

Using Differences in State Law to Test Whether LLMs Reason or Remember

Colin Doyle

Loyola Law School, Los Angeles

Aaron Tucker

FAR.AI

Frontier models score well on legal benchmarks.

**But is that evidence of a legal reasoning capability
or pattern-matching?**

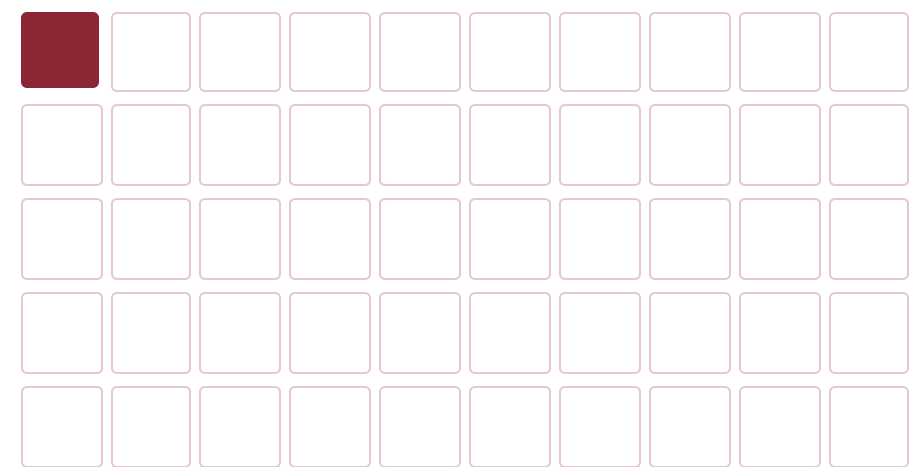
Differences in state law provide a natural experiment.

MAJORITY RULE



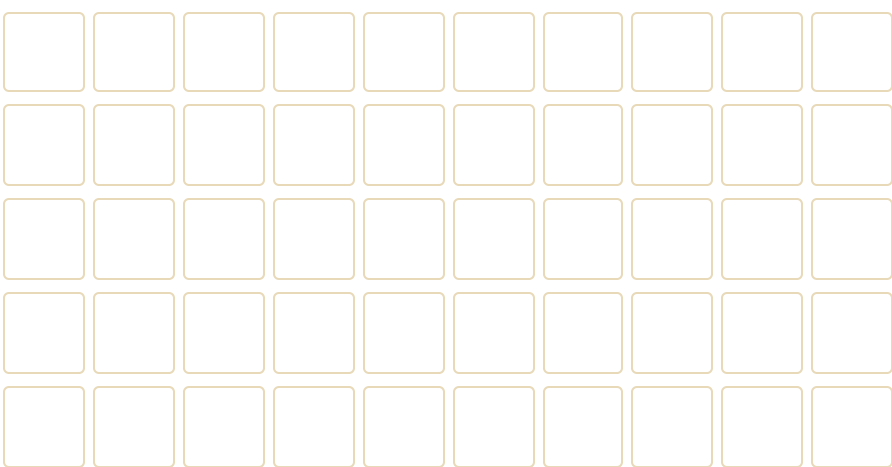
Heavily represented in training data

MINORITY RULE



Less represented in training data

HYPOTHETICAL RULE



Least represented in training data

Three legal domains

Damages

Hearsay

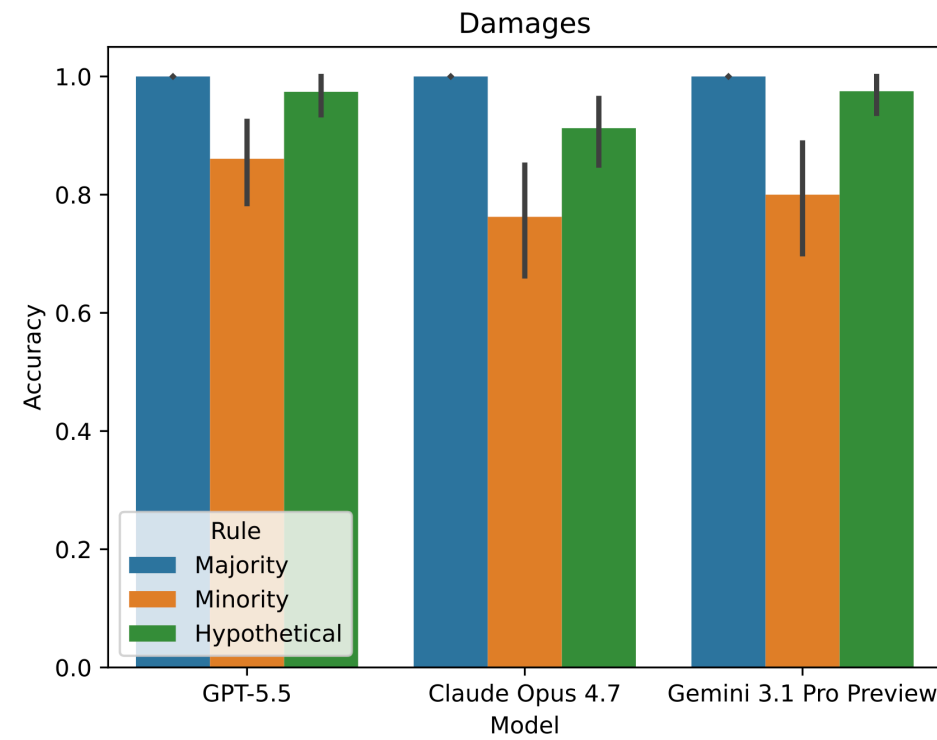
Noncompete agreements

LLM Query Structure

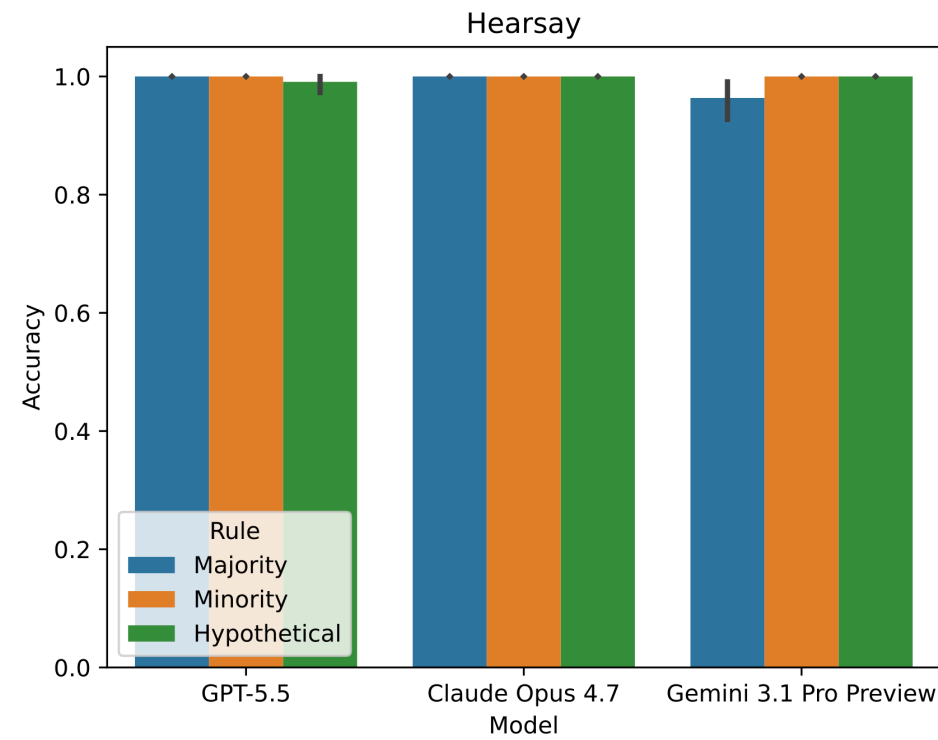


Results

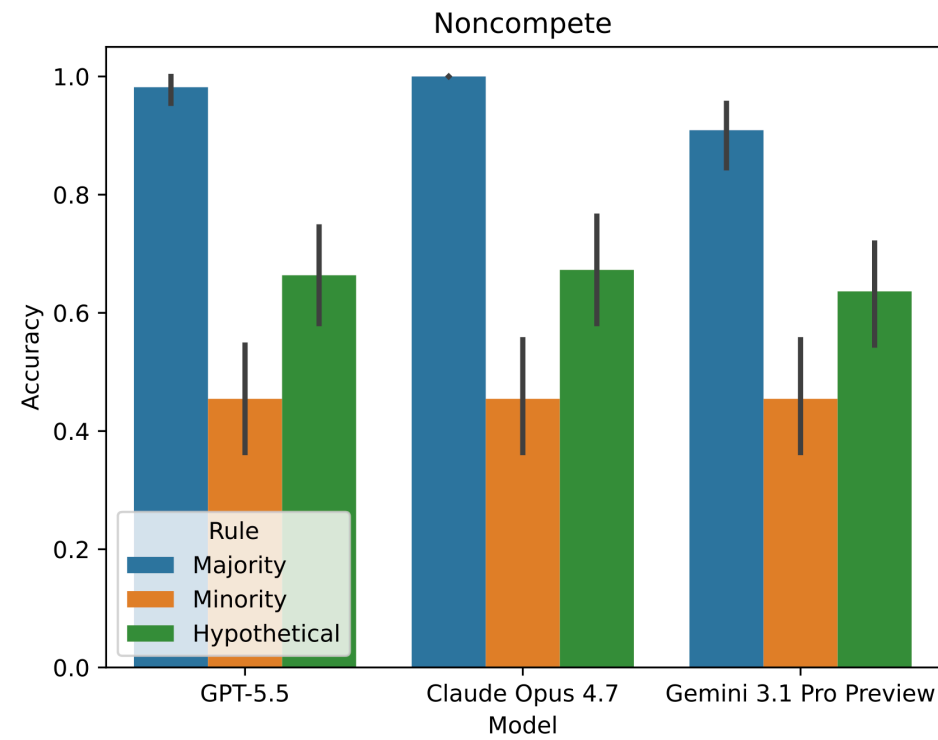
Damages



Hearsay



Noncompetes



Takeaways & Future Work

- Benchmarks built on common legal materials likely overstate legal reasoning performance.
- Hypothetical rules are resistant to training-data contamination and reflect how law students are tested.
- Variations in law across jurisdictions may provide opportunities for other natural experiments.

Thank you

Colin Doyle
colin.doyle@lls.edu

Aaron Tucker
aaron@far.ai