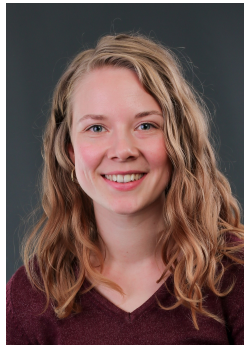



Workshop “Large Language Models in Education”

Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation

Maja Stahl, Leon Biermann, Andreas Nehring, **Henning Wachsmuth**

Slides by Maja Stahl, presented by Henning Wachsmuth

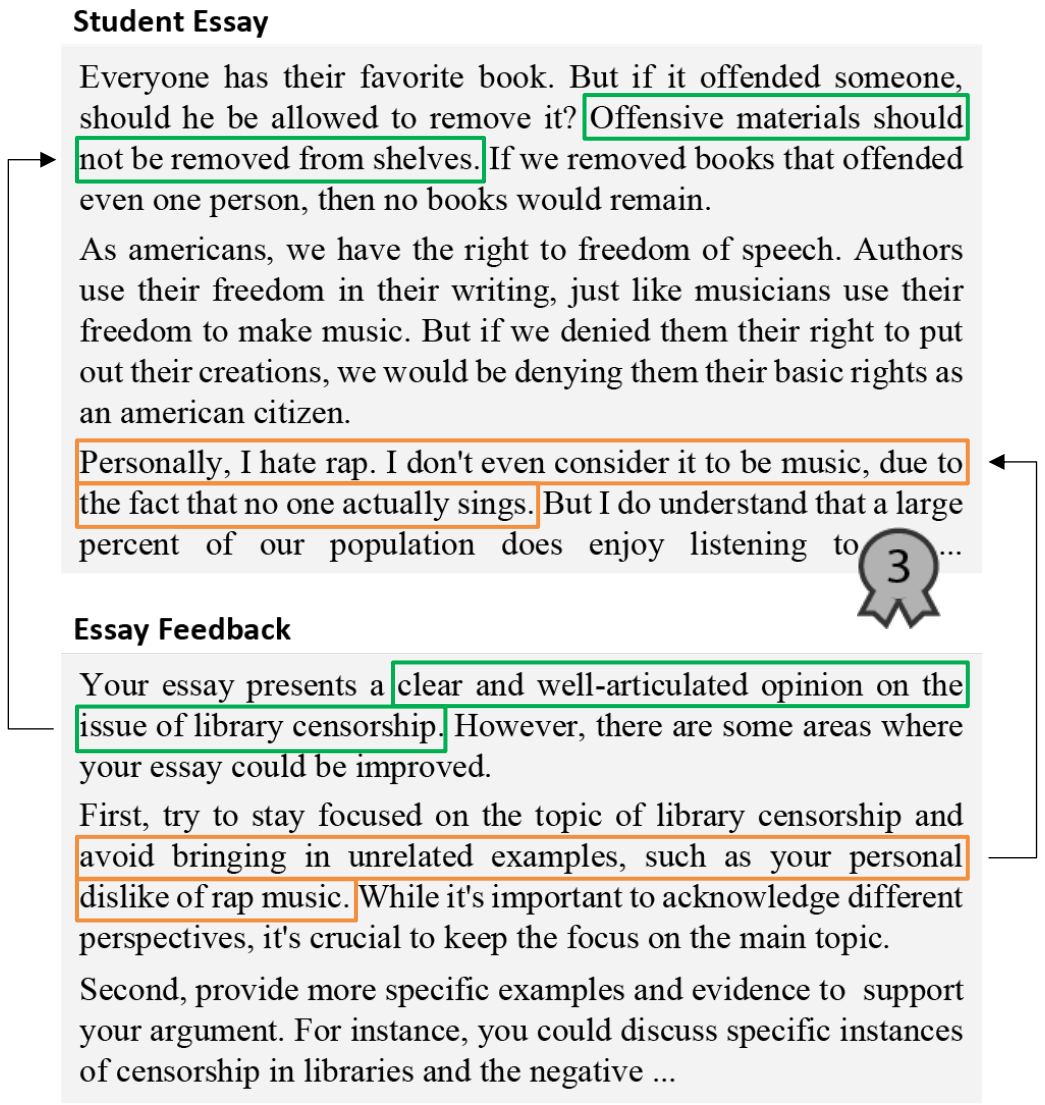
Best Long Paper Award at the
19th Workshop on Innovative Use of NLP for Educational Applications (BEA 2024)

Natural Language Processing

Leibniz Universität Hannover

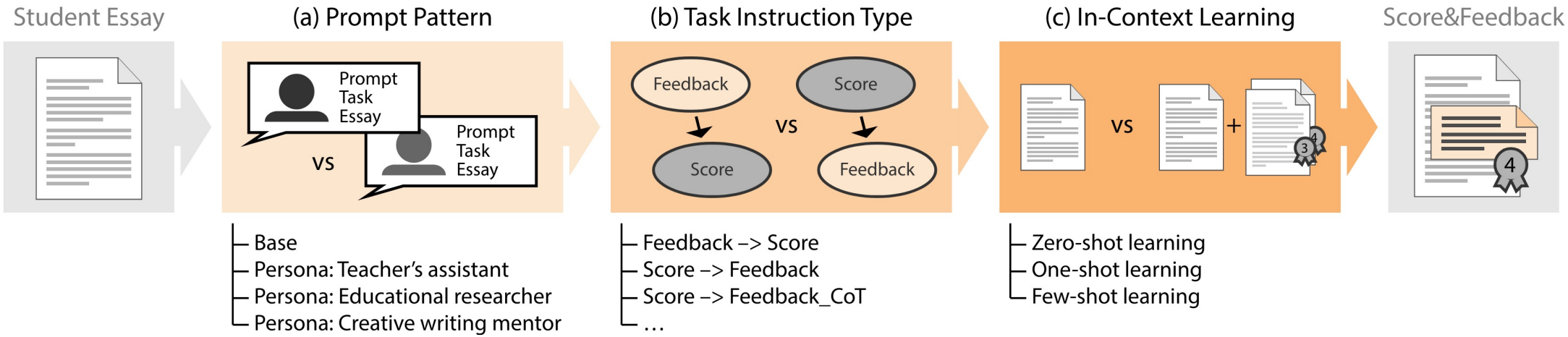
Introduction

- **Essay scoring**
  - Score essay quality using a predefined scale and rubric
  - Holistically or criteria-based
- **Essay feedback generation**
  - Generate textual feedback on an essay
  - The goal is to help students improve their essays
- **Research questions**
  1. How to prompt LLMs to create helpful essay feedback?
  2. Can feedback generation benefit essay scoring?
  3. Can essay scoring benefit feedback generation?



Study

- Compare LLM prompting strategies for essay scoring and feedback generation



Base: "You are given an essay written by a student and the corresponding prompt for the 7th to 10th grade student…"

Teacher's Assistant: "Imagine you are a teacher's assistant in a middle school tasked with reviewing a 7th to 10th grade student's essay…"

Feedback_dCoT→Scoring: "[…] Let's think step by step. First, analyze the quality of the essay in terms of the given rubric. Then, give feedback to the student that explains their mistakes and errors and additionally gives them tips to avoid them in the future. As a final step, output the score at the end."

Explanation→Scoring: "Analyze the given essay using the following rubric: *{rubric}*. To do this, first explain using the scoring rubric why you chose the score. After you analyzed the essay, give a final grade."

Exemplary essays, together with their score and a reasoning for the score.

- **One-shot:** Randomly select an essay with a medium score.
- **Few-shot:** Sample essays with the highest and lowest scores first, then cover medium scores.

Experimental Setup

- **ASAP corpus** (Hamner et al., 2012)
  - 12,980 essays written by school students
  - 8 essay datasets that differ by
    - Essay prompt
    - Scoring range
    - Rubric used by the raters
- **Instruction-following LLM**
  - Mistral with 7 billion parameters
    Used model: Mistral-7B-Instruct-v0.2
- **Scoring metric**
  - QWK (quadratic weighted kappa)
  - From –1 (worst) to 1 (best)

Rubric Guidelines

Score 3: The response demonstrates an understanding of the complexities of the text.

- Addresses the demands of the question
- Uses expressed and implied information from the text
- Clarifies and extends understanding beyond the literal

Score 2: The response demonstrates a partial or literal understanding of the text.

- Addresses the demands of the question, although may not develop all parts equally
- Uses some expressed or implied information from the text to demonstrate understanding
- May not fully connect the support to a conclusion or assertion made about the text(s)

Score 1: The response shows evidence of a minimal understanding of the text.

- May show evidence that some meaning has been derived from the text
- May indicate a misreading of the text or the question
- May lack information or explanation to support an understanding of the text in relation to the question

Score 0: The response is completely irrelevant or incorrect, or there is no response.

Exemplary rubric from essay dataset 3

Essay Scoring: Evaluation

- **What works best for essay scoring?**
  - Prompt pattern. Personas *educational researcher (ER)* and *teacher's assistant (TA)*
  - Task instruction type. First follow task-specific steps (*Feedback_dCoT→Scoring*) or first give explanation (*Explanation→Scoring*), then score essay
  - In-context learning. Giving examples of scored essays (*One-shot*, *Few-shot*)

Task Instruction Type	Essay Set							
	1	2	3	4	5	6	7	8
Scoring	.448	.585	.479	.596	.557	.649	.438	.48
Scoring→Feedback	.510	**.615**	.439	.530	.489	.621	.449	.48
Feedback→Scoring	.388	.561	.484	.600	**.622**	.630	.385	**.545**
Scoring→Feedback_CoT	.538	.595	.422	.494	.530	.635	.458	.477
Feedback_dCoT→Scoring	**.546**	.564	.424	.558	.581	.628	**.477**	.489
Scoring→Explanation	.466	.580	.472	.565	.541	.639	.420	.417
Explanation→Scoring	.470	.553	**.488**	.636	.571	**.675**	.384	.484

Pattern	Essay Set								
	1	2	3	4	5	6	7	8	Mean
Base	.495	.532	.405	.495	.497	.601	.436	.377	.480
TA	**.536**	**.603**	.408	.499	.512	**.625**	.443	.439	.508
ER	.436	.554	**.460**	**.560**	**.553**	.620	.418	**.467**	**.509**
CWM	.484	.588	.382	.434	.507	.596	**.471**	.352	.477

Context	Essay Set								
	1	2	3	4	5	6	7	8	Mean
Zero-shot	.510	.615	.439	.530	.489	.621	.449	**.481**	.517
One-shot	**.565**	**.619**	**.523**	**.600**	.606	.665	**.509**	.233	**.540**
Few-shot	.558	.586	.515	.586	**.618**	**.671**	.472	.297	.538

Feedback Generation: Automatic evaluation

- **Automatic evaluation**
  - LLM evaluates the generated feedback from 1 (not helpful) to 10 (very helpful)
  - Model used: Mistral
- **What works best for feedback generation?**
  - Prompt pattern. Using the persona *educational researcher (ER)*
  - Task instruction type. Generating feedback only (*Feedback*)
  - In-context learning. Giving examples of scored essays with explanation (*One-shot*, *Few-shot*)

Prompt Pattern	Mistral
Base	7.78
Teacher's assistant (TA)	7.90
Educational researcher (ER)	**8.26**
Creative writing mentor (CWM)	7.83

Task Instruction Type	Mistral
Feedback	**8.96** ±.25
Scoring→Feedback	8.04 ±.44
Feedback→Scoring	8.27 ±.38
Scoring→Feedback_CoT	7.30 ±.63
Feedback_dCoT→Scoring	8.53 ±.66
Scoring→Explanation	7.22 ±.45
Explanation→Scoring	7.27 ±.63

In-Context Learning	Mistral
Zero-shot learning	8.04 ±.44
One-shot learning	8.39 ±.54
Few-shot learning	**8.42** ±.56

Feedback Generation: Manual evaluation

- **Manual evaluation**
  - 3 best prompting strategies from automatic evaluation
  - 12 human annotators, 24 randomly-selected feedback texts
  - 5 statements judged on 7-point Likert scale from 1 (I strongly disagree) to 7 (I fully agree)
- **Statements**
  - S1. The feedback clearly **points out mistakes** that were made in the essay.
  - S2. The feedback explains exactly **why the errors are errors**.
  - S3. The feedback **is very clear and precise** so that the student can understand it.
  - S4. The feedback is absolutely **suitable for students** from 7th to 10th grade.
  - S5. Overall, the feedback is **very helpful**.
- **Findings**
  - All feedback generally perceived as rather helpful
  - *Feedback* only achieved the highest scores

Task Instruction Type	S1	S2	S3	S4	S5
Feedback	**5.88**	**5.71**	**6.04**	**5.75**	**6.08**
Feedback→Scoring	5.17	5.04	5.46	5.21	5.08
Feedback_dCoT→Scoring	5.50	4.92	5.29	4.83	5.00

Feedback Generation: Automatic vs. manual evaluation

- **How reliable is automatic evaluation?**
  - Pearson correlation of LLM-based helpfulness scores with manual statement judgments (S1–S5)
  - Two models used: Mistral, Llama-2

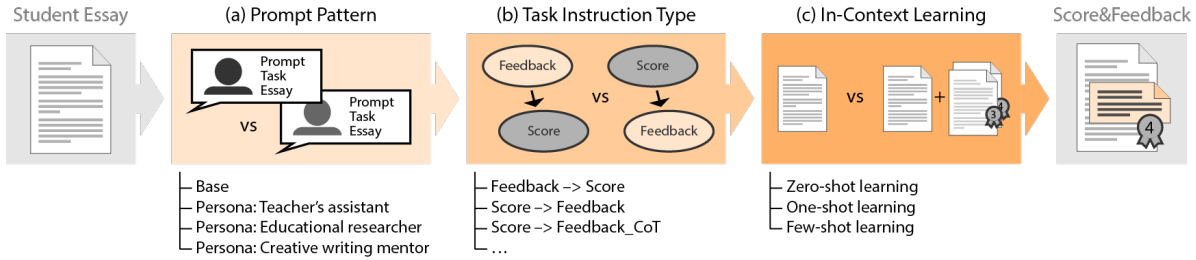
- **Statements**
  - S1. The feedback clearly **points out mistakes** that were made in the essay.
  - S2. The feedback explains exactly **why the errors are errors**.
  - S3. The feedback **is very clear and precise** so that the student can understand it.
  - S4. The feedback is absolutely **suitable for students** from 7th to 10th grade.
  - S5. Overall, the feedback is **very helpful**.

Autom. Evaluation	S1	S2	S3	S4	S5
Mistral	0.29	0.27	0.45	0.25	**0.61**
Llama-2	–0.11	–0.11	–0.02	0.07	–0.10

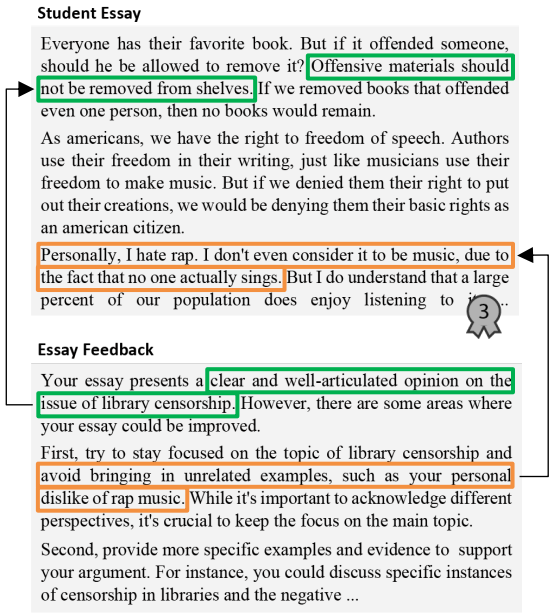
- **Findings**
  - Mistral can be useful for automatically evaluating feedback helpfulness
  - Overall helpfulness judgment has highest correlation

Summary

- **LLM prompting strategies**
  - Joint essay scoring and feedback generation
  - Systematic exploration of three main aspects of prompting strategies



- **Interaction of the two tasks**
  - Feedback generation benefits essay scoring
  - Impact of essay scoring on feedback generation seems rather low (if not negative)



- **More details in the paper**
  - Definitions of all task instruction types
  - Full prompts
  - Examples of scores and generated feedback

References

- **Hamner et al. (2012).** Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. 2012. The hewlett foundation: Automated essay scoring.