Expanding LLM-Assisted Translation of SQL Queries with Applications in Database Education

Matthew Eisenberg, Computer Science (Software Engineering)
Mentor: Dr. Jia Zou, Assistant Professor
School of Computing and Augmented Intelligence (SCAI)

Introduction

Current LLMs struggle to accurately grade student SQL queries, often failing to account for logically equivalent solutions. This study examines the conditions under which LLMs can most reliably evaluate SQL submissions, comparing their assessments against human grader benchmarks to identify the most effective approach.

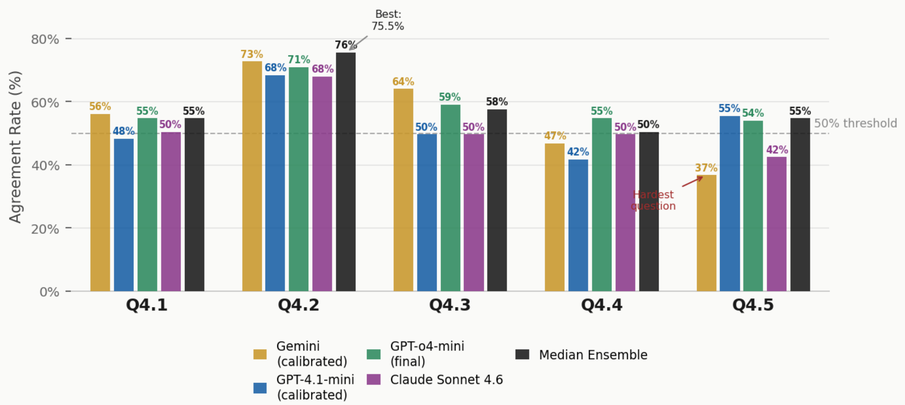
Materials & Methods

Four commercial LLMs were evaluated: Gemini Flash, GPT-4.1-mini, GPT-o4-mini, and Claude Sonnet 4.6. Strategies applied iteratively included: (1) deductions-only rubric formatting, (2) few-shot calibration examples, (3) strictness calibration notes, (4) anti-double-penalization rules, and (5) grade curving. A median ensemble combined scores across all models. All systems were benchmarked against 139 student SQL submissions from a midterm exam.

Results

Iterative prompt engineering significantly improved LLM grading accuracy across all models. The table visual compares each major strategy against the human grading baseline (mean = 43.4, SD = 6.8).

Question-Level Agreement by Model



Grading Strategy Comparison

Strategy	Model	Mean Score	Std Dev	Diff vs. Human
Human (reference)	–	**43.4**	6.8	–
Baseline	Gemini	37.5	10.3	−5.9
Baseline	GPT-4o-mini	23.5	3.8	−19.8
+ Deductions-only rubric	Gemini	35.5	10.7	−7.9
+ Deductions-only rubric	GPT-4.1-mini	34.1	10.9	−9.3
+ Strictness calibration + grade curve	Gemini	42.6	6.9	**−0.8**
+ Strictness calibration + grade curve	GPT-4.1-mini	41.2	7.4	−2.2
+ Strictness calibration + grade curve	**GPT-o4-mini**	45.4	5.0	**+2.0**
Median ensemble (all models)	All models	41.8	7.3	**−1.6**
Claude Sonnet 4.6	Claude Sonnet	39.7	8.0	−3.5
Final (updated prompts)	**GPT-o4-mini**	44.8	5.8	**+1.5**

Human reference | Best result | Mean score bar (below human avg) | Mean score bar (at/above human avg)
Diff vs. Human: blue = within ±2 pts | green = above human | red = large underestimate

Future Work

- Utilize an agentic framework for further grading improvement
- Combine grading pipeline with LLM-generated practice problems

Conclusion

Prompt engineering, particularly strictness calibration and anti-double-penalization rules, reduced Gemini's mean grade difference from −7.9 to −0.8 points. GPT-o4-mini achieved the highest overall agreement at 27.0%. The median ensemble provided the most consistent question-level performance. These results demonstrate LLMs can serve as viable grading supplements when carefully calibrated.

FURI

Acknowledgments

I would like to thank Dr. Jia Zou for her invaluable guidance and Dr. Chris Bryan for his advice during this research. I'm very grateful to ASU's FURI program as well.

Ira A. Fulton Schools of Engineering
Arizona State University