Contribution ID: **202** Type: **Oral Presentation**

Automated Scoring of Open-Ended Responses: Evaluating LLMs and Prompting Strategies

Wednesday 23 July 2025 18:00 (15 minutes)

Before the rapid development of artificial intelligence, standardized tests mainly relied on multiple-choice questions because evaluating open-ended tasks required significant resources. Modern large language models (LLMs), such as ChatGPT, Gemini, and Llama, now enable automated assessment of open-ended tasks. Unlike traditional machine learning or deep learning methods, foundational LLMs do not require labeled datasets or extensive expertise in data science and programming. Users only need to create a well-structured prompt and verify alignment with human raters on a small sample.

This study aims to assess the feasibility of using LLMs to score open-ended tasks through prompting. Additionally, it explores effective prompting strategies. Several LLMs were compared, with ChatGPT-4o serving as the baseline model. It was evaluated against ChatGPT-o3-mini, which features advanced latent reasoning. YandexGPT4 was also included, as it has been specifically trained in Russian, the language used by respondents in this study. To explore alternatives to cloud-based solutions, the DeepSeek r1 8B and Llama 3.1 8B models were tested, as they can run locally on a computer, reducing evaluation costs and ensuring confidentiality.

To evaluate the effectiveness of different prompting strategies, both analytical and holistic rubrics were used. Techniques included zero-shot (no examples), one-shot (one example), and few-shot (multiple examples). Additional strategies included chain-of-thought prompting (reasoning before making a decision), tree-of-thought reasoning (deliberation among "multiple evaluators" before deciding), and enhancing LLM "motivation" using specific phrases.

The models were tested on three types of tasks: (1) short-answer reading comprehension tasks for children aged 10-11 (1-2 sentences); (2) a prompt engineering task for undergraduate students (up to 150 words); and (3) an economics essay task for economics students (up to 144 words). LLM performance was measured using Weighted Quadratic Cohen's Kappa (WQK) and Mean Absolute Error (MAE). Scores were compared with human raters, who evaluated 50 randomly selected responses for each task type.

The presentation will compare five LLMs across three types of open-ended tasks. Results show that LLMs can achieve sufficient accuracy for use in low-stakes assessment scenarios. Additionally, findings on effective prompting strategies will be shared, identifying best practices. Specifically, analytical rubrics combined with multiple examples provided the most accurate assessments. Interestingly, phrases like "focus and take a deep breath before answering" or "I will pay you for good work" resulted in accuracy comparable to example-based prompting. These results highlight the potential of LLMs for scoring open-ended responses. At this stage, they can be used in low-stakes assessments or as an assistant to human raters. Furthermore, LLMs can support data annotation and the creation of training datasets for other machine learning models.

Oral presentation

Automated Scoring of Open-Ended Responses: Evaluating LLMs and Prompting Strategies

Author

Daniil Talov

Affiliation

HSE University

Abstract

Before the rapid development of artificial intelligence, standardized tests mainly relied on multiple-choice questions because evaluating open-ended tasks required significant resources. Modern large language models

(LLMs), such as ChatGPT, Gemini, and Llama, now enable automated assessment of open-ended tasks. Unlike traditional machine learning or deep learning methods, foundational LLMs do not require labeled datasets or extensive expertise in data science and programming. Users only need to create a well-structured prompt and verify alignment with human raters on a small sample.

This study aims to assess the feasibility of using LLMs to score open-ended tasks through prompting. Additionally, it explores effective prompting strategies. Several LLMs were compared, with ChatGPT-4o serving as the baseline model. It was evaluated against ChatGPT-o3-mini, which features advanced latent reasoning. YandexGPT4 was also included, as it has been specifically trained in Russian, the language used by respondents in this study. To explore alternatives to cloud-based solutions, the DeepSeek r1 8B and Llama 3.1 8B models were tested, as they can run locally on a computer, reducing evaluation costs and ensuring confidentiality.

To evaluate the effectiveness of different prompting strategies, both analytical and holistic rubrics were used. Techniques included zero-shot (no examples), one-shot (one example), and few-shot (multiple examples). Additional strategies included chain-of-thought prompting (reasoning before making a decision), tree-of-thought reasoning (deliberation among “multiple evaluators” before deciding), and enhancing LLM “motivation” using specific phrases.

The models were tested on three types of tasks: (1) short-answer reading comprehension tasks for children aged 10-11 (1-2 sentences); (2) a prompt engineering task for undergraduate students (up to 150 words); and (3) an economics essay task for economics students (up to 144 words). LLM performance was measured using Weighted Quadratic Cohen’s Kappa (WQK) and Mean Absolute Error (MAE). Scores were compared with human raters, who evaluated 50 randomly selected responses for each task type.

The presentation will compare five LLMs across three types of open-ended tasks. Results show that LLMs can achieve sufficient accuracy for use in low-stakes assessment scenarios. Additionally, findings on effective prompting strategies will be shared, identifying best practices. Specifically, analytical rubrics combined with multiple examples provided the most accurate assessments. Interestingly, phrases like “focus and take a deep breath before answering” or “I will pay you for good work” resulted in accuracy comparable to example-based prompting. These results highlight the potential of LLMs for scoring open-ended responses. At this stage, they can be used in low-stakes assessments or as an assistant to human raters. Furthermore, LLMs can support data annotation and the creation of training datasets for other machine learning models.

Keywords

LLM, Automated Scoring,Open-Ended Tasks

Primary author: TALOV, Daniil (HSE University)

Presenter: TALOV, Daniil (HSE University)

Session Classification: Session 21 : ”Psychometric Innovations and Diagnostic Methodologies”

Track Classification: Measurement: Measurement