LLM as a Supreme Court Judge

Simulating Legal Expert Reasoning in Complex Judicial Scenarios

Mandira Sawkar {ms7201@rit.edu}
MSAI, Rochester Institute of Technology

Introduction

Problem Statement:

- C-o-T reasoning in complex, nuanced situations like Supreme Court debates is underexplored.

Motivation:

- Legal decisions require nuanced deliberation, logical reasoning, and ethical considerations.
- Complex multi-turn debates remain untested with LLMs.

Research Objective: Can LLMs:

- Simulate multi-turn legal debates?
- Emulate reasoning patterns of real legal experts?
- Reach accurate verdicts compared to human decisions?
- Faithfully simulate deliberations using multi-turn dialogue and personas?

Dataset:

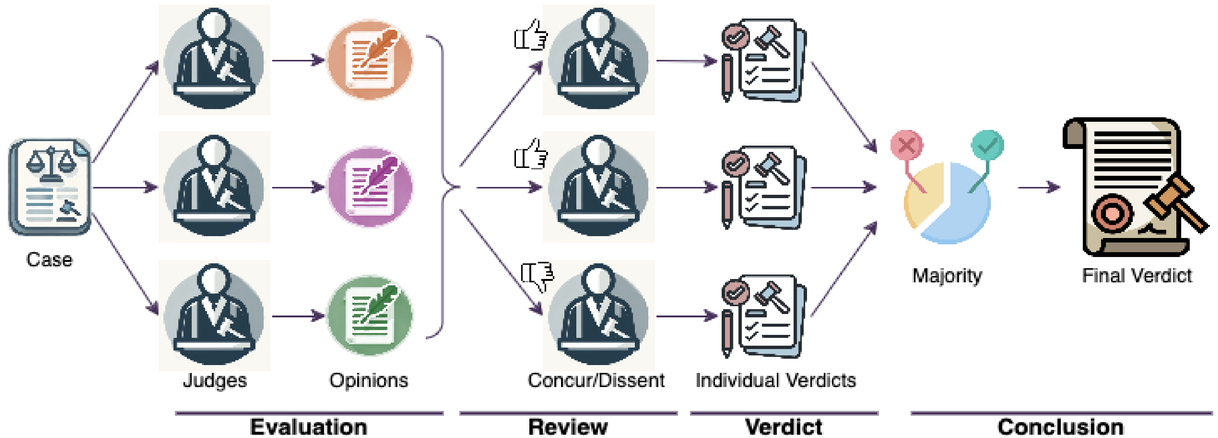
- JUSTICE [1] with Supreme Court case data, with facts, judges, votes, decision, etc.
- Filtered to retain 200 cases after preprocessing (cleaning, sampling).

Methodology

Experiments:

1. Variable number of judges (1 & 9) without personas.
2. Simultaneous and Sequential prompting
3. Generic judges without ideologies compared to specific SC judges with ideologies.
4. Different models used to measure capabilities.

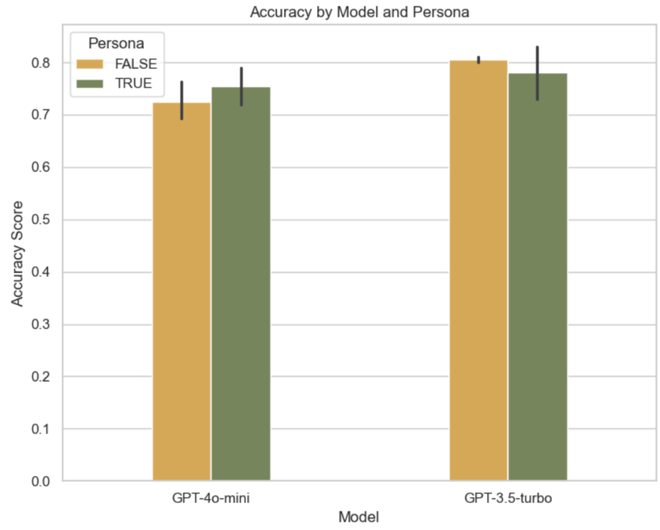
Methodology



Results

- Personas improve reasoning consistency.
- Models like GPT-4-o-mini outperform others in complex settings.
- Larger judge panels improve verdict stability.
- LLMs capture some nuances of judicial reasoning.
- Sequential debates simulate real-world deliberations effectively.

MODEL	PERSONA	JUDGES (#)	ACCURACY	F1 SCORE
PARALLEL STYLE				
GPT-4o-mini	FALSE	1	0.71	0.42
GPT-4o-mini	FALSE	9	0.78	0.44
GPT-3.5-turbo	FALSE	9	0.8	0.44
GPT-4o-mini	TRUE	6-9	0.72	0.42
GPT-3.5-turbo	TRUE	6-9	0.79	0.44
SEQUENTIAL STYLE				
GPT-4o-mini	FALSE	1	0.68	0.4
GPT-4o-mini	FALSE	9	0.73	0.42
GPT-3.5-turbo	FALSE	9	0.81	0.45
GPT-4o-mini	TRUE	6-9	0.79	0.44
GPT-3.5-turbo	TRUE	6-9	0.83	0.45



Ethics & Future Scope

Ethical Considerations:

- The setup must always be used with humans in the loop.
- It is to be used as a tool for knowledge and debate enhancement rather than decision-making.

Scope:

- Dataset Expansion: Include a broader range of cases and jurisdictions.
- Models: Evaluate emerging reasoning-optimized LLMs.
- Additional Metrics: Cosine or embedding similarity between human and AI verdicts.
- Practical Applications: Deploy in lower courts or for juries to understand legal precedents, procedures and deliberation.

Conclusion

- LLMs show potential in simulating judicial deliberations and their reliability in high-stakes decision-making.
- Personas and multi-turn dialogue improve verdict alignment and consistency.
- Judge count impacts majority stability.

References

- Alali et al. *JUSTICE: A Benchmark Dataset for Supreme Court's Judgment Prediction*. 2021.
- Zheng et al. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. 2023.
- Sulea et al. *Predicting the Law Area and Decisions of French Supreme Court Cases*. 2017.