Can Large Language Models Assist in the Evaluation of Pharmacy Accreditation Standards?
A Preliminary Study on Standard 18

Mehmet Arun[1], Vedat Eğilmez[2], Zerrin Sezgin Bayındır[3], **Sibel Suzen**[3], **Aysun Pabuccuoğlu**[1], Seçkin Özden[4]

[1] Ege University Faculty of Pharmacy, Izmir, Türkiye, [2] Pharmaceutical Industry Employers' Union, İstanbul, Türkiye, [3] Ankara University Faculty of Pharmacy, Ankara, Türkiye,
[4] Lokman Hekim University, Faculty of Pharmacy, Ankara, Türkiye

Background

ECZAKDER (Association for Evaluation and Accreditation of Pharmacy Education Programs) is Turkey's accreditation body for pharmacy education, established in 2014. The current accreditation standards (Version 6.0, January 2024) comprise 20 standards organized under 7 main domains:

Domain	Standarts
1. Mission & Objectives	S.1 - S.3
2. Organization & Governance	S.4 - S.6
3. Curriculum	S.7 - S.12
4. Student Standards	S.13- S.15
5. Academic Staff	S.16 - S.17
6. **Facilities & Learning Resources**	**S.18** - S.19
7. Financial Resources	S.20

Each faculty undergoing accreditation prepares a self-evaluation report (ÖDR) demonstrating compliance with all standards. An evaluation committee (DİZE) of three members independently reviews the report, writes justifications, and scores each standard on a 1–5 scale. The process requires at least 10 semesters (300 ECTS credits) of pharmacy education.

This labor-intensive process requires domain expertise. This preliminary study explored whether large language models (LLMs) can assist evaluators in assessing Standard 18 (Learning Resources), a standard with a relatively structured evaluation framework.

Why Standard 18?

Among the 20 standards, S.18 has clearly defined, observable sub-criteria (infrastructure, technology, feedback mechanisms) making it suitable for a pilot AI-assisted evaluation study. Standards involving complex qualitative judgments. (e.g., S.7 Curriculum, S.10 Teaching Processes) would require significantly more contextual understanding.

S.18.1 Library facilities, information technologies, internet access, and distance education technologies shall be of adequate quantity and quality.

S.18.2 Remote access technologies facilitating off-campus library use shall be available, and students shall be informed and trained on these resources.

S.18.3 Feedback from students and academic staff on the adequacy of learning resources shall be continuously collected, and necessary improvements shall be implemented.

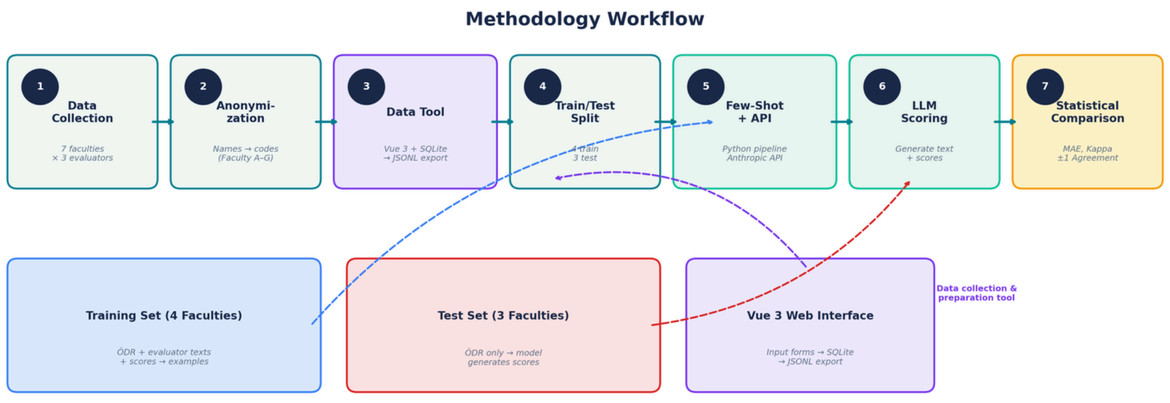
Methodology

Self-evaluation texts for Standard 18 from seven pharmacy faculties were compiled along with justification texts and scores from three evaluators per faculty. All data were anonymized. Four faculties served as the training set and three as the test set. A Python-based pipeline was developed to automate the workflow: extracting and anonymizing report data, constructing few-shot prompts with training examples, and calling the Anthropic API to generate justifications and scores for test faculties.

Three model configurations were tested:

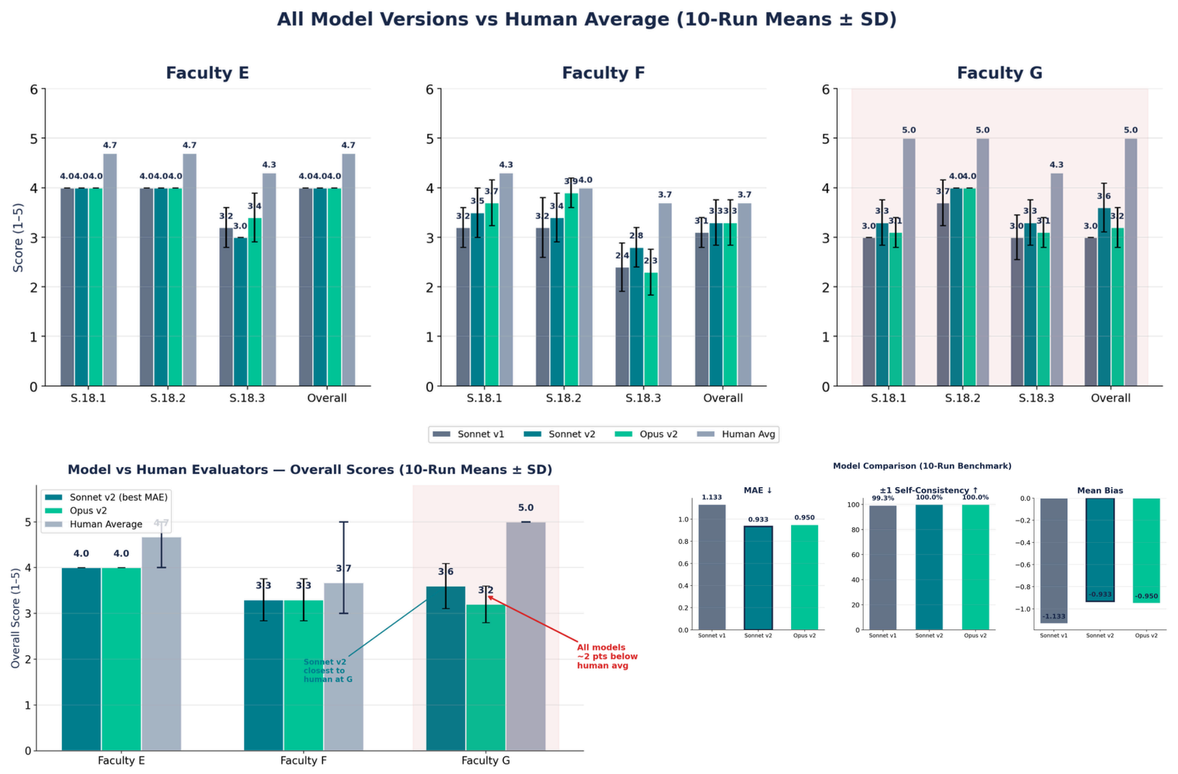
- Claude Sonnet 4 — original prompt (sonnet v1)
- Claude Sonnet 4 — calibrated prompt (sonnet v2)
- Claude Opus 2 — calibrated prompt (opus v2)

Model outputs were compared against human scores using MAE, agreement rates, and weighted Cohen's Kappa.



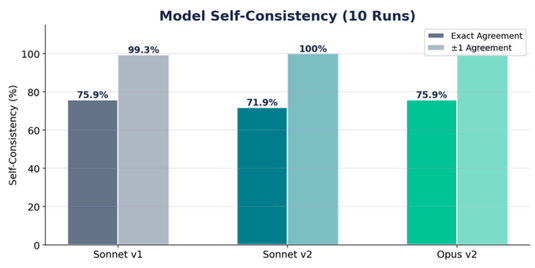
Results

Based on 10-run benchmark averages, Claude Sonnet 4 with the calibrated prompt achieved the best overall performance (MAE = 0.93, ±1 self-consistency = 100%), particularly excelling at Faculty G where it scored closest to human evaluators (3.6 vs 5.0). Claude Opus 4 with the calibrated prompt also performed well (MAE = 0.95, ±1 self-consistency = 100%), showing the strongest alignment with human scores on S.18.1 and S.18.2 sub-standards. Calibrated prompts and more capable models both contributed to better alignment with human evaluators.



The self-evaluation report for Faculty G mentioned that a computer laboratory had not yet been established. All three model versions interpreted this as a critical deficiency, scoring 2 points below the human average.

Human evaluators, however, took a holistic view, considering the overall learning resource adequacy rather than penalizing a single unfinished facility. This reveals a key limitation: LLMs are highly sensitive to explicitly stated deficiencies and struggle with holistic assessment.



Each model configuration was run 10 times with identical inputs to measure output variability.

Key findings:

- **Structured justifications:** LLMs produced coherent, criteria-aligned evaluation justifications comparable to human evaluator texts
- **Conservative scoring:** All models systematically scored lower than human average (mean bias −0.86 to −1.11)
- **Deficiency sensitivity:** Explicitly mentioned shortcomings were heavily penalized, even when human evaluators took holistic views
- **Strict baseline rater:** Model scores aligned most closely with the most conservative human evaluator

Conclusion

LLMs show promise as assistive tools in accreditation, not replacing human judgment, but supporting it with consistent, structured preliminary assessments.

The conservative bias may be a feature, an LLM that identifies potential weaknesses can serve as a valuable pre-screening tool.

Future Directions

- **Expanding dataset:** We are planing to include more faculties and more standards.
- **Optimizing prompts:** Reduce conservative bias through better calibration.
- **Evaluator assistant:** LLM as pre-screening tool to reduce workload
- **ÖDR feedback tool:** Help faculties identify report weaknesses before submission