INVESTIGATION OF THE INTER-RATER RELIABILITY BETWEEN LARGE LANGUAGE MODELS AND HUMAN RATERS IN QUALITATIVE ANALYSIS

Nikhil S Borse, Ravishankar Chatta Subramaniam, N. Sanjay Rebello
Department of Physics and Astronomy, Purdue University, West Lafayette, IN, 47906

MOTIVATION

- Human coding in qualitative research is time intensive
- Can LLMs like ChatGPT-4.5 and ChatGPT-4o offer an efficient approach (OpenAI, 2024)?
- Inter-rater reliability (IRR) is crucial for trustworthy results (Tinsley and Weiss, 1975)

METHODS

- Setting: Calculus-based physics lab with 14 student groups
- Students chose a project on Engineering Design **(ED)**
- Team project in groups of 2 to 3 students
- Design Project spanning 4 weeks – Problem definition, 1st solution, iteration, Summary
- Every week Students recorded audio files of peer discussions on ED project strategies
- Two expert human raters coded transcripts using validated rubric (Subramaniam, 2024)
- They discussed and coded to consensus
- An LLM was prompted to code the transcripts
- The human consensus coding was then compared with LLM coding

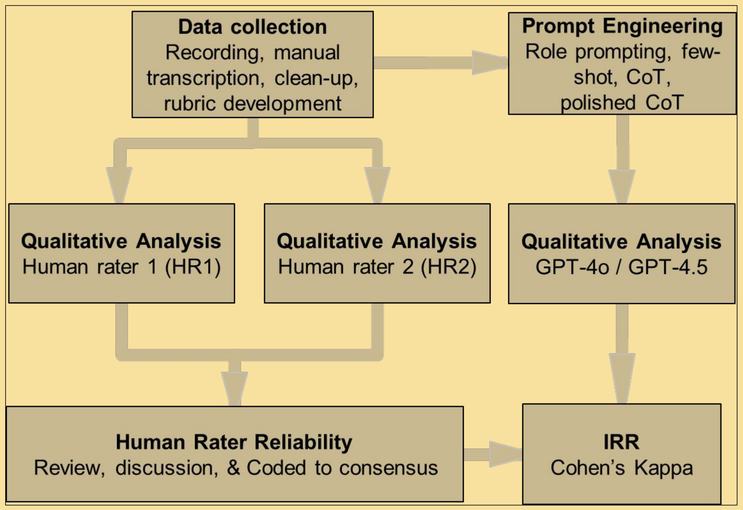

FIG. 1: Methodological Flow: Inter Rater Reliability

RESEARCH QUESTIONS

(i) What is the inter-rater reliability (IRR) between state-of-the-art LLMs such as GPT-4.5 and GPT-4o and expert human raters for coding audio transcripts of students engaged in a group discussion during a lab activity in the context of Engineering Design (ED)?

(ii) To what extent can the IRR of LLMs such as GPT-4.5 and GPT-4o be improved by prompt engineering and optimizing their hyperparameters through the OpenAI API (Application Programming Interface)?

FINDINGS & DISCUSSION

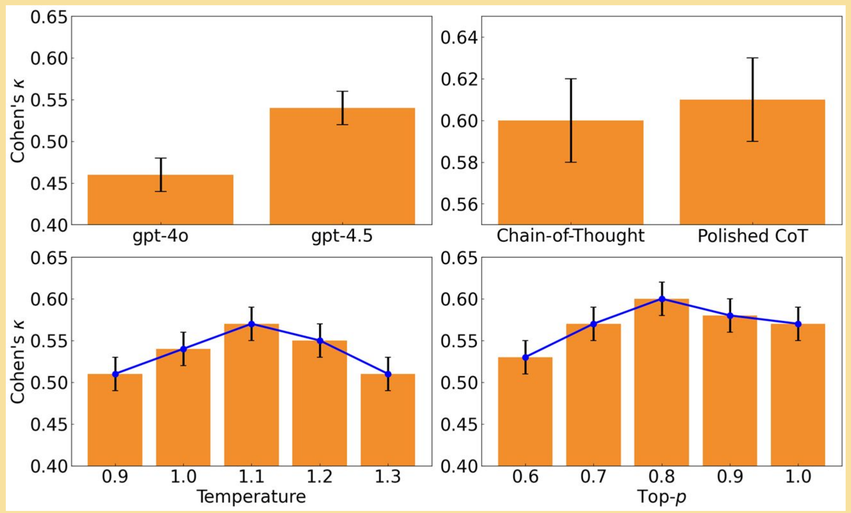

FIG. 2: Optimization of Model Performance for Engineering Design Theme

- Polished few-shot prompts and decomposed coding were key
- GPT-4.5 performed better for **ED,** GPT-4o did better for domain specific themes like PC,MC
- Hyperparameters: **Temperature** and **Top-p**
- **Temperature** controls the variability or randomness of the LLM generated response
- **Top-p** controls the number of most probable words sampled (B. Chen, 2023)

MODEL PERFORMANCE BY THEME AFTER OPTIMIZATION

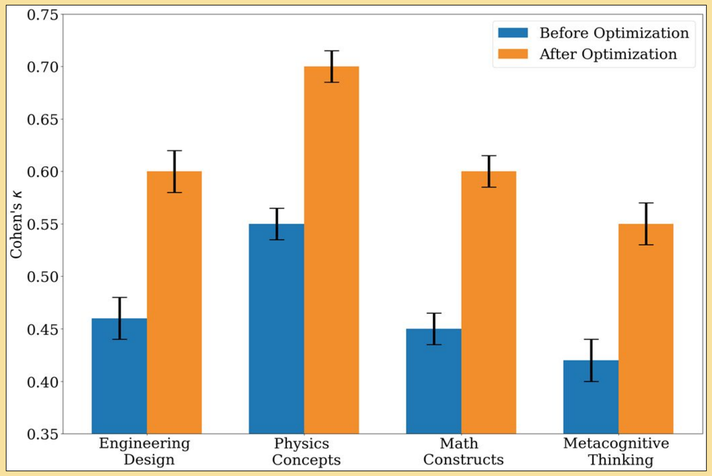

FIG 3: Cohen's Kappa by Theme

- Substantial agreement (κ > 0.6) for ED, PC, and MC with optimized LLMs
- Moderate agreement (κ ≈ 0.55) for MT

CONCLUSIONS & IMPLICATIONS

- LLM coding (K > 0.6) comparable to human raters for domain specific themes (PC, MC)
- Human oversight remains important for MT
- LLMs can scale up qualitative research

LIMITATIONS AND FUTURE WORK

- Study analyzed small subset (14 groups) of large audio dataset, scale up to large datasets
- Risk of overfitting in optimization and generalizability tests required
- Only OpenAI models tested; other LLMs or ML tools like DeepSeek-R1 can be explored

REFERENCES

- H. E. Tinsley and D.J. Weiss, Interrater reliability and agreement of subjective judgements., Journal of Counseling Psychology 22, 358 (1975)
- OpenAI, Gpt-4o technical report, https://openai.com/index/gpt-4o (2024).
- OpenAI, Gpt-4 technical report (2024)
- B. Chen, Z. Zhang, N. Langrené, and S. Zhu, Unleashing the potential of prompt engineering in LLMs: a comprehensive review, arXiv preprint arXiv:2310.14735 (2023).
- R. C. Subramaniam, C. LaFontaine, A. Bralin, J. W. Morphew, C. M. Rebello, and N. S. Rebello, Characterising stem ways of thinking in engineering design (ed)-based tasks, in 2024 PERC Proceedings [Boston, MA, July 10-11, 2024], edited by Q. X.Ryan, A. Pawl, and J. P. Zwolak (2024).

This work is supported in part by U.S. National Science Foundation Grant 2300645