LLMs in Economics Research

Measurement error, rectification, and external validity

Lecture 7

Observations and predictions are not the same

LLM outputs can be impressively close to real observations.

But an observation is a realized outcome in the world, while an LLM output is a model-based prediction generated from training data, prompts, and decoding choices.

- observations are realized data
- LLM outputs are conditional proxies
- closeness does not remove measurement error
- using predictions still needs validation and scope checks

Roadmap

Objects and uncertainty

LLMs as measurement tools

Simulated agents and survey synthesis

Research protocol

Objects and uncertainty

Three empirical tasks

Task	Core question
Prediction	can the model forecast a realized outcome?
Measurement	can the model proxy a defended human label y_i from text?
Simulation	can the model approximate a person's survey response?

- Ludwig, Mullainathan, and Rambachan (2025) focus on measurement
- Horton (2023) and Krsteski et al. (2025) focus on simulation

Common response notation

For units $i = 1, \ldots, N$, let

$$x_i = \text{available inputs}$$

where x_i can include text, demographics, survey history, or question context.

Let

$$y_i = \text{human target on a common scale}$$

and

$$\hat{y}_i = m(x_i; \theta, c, \tau)$$

for the synthetic response generated by the LLM. Here m is the end-to-end prompting pipeline, θ are model weights, c is the prompt context, and τ is the decoding rule.

In a text-labeling task, y_i is the gold-standard human code. In a survey-simulation task, y_i is the human survey response or its mapped score

Human target, held-out human sample, unlabeled frame

- **Human target** y_i: the defended gold-standard response or label
- **Held-out human sample** H: units for which both y_i and $\hat{y}_i$ are observed
- **Unlabeled frame** U: units for which we have inputs x_i and synthetic responses $\hat{y}_i$, but not new human responses

Krsteski et al. (2025) analyze both open-ended and multiple-choice responses by mapping them to a common scale before estimation. For labeling tasks, the same move lets us treat a human code as the target response

$$y_i \in \mathcal{Y}$$

where $\mathcal{Y}$ might be binary, Likert, or a numeric scale

How prompting works technically

Earlier we wrote

$$\hat{y}_i = m(x_i; \theta, c, \tau).$$

This shorthand hides three steps.

- the weights θ determine token probabilities conditional on the task input x_i and prompt context c
- the decoding rule τ turns those probabilities into a concrete answer
- the researcher then parses or scores that answer to obtain $\hat{y}_i$
- prompting changes c, but it does not update θ

Zero-shot, one-shot, and few-shot prompting

- **Zero-shot**: give task instructions, but no worked example
- **One-shot**: give one worked example in the prompt
- **Few-shot**: give several worked examples in the prompt

In all three cases, the examples enter through the prompt context c rather than through the model weights θ

Prompting versus fine-tuning

- **Prompting / in-context learning**
  - adaptation comes from tokens placed in the context window
  - it is limited by the model's maximum context size
  - that limits how many examples we can include
  - it often requires careful prompt engineering
  - it does not create a standalone reusable task model
- **Fine-tuning**
  - use labeled data to update model parameters for a new task
  - task information moves into the weights rather than the prompt
  - this can produce a reusable model, but it needs training data and held-out evaluation

Sources of statistical uncertainty

- **Sampling uncertainty**: finite held-out human samples and finite primary samples
- **Target uncertainty**: coder disagreement or respondent instability in the human target y_i
- **Configuration uncertainty**: lack of knowledge about the right model, prompt, persona, or fine-tuning setup
- **Decoding uncertainty**: repeated stochastic generations can differ even with the same (x_i, θ, c)

These remain even when the research design is conceptually valid

Design failures are not another variance term

- **Training leakage or contamination**: the model has effectively seen the evaluation texts already
- **Construct mismatch**: the defended human target y_i or the prompt operationalizes a different construct from the one claimed in the paper
- **Held-out sample mismatch**: the held-out human sample H comes from a different population, period, or prompt regime, so the estimated bias relation does not match the main sample
- **Evaluation reuse**: the same human labels are used for prompt design or fine-tuning and then reused as if they were fresh held-out labels, which makes both performance and correction look too optimistic

These are assumption failures. No standard error fixes them

LLMs as measurement tools

From synthetic responses to estimator bias

This block follows Ludwig, Mullainathan, and Rambachan (2025). We now focus on the measurement task. Define

$$\delta_i = \hat{y}_i - y_i$$

as the synthetic-minus-human error.

On the held-out human sample H, the researcher observes both y_i and $\hat{y}_i$.

The economic problem is not whether $\hat{y}_i$ looks plausible. It is whether replacing y_i with $\hat{y}_i$ changes the probability limit of the estimator. For concreteness, the next slides use one linear proxy-as-outcome example

Running example: hawkishness from central-bank text

- each record i is a paragraph from a central-bank statement
- $y_i = 1$ if the defended coding rule says the paragraph is hawkish
- y_i comes from trained human coders applying that rule
- $\hat{y}_i$ comes from a prompted LLM asked for the same label
- the downstream regression asks whether hawkish language predicts yield surprises

This is the concrete object behind the next three slides

Special case: synthetic response as outcome

To keep the notation light, let z_i be one regressor, such as a crisis indicator.

Suppose the target regression is

$$y_i = \alpha + \beta z_i + \varepsilon_i, \qquad \mathbb{E}[z_i \varepsilon_i] = 0.$$

If we regress $\hat{y}_i$ on z_i instead, then

$$\hat{y}_i = \alpha + \beta z_i + \varepsilon_i + \delta_i$$

so

$$\operatorname{plim} \hat{\beta}^{plug} = \beta + \frac{\mathbb{E}[z_i \delta_i]}{\mathbb{E}[z_i^2]}.$$

The extra term is the bias. It is large when label errors move systematically with the regressor

Toy example: high accuracy, biased coefficient

- in calm periods, hawkish language is common and easy to classify
- in crisis periods, hawkish language is rarer and phrased differently
- the LLM then under-calls hawkish text exactly when the crisis indicator is high
- headline accuracy can still look excellent because most paragraphs are easy negatives
- the coefficient is biased because δ_i moves with the crisis indicator

What matters is the error structure

Ludwig, Mullainathan, and Rambachan (2025) push a very economist point

- accuracy averages mistakes across observations
- inference depends on how mistakes covary with economically relevant covariates
- rare but policy-relevant cases can dominate coefficient bias
- two prompts can have similar headline accuracy and radically different downstream regression estimates

The right object is not average agreement alone. It is the full error structure of δ_i

Prediction problems have a different contract

For prediction, Ludwig, Mullainathan, and Rambachan (2025) emphasize training leakage instead

- a model can look miraculous on held-out researcher data because the text was already in its pretraining corpus
- this is especially dangerous for closed APIs with unknown data provenance
- therefore prediction tasks require no leakage between the model's training data and the research sample
- the practical recommendation is open, time-stamped models for serious forecasting exercises

Using a held-out sample to estimate coefficient bias

For the linear regression example on the previous slides, let the held-out human sample H reveal both y_i and $\hat{y}_i$. Estimate

$$\hat{y}_i - y_i = \gamma_0 + \gamma_1 z_i + u_i$$

on that sample, then correct the plug-in estimator by

$$\hat{\beta}^{db} = \hat{\beta}^{plug} - \hat{\gamma}_1.$$

In this linear special case, Ludwig, Mullainathan, and Rambachan (2025) show that the correction restores consistency and correct coverage under the held-out-sample assumptions

Running example: what the held-out sample buys you

- relabel a subset of paragraphs with the defended human protocol
- estimate whether the LLM misses hawkish language more often in crisis periods
- use that estimated error-covariate relation to adjust the plug-in coefficient
- the principle is general, but the exact correction formula changes with the downstream estimand

What must the held-out sample identify?

The correction works only if the held-out human sample is informative about the error in the main sample

- same construct definition
- same prompt and model configuration
- same relation between δ_i and z_i in the held-out and primary samples
- held-out labels must be genuinely held out, not reused from prompt tuning or fine-tuning

This is why “gold standard” does not mean “humans are perfect”. It means “this is the measurement rule we are willing to defend”

The real value of the human labels

A held-out human sample does two different jobs

1. it measures raw label quality
2. it identifies the bias correction needed for the downstream estimator

Ludwig, Mullainathan, and Rambachan (2025) argue that researchers often focus too much on the first and underappreciate the second

That is why even a small held-out human sample can be more valuable than a large amount of prompt tweaking

Simulated agents and survey synthesis

What does Horton actually show?

- Horton is not mainly showing that LLMs have deep preferences
- in a simulation task, the model produces synthetic responses $\hat{y}_i$ for respondent profiles and scenarios
- these synthetic responses can approximate some observed human regularities
- it can therefore be useful for piloting experiments or generating hypotheses
- but matching responses is weaker than identifying the preferences or decision rules behind human behavior

Source: Horton (2023), read here as evidence about conditional response fit rather than human-like preferences

Matching responses is weaker than matching preferences

Let s denote an economic scenario and define

$$\mu_H(s) = \mathbb{E}[y_i \mid s], \qquad \mu_L(s) = \mathbb{E}[\hat{y}_i \mid s].$$

If

$$\mu_L(s) \approx \mu_H(s)$$

on observed scenarios, the model matches average response patterns.

That still does not show that the model has the same utility, beliefs, or counterfactual response to a new scenario s'

Novel shocks are an off-support problem

Let $\mathscr{S}$ denote the set of scenarios on which we have observed human responses and assessed model fit

If a new policy or shock s' lies outside $\mathscr{S}$, then simulation requires extrapolation rather than interpolation

- that is an external-validity and extrapolation problem, not only a prediction problem
- the required assumption is now cross-environment stability
- this assumption is usually untestable without fresh human data

External validity: The ability of a mapping learned in one environment to transport to a new population, institution, or shock.

Special case: rectification for a survey mean

If raw survey answers are not already numeric, Krsteski et al. (2025) first use a coding map $\phi(\cdot)$ to turn each answer into the scalar target y_i. For that scalar target, they write

$$\theta^* = \frac{1}{|U|} \sum_{i \in U} y_i$$

and estimate it with

$$\hat{\theta}_\lambda = \frac{1}{|U|} \sum_{i \in U} \hat{y}_i + \lambda \left(\frac{1}{|H|} \sum_{i \in H} y_i - \frac{1}{|H|} \sum_{i \in H} \hat{y}_i \right).$$

The first term uses the large synthetic frame. The second term adds the estimated human-minus-synthetic gap from the held-out human sample

Toy example: what the rectification weight does

- suppose the synthetic frame predicts 62% support for a policy
- in the held-out human sample H, humans average 55% while the model averages 60%
- the estimated human-minus-model gap is therefore minus 5 percentage points
- $\lambda = 1$ subtracts the full gap from the synthetic estimate
- a tuned λ shrinks or enlarges that adjustment depending on proxy quality and finite-sample noise

What Krsteski et al. (2025) establish

Krsteski et al. (2025) turn this idea into a concrete survey-design result

- synthesis alone introduces substantial bias, 24 to 86% in their two surveys
- with rectification, bias falls below 5%
- effective sample size gains reach about 14% in their settings
- under a fixed human-data budget, spending most labels on a held-out correction sample can beat spending them all on fine-tuning

Source: Krsteski et al. (2025), abstract

Three levers in Krsteski et al. (2025)

- **Persona-guided prompting**: condition on respondent backstories or observable attributes
- **Supervised fine-tuning**: adapt model weights to a labeled human-response dataset
- **Rectification**: use held-out human responses to correct the final synthetic estimator
- empirical lesson: upstream improvements help, but valid estimation still depends on the held-out correction stage

A fixed human-label budget creates a real trade-off

Suppose the human budget is

$$n = n_{FT} + n_H, \qquad n_H = |H|.$$

- increasing n_{FT} may improve the proxy model
- increasing n_H improves identification of the correction term
- if the proxy is already fairly correlated with truth, the marginal return to more held-out labels can exceed the return to more fine-tuning labels

Their experiments often favor allocating a majority of the human budget to the held-out correction sample rather than spending it all upstream

Research protocol

A more technical workflow for economists

1. Define the estimand before prompting the model
2. Separate prediction tasks from measurement tasks from simulation tasks
3. Treat prompt and model choice as part of the research design
4. Hold out a genuine human sample H for any estimation use
5. Report sensitivity across prompts, models, and subgroup slices
6. Treat new-shock exercises as external-validity problems, not automatic evidence

Summary

- LLMs are useful for economists when we are explicit about the target object and the uncertainty around it
- Ludwig, Mullainathan, and Rambachan (2025) show that plug-in labels create nonclassical measurement error in downstream estimation
- Held-out human data identify the downstream correction, not only headline accuracy
- Horton (2023) is most useful here as a conditional-response model with earned external validity
- Krsteski et al. (2025) show that rectification can recover valid population estimates and meaningful ESS gains
- The right habit is to decompose uncertainty rather than hide it behind one final standard error