Law and Large Language Models

Noah Budnitz (Wharton, 2026) and Rishabh Mandayam (Wharton & Engineering, 2025)
Mentored by Professor David Abrams

PURM
Penn Undergraduate Research Mentoring Program

Introduction

Large language models like ChatGPT present numerous opportunities in the realm of Law & Economics as a tool to assist in research. Specifically, these AI models have particular utility in their potential to automate hand-coding tasks for large datasets.

In this project, we explore the feasibility of utilizing generalized large language models for the automation of classification, named entity recognition, and advanced language processing tasks in law-specific research. We do so by comparing the output of OpenAI's GPT 3.5 and GPT 4 models in comparison to that of legal experts in analyzing aspects of Supreme Court cases.

Methods

The Supreme Court Database (SCDB) is a multi-decade project analyzing every Supreme Court case since 1950. Each case is assigned more than 50 expertly-coded classifications and language variables. As the variables were determined by an evolving team of legal experts, they served as the a "gold standard" for our comparison and a benchmark for how well the language model performed in comparison.

Over 28,000 cases and orders from the Supreme Court were part of our analysis, the complete pdfs of which were queried from the open access project "case.law." Minor processing was done to turn these into raw text, which was then matched with its corresponding variable set from the SCDB for later comparison.

Methods

For each SCDB variable of interest, questions were created based on the set of guidelines in the SCDB casebook. These questions were tuned to match the SCDB's style of notation and answer qualifications. The full text of the relevant SCOTUS text was then fed to the GPT 3.5 API with the previously crafted question.

CaseNum	Start Date	End Date	Judge	Prosecutor
0001	01/03/2020	03/07/2020	Justice A	Prosecutor A
0002	02/07/2021	08/10/2021	Justice A	Prosecutor B
0003	08/13/2022	09/22/2022	Justice B	Prosecutor B

Figure 1: Example output from model

The SCDB dataset codes for variables numerically (i.e. 1=Overturned, 2=Upheld). As such, the GPT 3.5 query included instructions to only return a single number for each question. The numbers could then be converted back into human readable language afterwards.

We iterated thrice on the method of input/prompt to the GPT 3.5 API. The second revision focused on ensuring the model only returned what was asked of it, by providing within the prompt more explicit exclusion instructions and sample responses.

The third iteration utilized more advanced pre-processing techniques for the full case text. Instead of passing the entire case to the LLM, we generated snippets of relevant text around keywords selected for each variable and passed this set to the model instead.

Results

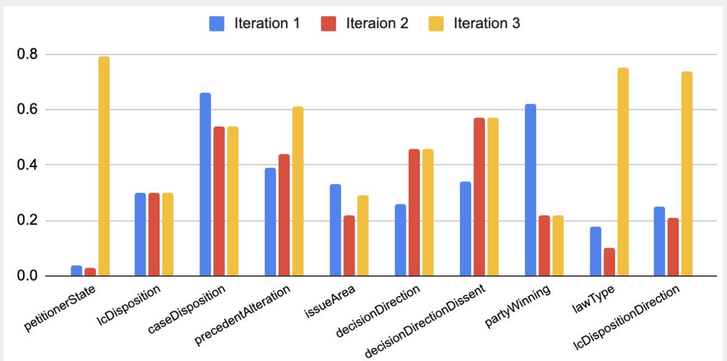

Figure: Accuracies of model iterations

Discussion & Future Research

Our project iterations demonstrated ways by which general LLM models can be improved for law research tasks. However, even with this tuning, we were unable to bring accuracy levels to parity with that of a human reviewer. This is likely due to selection of model (GPT 3.5) and with the inherent limitations of a general, as opposed to specifically trained), model, presenting plenty of opportunities open for future research.

Citations

Caselaw Access Project. (2018). Retrieved 2023, from case.law.

Harold J. Spaeth, Lee Epstein, et al. 2022 Supreme Court Database, Version 2022 Release 1. URL: http://Supremecourtdatabase.org

Funding provided by Penn Undergraduate Research Mentoring Fund (PURM)
Aspects of project were developed in collaboration with a team at Lehigh led by Professor Brian Chen