NEURAL INFORMATION PROCESSING SYSTEMS

NarrativeBench: Benchmarking Multi-trait Automated Scoring and Feedback Efficacy of Large Language Models for Chinese Narrative Essays

Jin Wu, East China Normal University

華東師範大學 EAST CHINA NORMAL UNIVERSITY

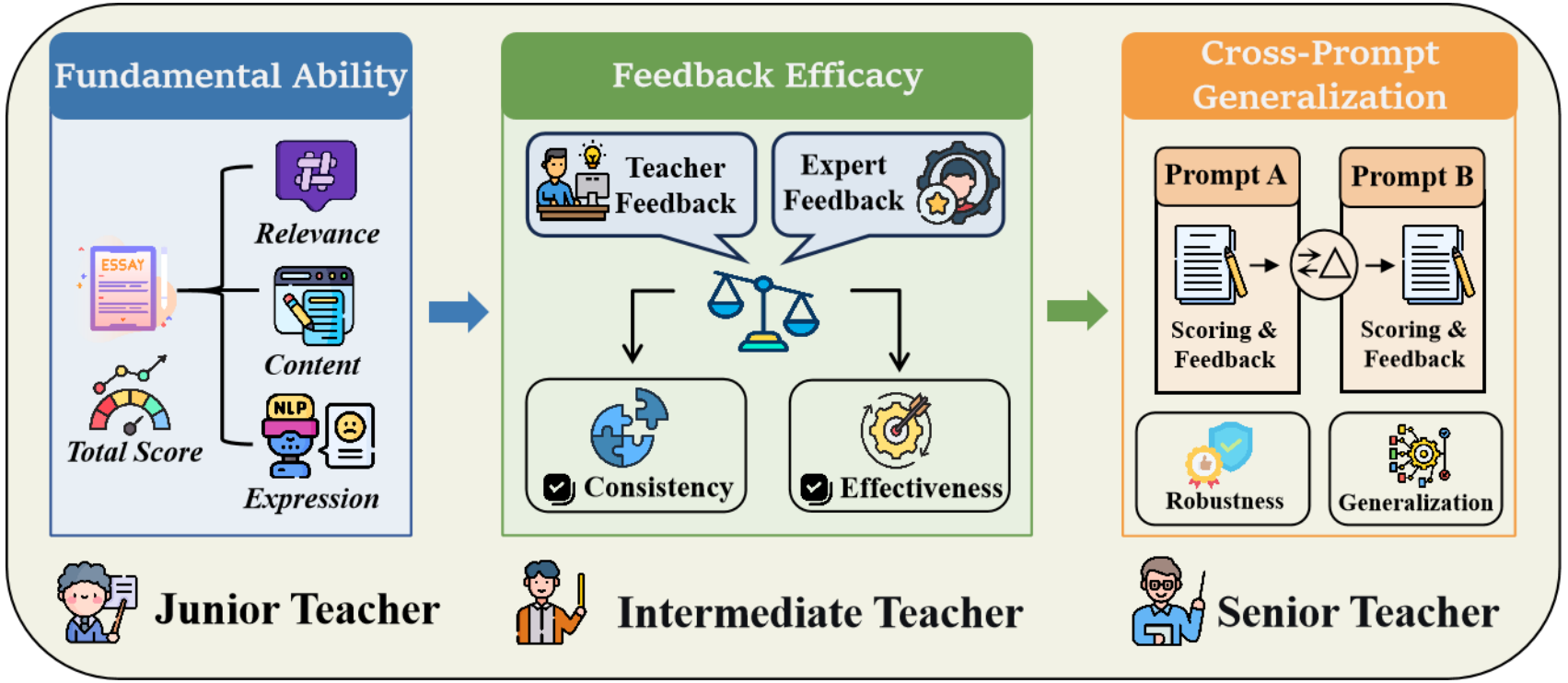
AIE Shanghai 上海智能教育研究院 INSTITUTE of AI EDUCATION, SH

Background: While LLMs have demonstrated transformative progress, **existing essay benchmarks face some challenges**:

- ✓ **Standard Alignment**: The varied complexity of scoring traits and rubrics across different benchmarks;

- ✓ **Evaluation Scope**: Current benchmarks focus on scoring accuracy for LLMs, yet they lack a systematic framework for assessing the quality of generative feedback.

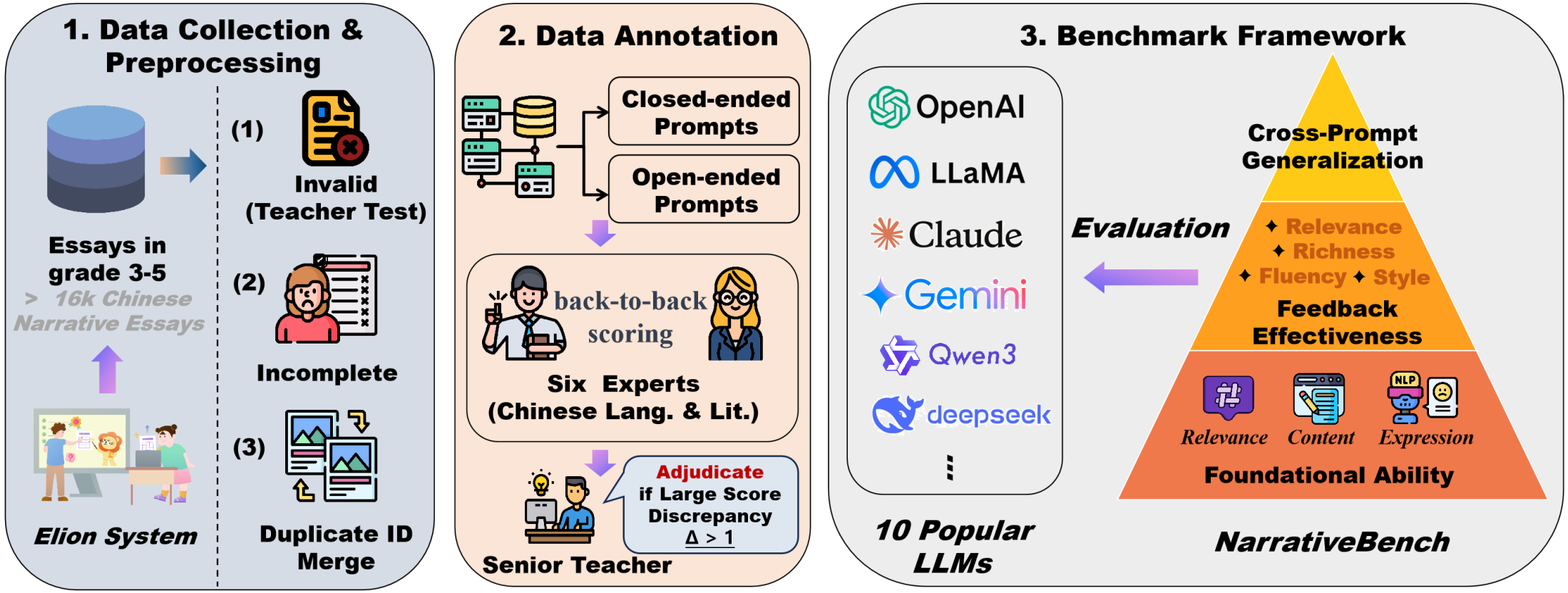
- ✓ **Prompt Coverage**: There is a lack of CNE datasets that cover both closed-ended and open-ended prompts.



(1) ***Fundamental Ability***: evaluating the holistic scoring performance and multi-trait capabilities of LLMs in essays;

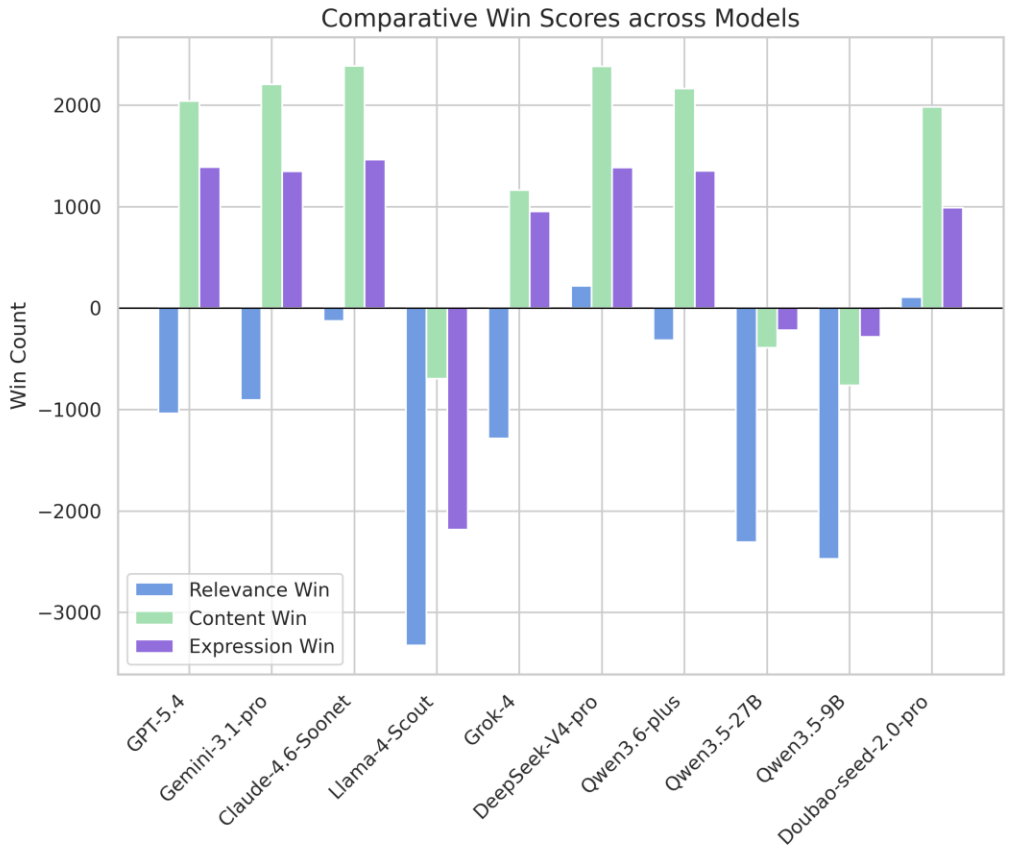
(2) ***Feedback Efficacy***: assessing the alignment and effectiveness of LLM-generated feedback compared to expert-annotated gold standards; and

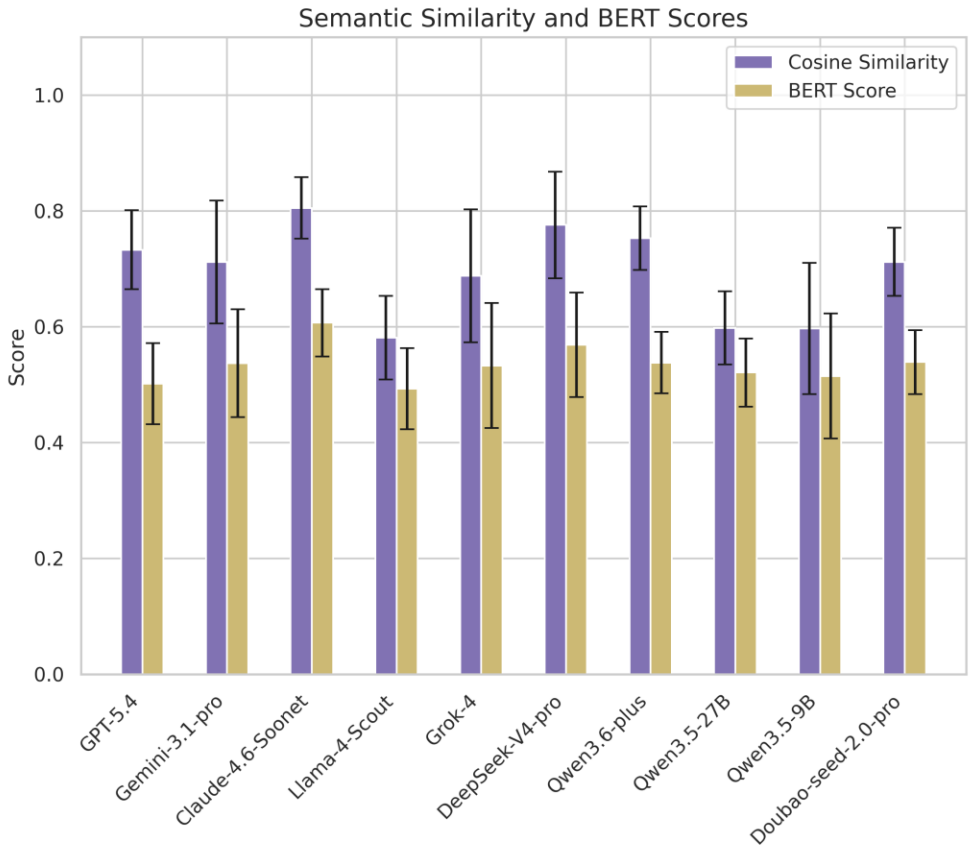
(3) ***Cross-Prompt Generalization***: examining the performance disparities of LLMs under different prompts to examine their generalizability and transferability.



- *we identify relevance, content, and expression as the shared key dimensions for Chinese Narrative Essay (CNE) evaluation.*

- *We introduce NarrativeBench, a comprehensive benchmark comprising 4,500 Chinese narrative essays from grades 3 to 5 across 14 distinct writing prompts (both closed-ended and open-ended).*





- Current models fall substantially short of human expert-level scoring: the best-performing LLM (Claude-4.6-Sonnet) remains far below the LoRA-tuned supervised baseline.

- Results show that LLM limitations stem from insensitivity to linguistic nuance, misunderstanding of prompt intent, and distributional mismatch with elementary-level writing.

YOUR LOGO

THANK YOU