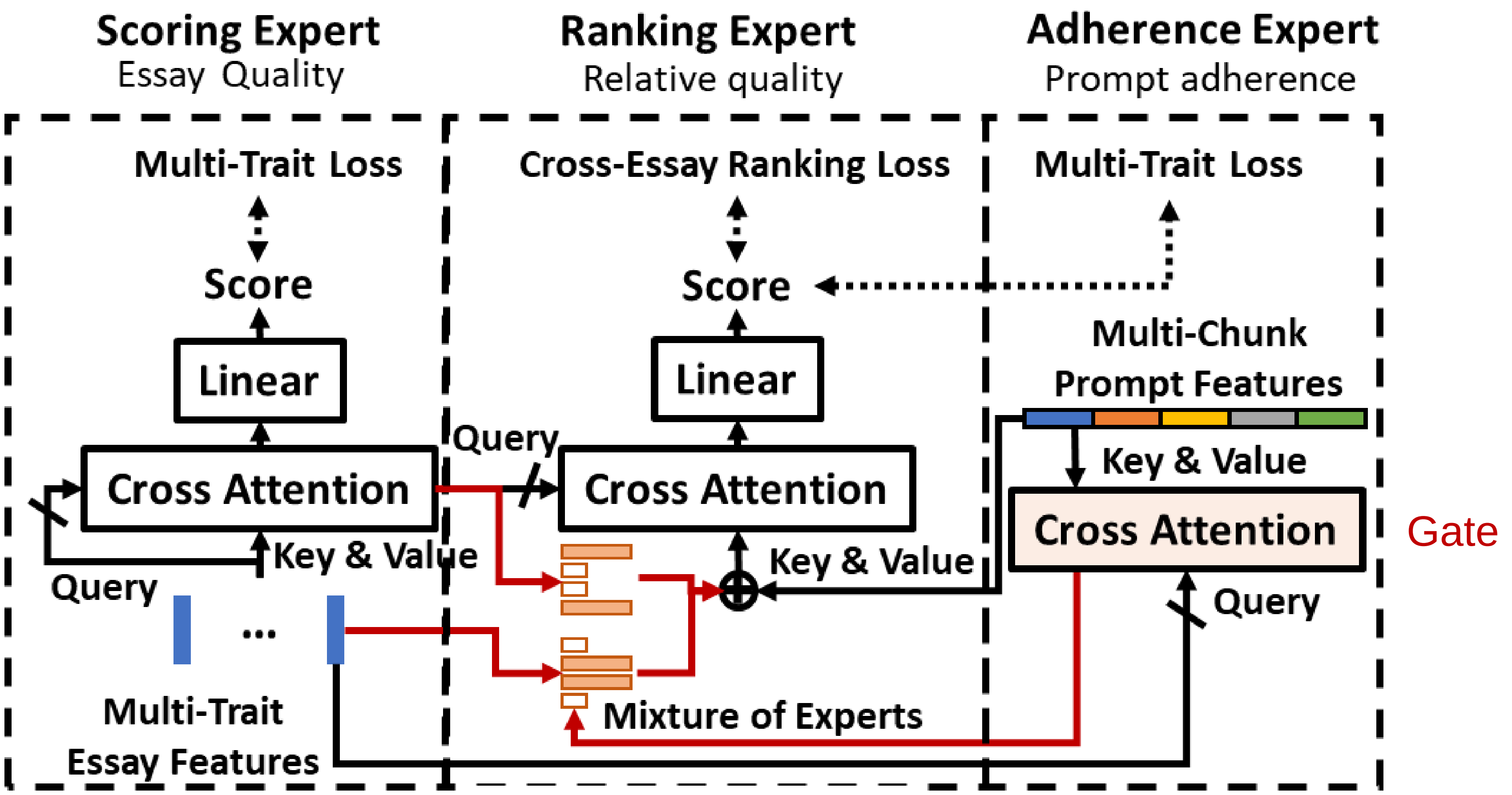
- Solution:
  - Apply a **stop-gradient** operation to the **query**, preventing its representation from being updated during backpropagation.
  - The **fixed query** serves as a stable anchor for **retrieving relevant scoring cues**.



Novelty 3: Mixture of Ordered Scoring Experts



Novelty 3: Mixture of Ordered Scoring Experts



$$y = \sum_{i=1}^{n} G(x)_i \cdot E_i(x)$$

$$y = \sigma(CA(SG(F_{e1}), F_p)) \cdot E_1(F_{e1}) + (1 - \sigma(CA(SG(F_{e1}), F_p))) \cdot E_2(F_{e2})$$

Experiment settings

- Dataset: ~13,000 essays from ASAP++ (Mathias & Bhattacharyya, LREC 2018)

Prompt	Essay Type	Content	Organization	Word Choice	Sentence Fluency	Conventions	Prompt Adherence	Language	Narrativity
1	Argumentative	✓	✓	✓	✓	✓			
2	Argumentative	✓	✓	✓	✓	✓			
3	Response (Source-Dependent)	✓					✓	✓	✓
4	Response (Source-Dependent)	✓					✓	✓	✓
5	Response (Source-Dependent)	✓					✓	✓	✓
6	Response (Source-Dependent)	✓					✓	✓	✓
7	Narrative	✓	✓			✓			
8	Narrative	✓	✓	✓	✓	✓			

- Cross-prompt setting:
  - Leave-one-prompt-out
  - Train on 7 prompts, test on 1 unseen prompt
- Evaluation metric:
  - Quadratic Weighted Kappa (QWK)

Comparisons with State-of-The-Arts

Model	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Prompt 6	Prompt 7	Prompt 8	AVG	STD
PAES (Ridley et al., 2020)	.605	.522	.575	.606	.634	.545	.356	.447	.536	.088
PMAES (Chen and Li, 2023)	.656	.553	.598	.606	.626	.572	.386	.530	.566	.078
CTS (Ridley et al., 2021)	.623	.540	.592	.623	.613	.548	.384	.504	.553	.076
RDCTS (Sun et al., 2024)	.651	.553	.608	.623	.651	.580	.375	.529	.571	.085
ProTACT (Do et al., 2023)	.647	.587	.623	.632	.674	.584	.446	.541	.592	.067
LLMs-based EPCTS (Xu et al., 2025)	.659	.609	.619	**.686**	.671	**.629**	.555	**.630**	.632	.038
OSE (Ours)	.679	.612	**.660**	.660	.686	.596	.581	.627	.638	.037
MOOSE (Ours)	**.685**	**.613**	.657	.652	**.700**	.615	**.592**	.621	**.642**	**.036**

Table 2: Comparison of average QWK for each prompt on the ASAP++ dataset, **bold font** indicates best performance.

Model	Overall	Content	Organization	WC	SF	Convention	PA	Language	Narrativity	AVG	STD
PAES (Ridley et al., 2020)	.657	.539	.414	.531	.536	.367	.570	.531	.605	.527	.075
PMAES (Chen and Li, 2023)	.671	.567	.481	.584	.582	.421	.584	.545	.614	.561	.060
CTS (Ridley et al., 2021)	.670	.555	.458	.557	.545	.412	.565	.536	.608	.586	.062
RDCTS (Sun et al., 2024)	.673	.561	.480	.591	.576	.426	.609	.560	.634	.568	.065
ProTACT (Do et al., 2023)	.674	.596	.518	.599	.585	.450	.619	.596	.639	.586	.058
LLMs-based EPCTS (Xu et al., 2025)	**.728**	.630	.606	.614	.617	.525	.630	.613	.647	.623	.035
OSE (Ours)	.677	.643	.639	**.641**	.635	.575	.637	.610	.649	.634	.023
MOOSE (Ours)	.650	**.651**	**.652**	.634	**.643**	**.604**	**.649**	**.624**	**.665**	**.641**	**.018**

↓ ~50%

Table 3: Comparison of average QWK for each trait on the ASAP++ dataset, **bold font** indicates best performance.

Analysis of cross-prompt essay scoring

Model	P1	P2	P3	P4	P5	P6	P7	P8
prompt as query	**.677**	.611	.643	.664	.646	.576	.480	.427
essay as query	.675	**.617**	**.654**	**.668**	**.686**	**.600**	**.528**	**.560**

Table 5: Analysis of query type on each prompt.

Model	P1	P2	P3	P4	P5	P6	P7	P8
scoring	.639	.593	.603	.604	**.657**	**.555**	.469	.594
cue retrieval	**.645**	**.616**	**.613**	**.617**	.648	.553	**.477**	**.600**

Table 6: Analysis of learning goal on each prompt.

Model	P1	P2	P3	P4	P5	P6	P7	P8
scoring experts	.648	.608	.592	.638	.651	.535	.484	**.616**
ranking experts	.630	.583	.636	.656	.683	.575	**.579**	.514
ordered experts	**.675**	**.617**	**.654**	**.668**	**.686**	**.600**	.528	.560

Table 7: Analysis of expert type on each prompt.

- scoring experts: multi-trait loss, multi-trait loss
- ranking experts: multi-trait loss + ranking loss, multi-trait loss + ranking loss
- ordered experts: multi-trait loss, multi-trait loss + ranking loss

- Using essay as query strongly improves the performance via estimating distribution of essay over prompt and essay.
- Reformulating learning goal to cue retrieval makes the model more robust on the unseen prompts.
- The ordered experts get outstanding performance on essay scoring by imitating scoring process of human raters, from holistic evaluation to ranking and then prompt adherence.

Analysis of trait scoring

Model	Overall	T1	T2	T3	T4	T5	T6	T7	T8
prompt as query	.631	.607	.547	.575	.552	.478	.628	.593	.645
essay as query	**.678**	**.627**	**.603**	**.634**	**.601**	**.522**	**.638**	**.610**	**.658**

Table 8: Analysis of query type on each trait.

Model	Overall	T1	T2	T3	T4	T5	T6	T7	T8
rank→score	.633	.595	**.581**	**.620**	**.619**	**.525**	.592	.588	.611
score→rank	**.649**	**.605**	.568	.577	.553	.506	**.622**	**.605**	**.646**

Table 9: Analysis of experts' order on each trait.

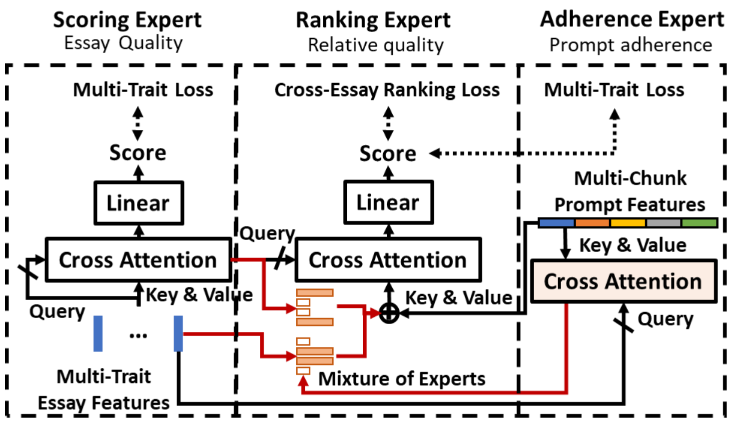
Model	Overall	T1	T2	T3	T4	T5	T6	T7	T8
scoring experts	.632	.603	.571	.628	.612	.509	.608	.591	.628
ranking experts	.666	.603	.569	.585	.567	.512	.613	.603	.628
ordered experts	**.678**	**.627**	**.603**	**.634**	**.601**	**.522**	**.638**	**.610**	**.658**

Table 10: Analysis of expert type on each trait.

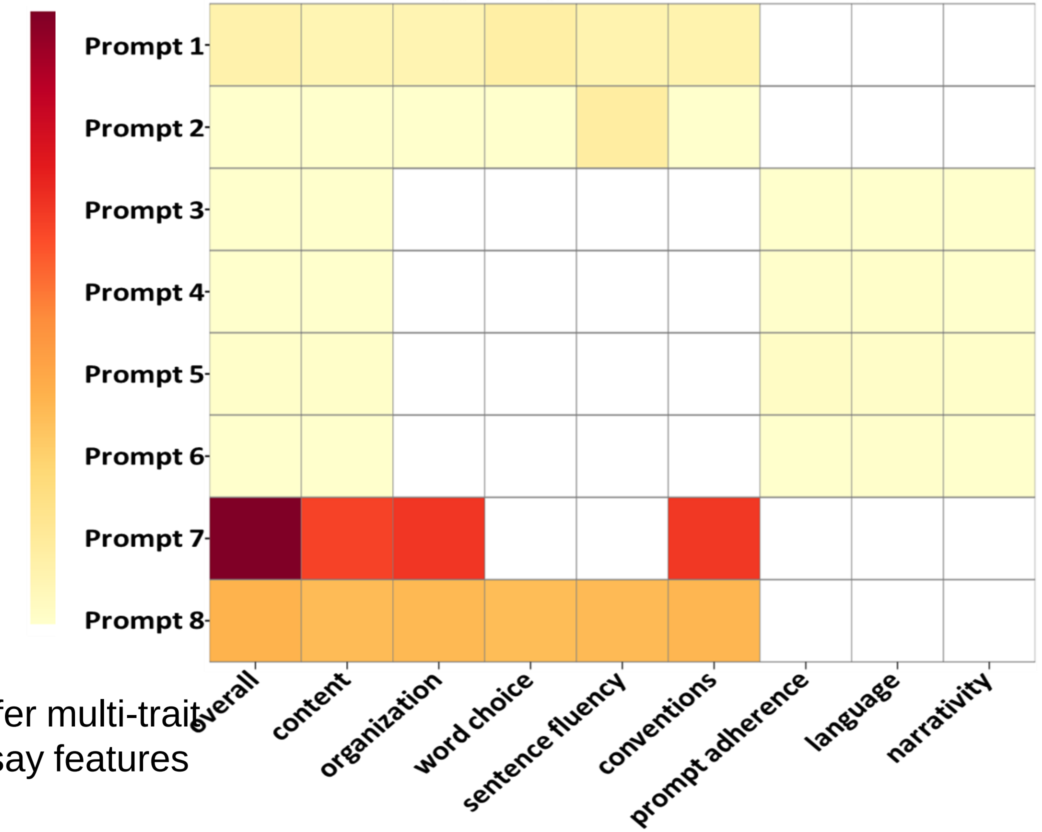
T1: Content T2: Organization T3: Word choice T4: Sentence Fluency	T5: Convention T6: Prompt adherence T7: Language T8: Narrativity

- **Essay-as-query** increase scoring ability in all of the traits.
- The effect of expert ordering:
  - rank → score: argumentative prompts on **Organization**, **Word Choice**, **Sentence Fluency**, **Convention**
  - score → rank: response prompts on **Prompt Adherence**, **Language**, and **Narrativity**.
- **Ordered Score Experts** achieve the best results for all traits, confirming that imitates the human scoring process is a promising strategy.

Visualization on MoE gating



Prefer features refined by scoring expert



Prefer multi-trait essay features

- **Narrative prompts (P7, P8):**
  - Prefer **refined** feature from scoring expert
    - require high-level semantic features (open-ended prompt).
- **Response prompt (P3~P6):**
  - Select **original** multi-trait essay features:
    - rely on original essay features (source-focused).
- **Argumentative prompt (P1, P2):**
  - **Moderate** preference for refined features
    - support opinions; need semantic cues sometimes

Conclusion

- MOOSE imitates the scoring process of human experts,
  - **a scoring expert** to assess the inherent quality of the essay,
  - **a ranking expert** to compare relative quality across different essays,
  - **an adherence expert** to measure the relation between the essay-prompt pair.
- We introduce essay query, query detach, and MoE techniques, which enable MOOSE to capture fine-grained features and focus on retrieving useful scoring cues.
- MOOSE achieves impressive performance on the ASAP++ cross-prompt essay trait scoring task, surpassing current SOTA built on LLMs.