KNOWLEDGE INNOVATION RESEARCH CENTER

KAIST ISE

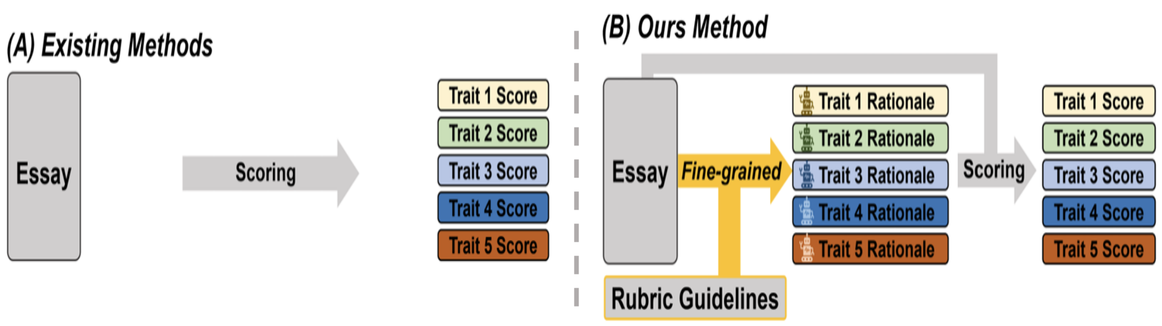
NAACL 2025

Rationale Behind Essay Scores: Enhancing S-LLM's Multi-Trait Essay Scoring with Rationale Generated by LLMs

Jong Woo Kim*, Seong Yeub Chu*, Bryan Wong, Mun Yong Yi
Korea Advanced Institute of Science and Technology (KAIST)

* : Co-Autor

Background



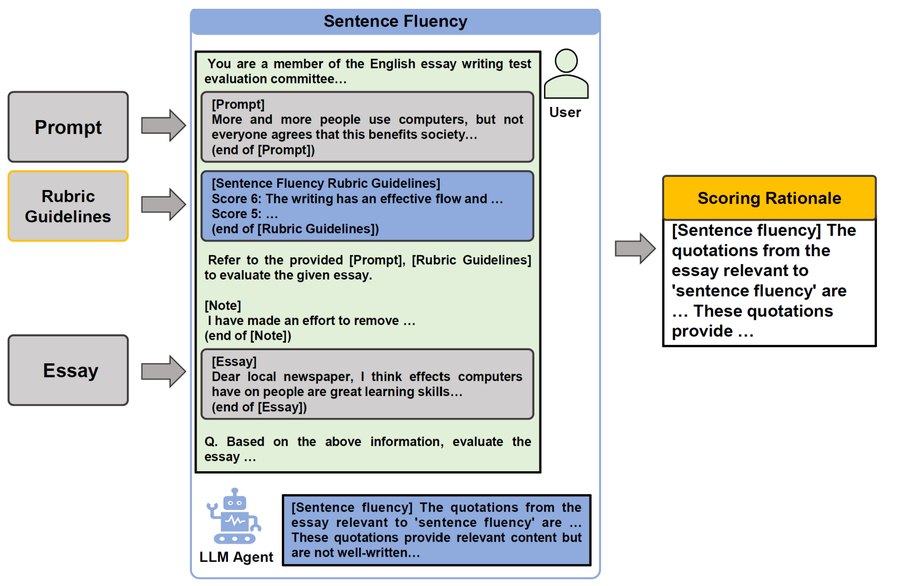
- Existing studies rely solely on essay text, lacking alignment with evaluation rubrics.
- Multi-trait scoring models fail to adequately consider interactions between evaluation criteria.

Research Question

RQ1. What are the key findings from the analysis of LLM generated rationales for essay evaluation?
RQ2. To what extent does incorporating rationales improve the reliability of multi-trait essay scoring using S-LLMs?

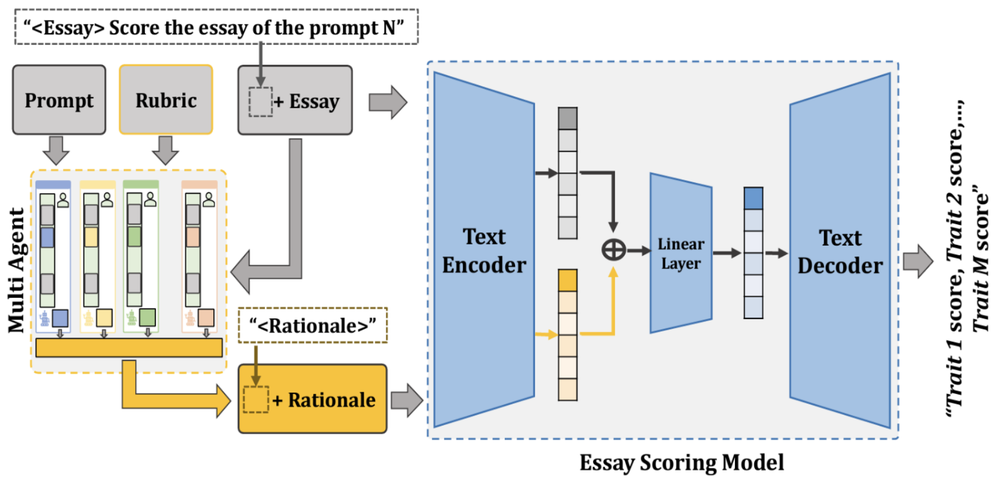
Methodology

LLM-based trait-wise rationale generation system



- Prompt: An introduction providing the topic, purpose, or specific direction for an essay.
- Rubric Guidelines: Descriptors for each trait.
- Essay : A structured piece of writing that argues a topic.

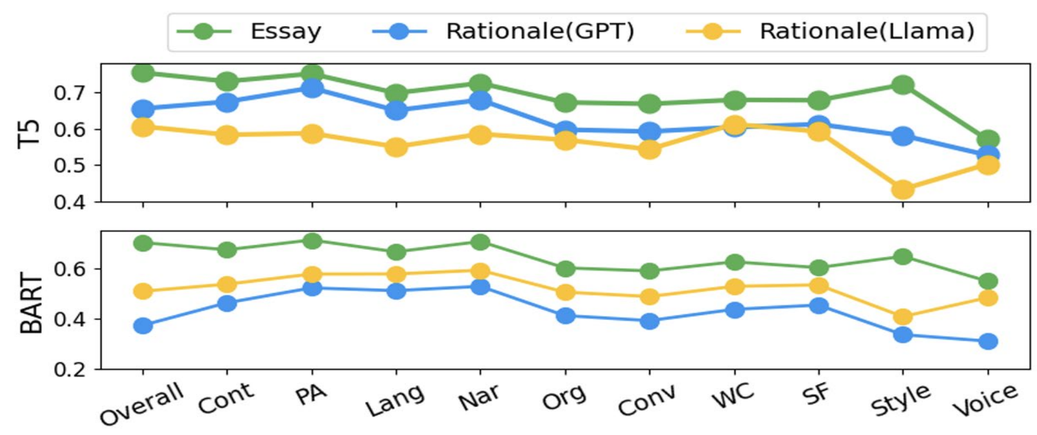
Representation Extraction and Scoring



- Text Encoder: Encodes essays and rationales into representation embeddings.
- Linear Layer: Aggregates the essay and the rationale into a unified representation.
- Text Decoder: Generates multi-trait score sequences.

Results

Faithfulness of rationales (RQ1)



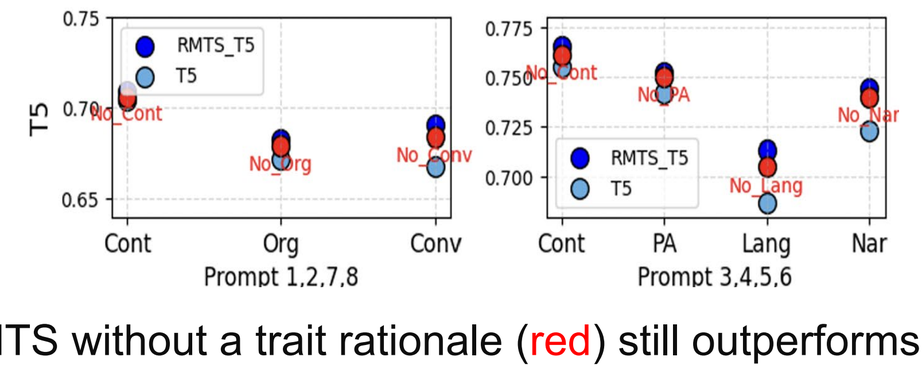
- Compare the performance of using only Essay (green) versus using only Rationale (blue, orange).
- Rationale positively impacts score prediction and effectively explains the basis of the Essay score.

Performance comparison (RQ2)

	Trait (Prediction Order: ←)											
Model	**Overall**	**Cont**	**PA**	**Lang**	**Nar**	**Org**	**Conv**	**WC**	**SF**	**Style**	**Voice**	**AVG↑ (SD↓)**
HISK	0.718	0.679	0.697	0.605	0.659	0.610	0.527	0.579	0.553	0.609	0.489	0.611 (0.004)
STL-LSTM	0.750	0.707	0.731	0.640	0.699	0.649	0.505	0.621	0.612	0.609	0.544	0.642 (0.073)
MTL-BiLSTM	**0.764**	0.685	0.701	0.604	0.668	0.615	0.560	0.615	0.598	0.632	0.582	0.639 (0.057)
PMAES	0.671	0.567	0.584	0.545	0.614	0.481	0.421	0.584	0.582	-	-	0.614 (-)
PLAES	0.673	0.574	0.601	0.554	0.631	0.491	0.447	0.579	0.580	-	-	0.631 (-)
T5 (ArTS)	0.754	0.730	0.751	0.698	0.725	0.672	0.668	0.679	0.678	**0.721**	0.570	0.695 (0.018)
+ RMTS(G) (+%)	0.755 (+0.1)	**0.737** (+0.7)	**0.752** (+0.1)	**0.713** (+1.5)	**0.744** (+1.9)	0.682 (+1.0)	**0.690** (+2.2)	**0.705** (+2.6)	**0.694** (+1.6)	0.702 (-1.9)	0.612 (+4.2)	**0.708** (0.043)
+ RMTS(L) (+%)	0.754 (+0.0)	0.730 (+0.0)	0.749 (-0.2)	0.701 (+0.3)	0.737 (+1.2)	0.675 (+0.3)	0.684 (+1.6)	0.690 (+1.1)	0.684 (+0.6)	0.696 (-2.5)	0.640 (+7.0)	0.704 (0.042)
Flan-T5	0.662	0.645	0.615	0.539	0.577	0.646	0.636	0.694	0.667	0.578	0.624	0.626 (0.064)
+ RMTS(G) (+%)	0.732 (+7.0)	0.733 (+8.8)	0.750 (+13.5)	0.708 (+16.9)	0.737 (+16.0)	**0.684** (+3.8)	0.680 (+4.4)	0.691 (-0.3)	0.680 (+1.3)	0.688 (+11.0)	0.563 (-6.1)	0.695 (0.048)
+ RMTS(L) (+%)	0.723 (+6.1)	0.717 (+7.2)	0.736 (+12.1)	0.696 (+15.7)	0.722 (+14.5)	0.663 (+1.7)	0.662 (+2.6)	0.673 (-2.1)	0.663 (-0.4)	0.695 (+11.7)	0.620 (-0.4)	0.688 (0.054)
BART	0.701	0.672	0.711	0.664	0.705	0.600	0.588	0.624	0.601	0.646	0.547	0.642 (0.054)
+ RMTS(G) (+%)	0.720 (+1.9)	0.710 (+3.8)	0.731 (+2.0)	0.683 (+1.9)	0.720 (+1.5)	0.651 (+5.1)	0.637 (+4.9)	0.685 (+6.1)	0.655 (+5.4)	0.661 (+1.5)	**0.649** (+10.2)	0.674 (0.046)
+ RMTS(L) (+%)	0.724 (+2.3)	0.704 (+3.2)	0.732 (+2.1)	0.677 (+1.3)	0.714 (+0.9)	0.658 (+5.8)	0.647 (+5.9)	0.671 (+4.7)	0.662 (+6.1)	0.673 (+2.7)	0.596 (+4.9)	0.678 (0.037)

- S-LLMs (T5, Flan-T5, BART) show relatively lower performance, but RMTS significantly improve results.
- GPT-based models (**G**PT-3.5-Turbo, **L**lama-3.1-8B-Instruct) demonstrate consistently high performance.

Ablation Study



- RMTS without a trait rationale (red) still outperforms vanilla models (sky) without any rationale input.
- Trait rationales not only influence their own assessments but also interact with and affect the evaluation of other traits.

Conclusion

- The study introduces RMTS, a framework combining LLM-generated rationales with essays for multi-trait scoring.
- It shows that trait-specific rationales improve S-LLMs' scoring and enhance trait evaluation, offering significant benefits for formative assessments.