Court Sanctions

and

Court Rules

Empirical Design

Types of studies

- Experiment: you control the intervention
- Observational study: you watch what happens “in the wild”
- Case study: deep dive into examples
- Demonstration: cool story, bro (but not necessarily generalizable)

Many AI “benchmarks” and “studies” are more like demonstrations than experiments even though they use the language of experiments.

What is an experiment?

An experiment is a structured way to learn whether a change in one thing causes a change in another.

Key move: isolate the causal story.

The core vocabulary

- Research question
- Hypothesis
- Independent variable (what you change)
- Dependent variable (what you measure)
- Controls (what you hold constant)
- Confounders (sneaky alternative explanations)

Research question

We want to test language models in some way related to a legal task.

Example research questions:

- Do judges apply a new Supreme Court standard differently depending on ideology?
- How does providing counsel for people in family court litigation affect case outcomes?
- What is the effect of adopting pretrial risk assessments on pretrial incarceration?

Let’s workshop a research question together

- Can AI be used to eliminate bias in legal decisions (or how much is bias baked in)?
- How will generative AI video and photo affect video evidence in court?
- Can AI predict the success of a cause of action or claim?
- To what degree does sanction severity deter future misconduct?
- Does AI reduce or reinforce inequality of outcomes for pro se litigants?

Hypotheses

Does AI reduce or reinforce inequality of outcomes for pro se litigants?

A hypothesis is a prediction. Must be capable of being disproven.

Potential hypotheses:
Providing pro se litigants access to AI results in fewer dismissals.
Providing pro se litigants access to AI results in more cases being pursued.
Providing pro se litigants access to AI results in more sanctions.

Independent variable

What do we change in this experiment?

Hypothesis: Providing pro se litigants access to AI results in fewer dismissals.

One group has access to AI and one does not.

Dependent variable

What do we measure?

Rate of dismissal between the two groups.

Controls

What do we try to hold constant and how?

Type of case

Type of AI.

Timing.

Confounders

What might be sneaky alternative explanations for our results?

Template for using LLMs for experiments

Research question:

Hypothesis:

Independent variable (what we change):

Dependent variable(s) (outcomes measured):

Task definition:

Test set:

Prompting plan:

Rubric:

Analysis plan:

Limitations:

Reliability and validity

Reliability: are we measuring consistently?

Validity: are we measuring what we think we're measuring?

Statistical thinking

Basic questions:

- Is there a difference?
- How big is it?
- Can we generalize from it?