Economics with LLMs: Measuring, Forecasting, and Simulating Economic Behavior

Day #3: Forecasting, Prediction, and Survey Augmentation

Sophia Kazinnik

Digital Economy Lab
Stanford University

August 19, 2026

Class Outline

- Forecasting:
  - *Simulating the Survey of Professional Forecasters* with Anne L. Hansen, John J. Horton, Daniela Puzzello, and Ali Zarifhonarvar
- Econometric contract (estimation and prediction tasks)

Why Forecasting?

Forecasting is one of the most natural things to ask an LLM to do.

- Give it economic data and ask about inflation
- Give it a company filing and ask about earnings
- Give it a political event and ask what happens next

Simple?

Two Things

At a basic level, a forecaster has:

- **An information set:** what the forecaster knows at the time of prediction
- **A loss function:** how forecast errors are scored

Forecasting is a disciplined prediction task, not just an opinion about the future.

Why LLM Forecasting Is Complicated

LLMs make both parts of forecasting harder.

Information set

- The model was trained on a huge and partly hidden corpus
- The training cutoff may not be enough to know what it has seen
- Historical forecasting can accidentally become memory, not prediction

Uncertainty

- If we ask for one number, we get a point forecast
- But we also need to know how uncertain that forecast is
- Different mistakes may have different costs

Turning an LLM into a Forecaster

The goal is not just to ask:

What will inflation be next year?

The goal is to make the model produce something we can evaluate.

- Ask for a **distribution** of possible outcomes
- Record the **information set** available at forecast time
- **Compare** against experts, benchmarks, or simple statistical models

Example: Professional Forecasters

Simulating the Survey of Professional Forecasters[1]

Anne Lundgaard Hansen[a], John J. Horton[b], Sophia Kazinnik[c],
Daniela Puzzello[d], and Ali Zarifhonarvar[d]

[a]*Federal Reserve Bank of Richmond*
[b]*MIT Sloan School of Management*
[c]*Stanford HAI*
[d]*Indiana University Bloomington*

First Draft: April 2024
This Draft: February 2025

We simulate economic forecasts of professional forecasters using large language models (LLMs). We construct synthetic forecaster personas using a unique hand-gathered dataset of participant characteristics from the Survey of Professional Forecasters. These personas are then provided with real-time macroeconomic data to generate simulated responses to the SPF survey. Our results show that LLM-generated predictions are similar to human forecasts, but often achieve superior accuracy, particularly at medium- and long-term horizons. We argue that this advantage arises from LLMs' ability to extract latent information encoded in past human forecasts while avoiding systematic biases and noise. Our framework offers a cost-effective, high-frequency alternative that complements traditional survey methods by leveraging both human expertise and AI precision.

Keywords: Large Language Models; Survey of Professional Forecasters; Behavioral Finance; Generative Artificial Intelligence; Simulated Economic Agents.

LLMs as Approximations of Humans

Growing body of literature shows that LLMs produce responses consistent with both economic theory and documented patterns of human behavior:

- behavioral econ experiments (Horton; 2023)
- consumer choice surveys (Brand, Israeli, and Ngwe; 2023)
- surveys on political biases (Argyle et al.; 2023)

Additionally, LLMs:

- can align with their Big Five assigned personality profiles (Jiang, Zhang, Cao, and Kabbara; 2023)
- and exhibit personality consistency (Frisch and Giulianelli; 2024)

LLMs are *human enough*.

Motivation

- Survey-based forecasts (e.g., SPF) are critical for policymakers and researchers (e.g., SPF is used by central banks, academia, practitioners)
- Survey data collection is costly (& infrequent); can't easily adapt questions
- LLMs can augment suvey data collection by simulating agent behavior (quickly and cheaply)

The Paper

Goal: Construct LLM-based synthetic forecasters mimicking the *Survey of Professional Forecasters* (SPF) participants

1. Build synthetic forecasters *at the individual level*, based on information of actual SPF participants
2. Use these synthetic personas, past median SPF forecasts, and real-time data as LLM inputs
3. Ask for point forecasts similar to the SPF instrument

The Paper

Goal: Construct LLM-based synthetic forecasters mimicking the *Survey of Professional Forecasters* (SPF) participants

4. Compare accuracy of LLM-based forecasts to SPF forecasts

Framework

- Forecasting process:

$$y_{t+H} = f(x_t, z_t) + \varepsilon_{t+H}$$

with t as current time period, H as forecast horizon, x_t as observable predictors, z_t as unobservable, and ε_t unpredictable with zero mean

- Unobservables z_t represent any additional information that can help predict y_{t+H} but is (very) hard to quantify, e.g.:
  - Private insights
  - Tacit domain knowledge
  - Internalized heuristics
  - Intuition

Humans, Algorithms, and AI

- **Humans** can access both x_t and z_t, but do so imperfectly:

$$h_{i,t} = f(x_t, z_t) + \Delta_{i,t}$$

  - $\Delta_{i,t}$ is human bias that may not have zero mean

- **Traditional algorithms** cannot access z_t but they process x_t efficiently (*direct mapping*):

$$m_t = \mathbb{E}\left[f(x_t, z_t) \mid x_t\right]$$

- **LLMs** are similar to traditional algorithms in that they only access x_t, but expectations are formed differently (*massive text-based probability distribution*):

$$m_t^{\text{AI}} = \mathbb{E}^{\text{AI}}\left[f(x_t, z_t) \mid x_t\right]$$

Humans vs. AI

- The distance between human and AI forecasts ultimately depends on the size of human bias ($\Delta_{i,t}$) relative to LLM's ($\Delta_t^{AI} = m_t^{AI} - f(x_t, z_t)$)

- We can minimize this distance by giving an LLM:

  (1) **Forecaster characteristics** to capture systematic patterns in biases:

  $$\Delta_{i,t} = \gamma(w_{i,t}) + e_{i,t},$$

  (2) **Past median SPF forecasts** to proxy unobservable z_t:

  $$\bar{h}_{t-1} = f(x_{t-1}, z_{t-1}) + \bar{\Delta}_{t-1}$$

- This helps LLMs mimic humans in their forecasting process:

$$m_{i,t}^{AI} = \mathbb{E}^{AI}\left[f(x_t, z_t) \mid x_t,\ f(x_{t-1}, z_{t-1}) + \bar{\Delta}_{t-1},\ w_{i,t}\right]$$

The Survey of Professional Forecasters

About the SPF

- Oldest quarterly survey of macroeconomic expectations in the U.S.
  - Launched in 1968
  - Conducted by the Federal Reserve Bank of Philadelphia since 1990
- Widely used by policy-makers and economic researchers
- Survey questions:
  - 23 point forecasts at nine horizons: the current quarter (nowcast), one to four quarters ahead, the current year, and one to three years ahead
- Survey responses are releases at the individual level, but without forecaster identifiers. However, published surveys include the names and affiliations of recent contributors ▸ Example of Acknowledgments

Data

- We focus on all point forecast variables:
  - U.S. business indicators (e.g., Nominal GDP; Unemployment Rate; T-Bill Rate, 3-month)
  - Real GDP and its components (e.g., Real GDP, Real Personal Consumption Expenditures)
  - Inflation measures (CPI, Core CPI, PCE, Core PCE)
- We forecast over five horizons: nowcast + one to four quarters ahead
- Sample: 1999-2023 + an out-of-sample validation for 2024

Simulating the SPF with LLMs

Synthetic Forecasters

We collect publicly available data (e.g., LinkedIn, personal websites) to:

- Create a set of **synthetic forecasters** by endowing them with:
  - Education, job title, affiliation, company location
  - Experience and possible geographic or sector biases
  - Social media presence, interviews, etc.
- These features vary widely across actual SPF participants individuals

Method

1. We use a set of LLMs (e.g., GPT-4o mini) and prompt them with:
   - **Synthetic forecaster personas** (i)
   - **Real-time data** (up to quarter t)
   - **Past SPF median forecasts**
2. The model is then instructed to forecast the same variables over the same horizons as human SPF forecasters
3. Evaluate LLM forecasts *versus* actual SPF and realized outcomes

Prompt

You are a participant on a panel of Survey of Professional Forecasters. Your name is [name], you graduated from [*alma mater*] with a [education] around [graduation year]. Today, you work as [title] at [affiliation]. It's [affiliation types] organization.
Your organization is based in [company location].
You are originally from [country of origin]. [social media status].
We are in [quarterly date]. You are about to fill out the forecast form for [quarterly date]. Using only the information available to you as of [quarterly date], please provide your best numeric forecasts for the following variables: [variables].
Do this for the following quarters: t (current quarter), t+1, t+2, t+3, and t+4, as well as annual forecasts for this and next year (annual averages). You have the most recent real-time data on key macroeconomics variables available to you as of today: [real-time data].
The forecasts made by the SPF panel during the previous quarter were as follows (for t-1, t, t+1, t+2, t+3, t+4; where t is previous quarter: [past median forecasts].
Do not incorporate any data that was not available to you beyond the current date in your forecasts. Do consider all relevant information on the broad economic conditions and current Federal Reserve actions (up to, but not beyond [release date]).
Use available information, and your professional judgment and experience. Your forecast is anonymous. Provide the forecasts as a sequence of numerical values only. Please only provide your forecasts in the format: (t, t+1, t+2, t+3, t+4, this year's average, next year's average).

Prompt

You are a participant on a panel of Survey of Professional Forecasters. Your name is [name], you graduated from [*alma mater*] with a [education] around [graduation year]. Today, you work as [title] at [affiliation]. It's a [affiliation types] organization. Your organization is based in [company location].
You are originally from [country of origin]. [social media status].

Prompt

You are a participant on a panel of Survey of Professional Forecasters...

We are in [quarterly date]. You are about to fill out the forecast form for [quarterly date]. Using only the information available to you as of [quarterly date], please provide your best numeric forecasts for the following variables: [variables].

Prompt

You are a participant on a panel of Survey of Professional Forecasters...
We are in [quarterly date]...

Do this for the following quarters: t (current quarter), t+1, t+2, t+3, and t+4, as well as annual forecasts for this and next year (annual averages). You have the most recent real-time data on key macroeconomics variables available to you as of today: [real-time data].

Prompt

You are a participant on a panel of Survey of Professional Forecasters...
We are in [quarterly date]...
Do this for the following quarters...

The forecasts made by the SPF panel during the previous quarter were as follows (for t-1, t, t+1, t+2, t+3, t+4; where t is previous quarter): [past median forecasts].

Prompt

You are a participant on a panel of Survey of Professional Forecasters...
We are in [quarterly date]...
Do this for the following quarters...
The forecasts made by the SPF panel during the previous quarter...

Do not incorporate any data that was not available to you beyond the current date in your forecasts. Do consider all relevant information on the broad economic conditions and current Federal Reserve actions (up to, but not beyond [survey release date]).

Prompt

You are a participant on a panel of Survey of Professional Forecasters...
We are in [quarterly date]...
Do this for the following quarters...
The forecasts made by the SPF panel during the previous quarter...
Do not incorporate any data that was not available...

Use available information, and your professional judgment and experience. Your forecast is anonymous. Provide the forecasts as a sequence of numerical values only. Please only provide your forecasts in the format: (t, t+1, t+2, t+3, t+4, this year's average, next year's average).

Prompt

You are a participant on a panel of Survey of Professional Forecasters. Your name is [name], you graduated from [*alma mater*] with a [education] around [graduation year]. Today, you work as [title] at [affiliation]. It's [affiliation types] organization.
Your organization is based in [company location].
You are originally from [country of origin]. [social media status].
We are in [quarterly date]. You are about to fill out the forecast form for [quarterly date]. Using only the information available to you as of [quarterly date], please provide your best numeric forecasts for the following variables: [variables].
Do this for the following quarters: t (current quarter), t+1, t+2, t+3, and t+4, as well as annual forecasts for this and next year (annual averages). You have the most recent real-time data on key macroeconomics variables available to you as of today: [real-time data].
The forecasts made by the SPF panel during the previous quarter were as follows (for t-1, t, t+1, t+2, t+3, t+4; where t is previous quarter: [past median forecasts].
Do not incorporate any data that was not available to you beyond the current date in your forecasts. Do consider all relevant information on the broad economic conditions and current Federal Reserve actions (up to, but not beyond [release date]).
Use available information, and your professional judgment and experience. Your forecast is anonymous. Provide the forecasts as a sequence of numerical values only. Please only provide your forecasts in the format: (t, t+1, t+2, t+3, t+4, this year's average, next year's average).

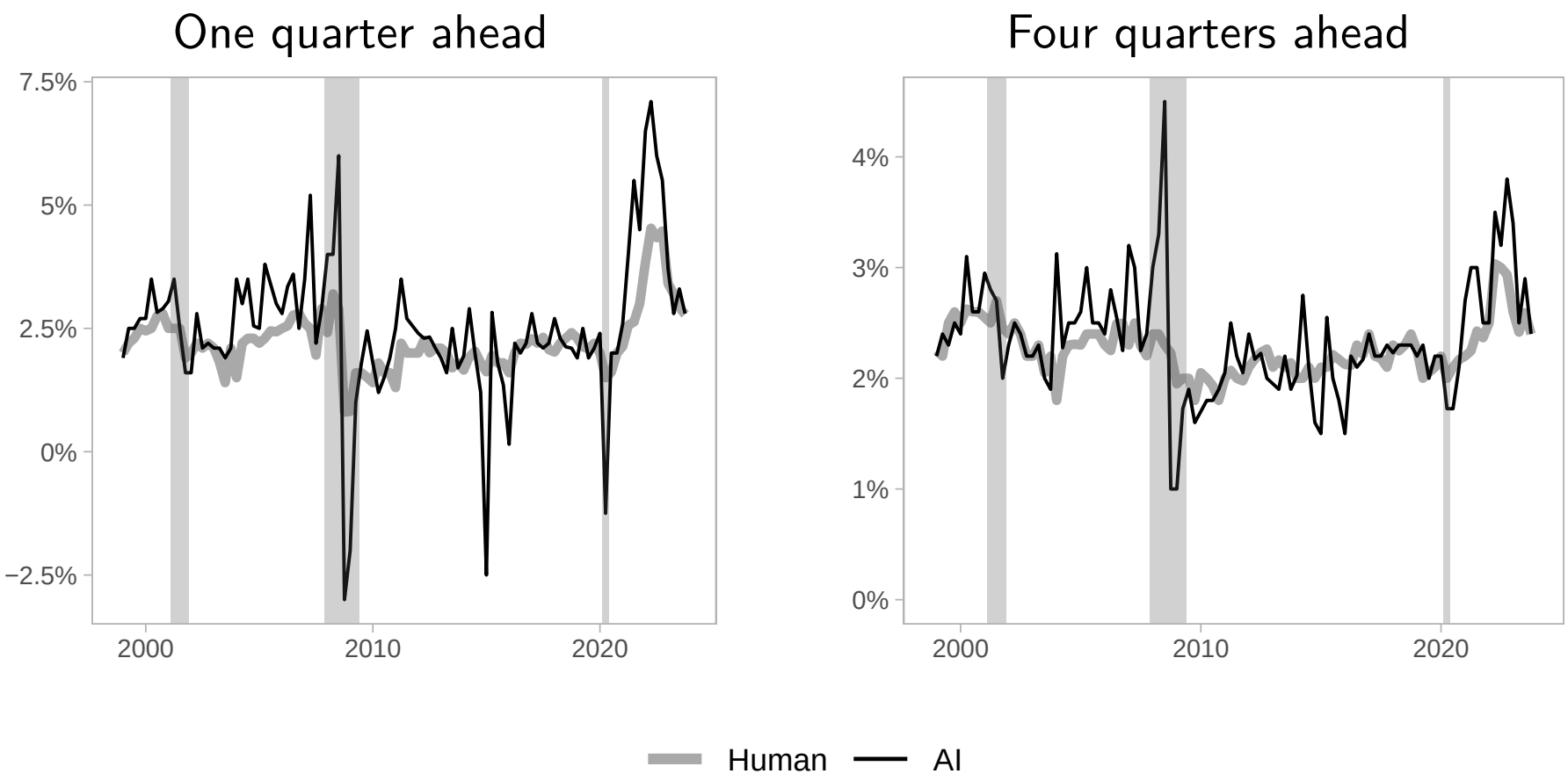
Results

Results

- Large data set comprising point forecasts for 20+ variables at different horizons for both human and AI forecasters
- **Three main take-aways:**
  - #1 AI $\approx$ humans: AI and human forecasts are qualitatively similar, but there are quantitative differences
  - #2 AI $\succ$ humans: AI often achieves lower forecasting errors
  - #3 AI $\succ$ humans | human input: Accuracy of AI hinges on human input in prompt

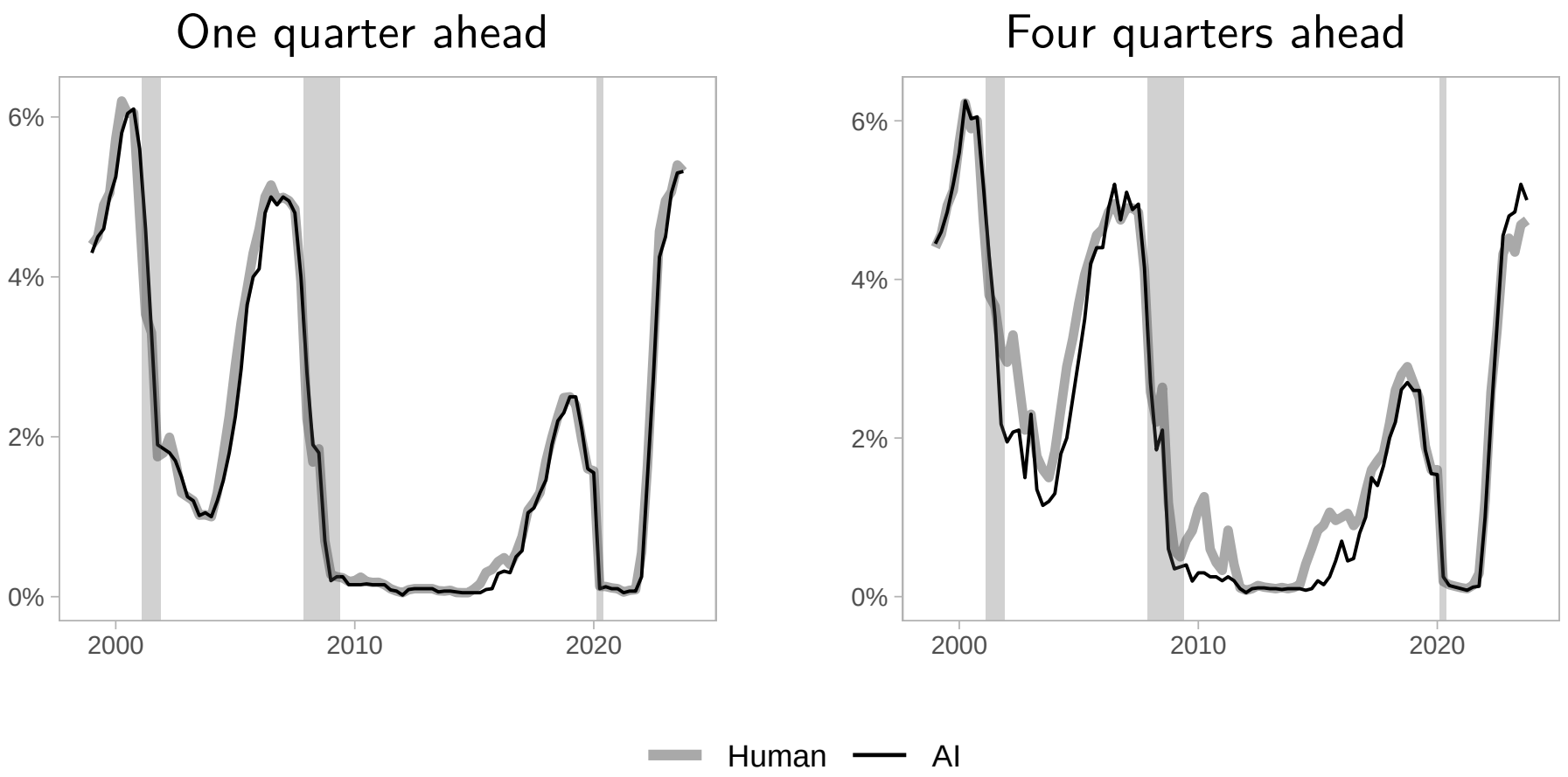
Result #1: AI $\approx$ humans

Median Forecasts: CPI Inflation



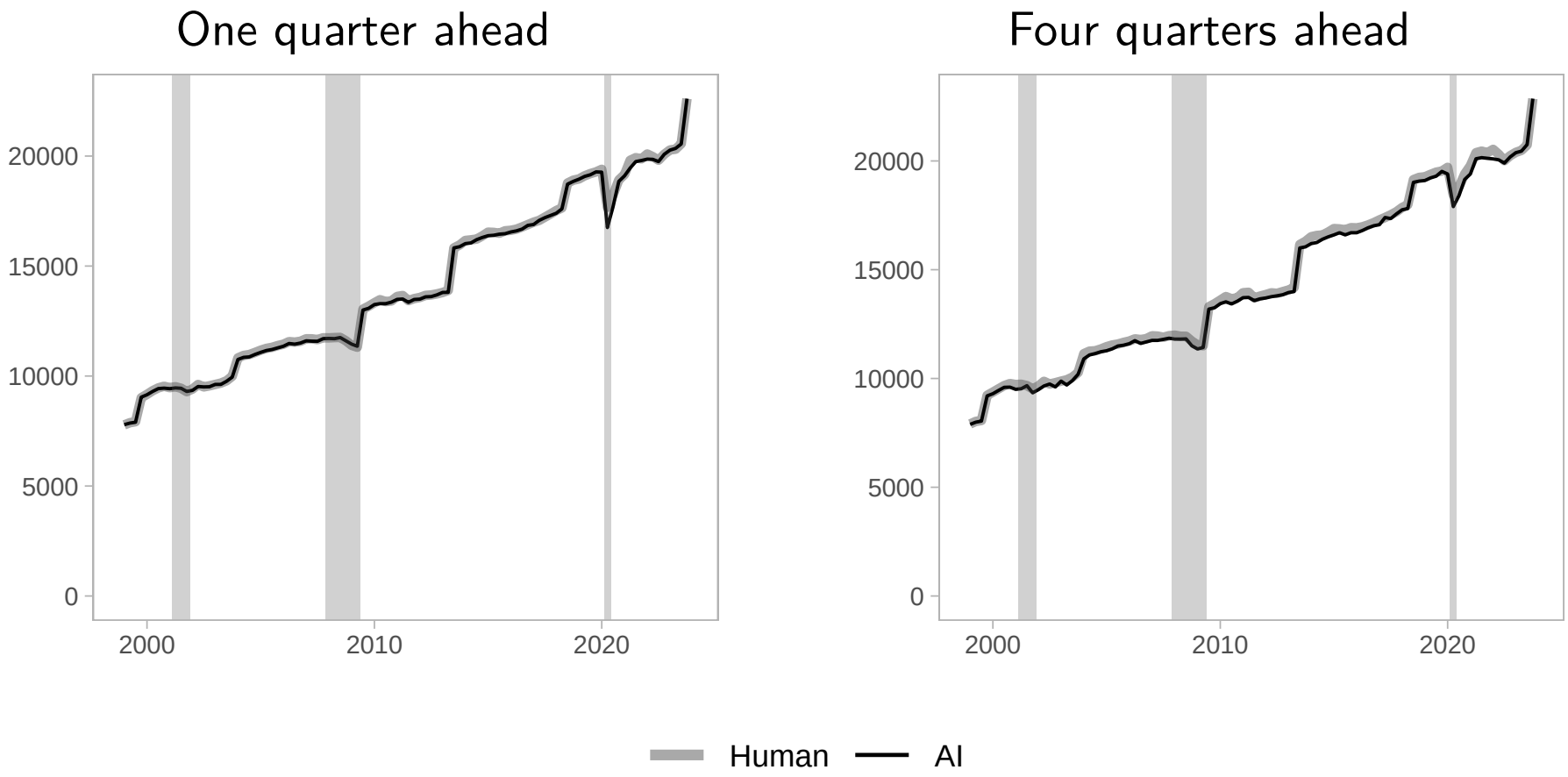
Shaded areas are NBER recessions

Median Forecasts: T-Bill Rate (3-month)



Shaded areas are NBER recessions

Median Forecasts: Real GDP



Shaded areas are NBER recessions

Result #2: AI $\succ$ humans

Forecast Accuracy (Mean Absolute Error)

Horizon (quarters)	0			1			4		
	AI	Human	P-val	AI	Human	P-val	AI	Human	P-val
Section 1: US Business Indicators									
Nominal GDP	248.09	**187.45**	0.90	**161.71**	178.31	0.23	**340.95**	379.87	0.00***
GDP Price	**21.87**	22.16	0.00***	**21.85**	22.20	0.00***	**21.68**	22.35	0.00***
Corporate Profits	87.60	**71.78**	0.01***	**61.80**	101.39	0.02**	**165.31**	186.30	0.21
Unemployment Rate	**0.31**	0.38	0.00***	**0.52**	0.57	0.00***	**0.91**	0.94	0.00***
Non-Farm Payroll	**252.33**	465.23	0.01**	933.18	**804.54**	0.05*	2327.18	**1938.36**	0.00***
Industrial Production	**0.54**	1.64	0.00***	**2.14**	2.99	0.00***	**4.92**	6.27	0.00***
Housing Starts	**0.05**	0.09	0.05**	**0.10**	0.12	0.02**	**0.16**	0.21	0.80
Treasury Bill Rate (3M)	0.35	**0.26**	0.31	0.53	**0.43**	0.01***	**1.15**	1.21	0.00***
AAA Corp Bond Yield	**0.18**	0.28	0.07*	**0.37**	0.44	0.00***	**0.59**	0.73	0.00***
Treasury Bond Rate (10Y)	0.36	**0.32**	0.63	**0.48**	0.51	0.00***	**0.76**	0.88	0.00***
Section 2: Real GDP and Its Components									
Real GDP	**90.82**	126.20	0.00***	**169.81**	209.17	0.00***	**524.39**	568.19	0.00***
Real PCE	1393.03	**90.64**	0.00***	1454.12	**130.32**	0.00***	1710.50	**330.69**	0.00***
Real Non-Res Fixed Inv	**21.91**	25.92	0.11	**36.11**	48.83	0.00***	**116.08**	133.96	0.00***
Real Res Fixed Inv	**10.40**	10.43	0.86	**13.89**	18.10	0.39	**49.81**	54.14	0.00***
Real Federal C&GI	9.94	**6.79**	0.84	**14.32**	19.34	0.04**	46.88	**45.62**	0.00***
Real State/Local C&GI	9.66	**8.09**	0.23	**20.04**	24.39	0.00***	**68.31**	69.76	0.00***
Real Change in Private Inv	35.35	**24.91**	0.19	**19.46**	38.13	0.10*	51.62	**48.89**	0.02**
Real Net Exports	26.97	**16.79**	0.54	**24.40**	42.52	0.02**	**90.27**	91.17	0.00***
Section 3: CPI and PCE Inflation									
CPI Inflation Rate	**1.84**	1.98	0.00***	2.36	**2.14**	0.01***	**2.03**	2.06	0.04**
Core CPI Inflation Rate	**0.67**	0.82	0.02**	0.92	**0.88**	0.00***	**0.97**	1.00	0.03**
PCE Inflation Rate	**2.40**	2.49	0.54	**2.27**	2.72	0.55	**3.13**	**3.13**	0.85
Core PCE Inflation Rate	2.37	**2.30**	0.01**	2.31	**2.21**	0.15	**2.12**	2.14	0.01**

Proportion of Quarters Where AI is More Accurate

Horizon (quarters)	0		1		4	
	Pct	P-val	Pct	P-val	Pct	P-val
CPI Inflation Rate	**0.69**	0.01***	0.47	0.74	**0.55**	0.78
T-bill	**0.51**	1.00	0.47	0.97	**0.60**	0.08*
Unemp	**0.81**	0.00***	**0.74**	0.01***	**0.63**	0.22
Real GDP	**0.70**	0.00***	**0.75**	0.00***	**0.63**	0.01**

Boldfaced values are ≥ 0.5. P-val reports significance of randomized tests of Pct$= 0.5$.

Forecast Accuracy (MAE)

- AI forecasts often outperform human forecasts, especially at longer horizons
- Gains are most pronounced for variables like real GDP and unemployment rate

"LLMs extract latent (z_t) information from human forecasts while also processing x_t more effectively."

Result #3: AI $\succ$ humans $\mid$ human input

AI Forecast Accuracy without Human Input

	Generic		Generic, w/o real-time data		Generic, w/o real-time data, w/o past SPF data	
Horzion	0	4	0	4	0	4
T-bill	**1.09**	**1.01**	0.74	**1.03**	**1.07*****	**1.08*****
Unemp	**1.02**	**1.02*****	**1.20**	**1.02**	**1.12**	**1.10*****
Real GDP	**1.15**	**1.04**	**1.37**	**1.08**	**7.57*****	**1.53*****
CPI	0.90	**1.02**	**1.09**	**1.02**	**1.09**	**1.13****
Average	**1.14**	**1.06**	**1.31**	**1.06**	**8.88**	**2.52**

Values are MAEs relative to MAEs of baseline AI forecasts. Boldfaced values are ≥ 1.
P-val reports significance of randomized tests of Pct$= 1$.

Addressing Temporal Leakage

- LLM might recall future data from its training set
- Mitigation:
  - Strict instructions to use only data *up to t*
  - Real-time "dated" data sets (no future info)
  - Out-of-sample test (e.g., 2024 data) outside model's training window
- Recall test: Ask the model to recall past realized values from the data set. On average, errors are 16x larger than our baseline nowcasting results.

Discussion

- Humans have access to unobservable insights but can suffer systematic biases
- LLMs see only structured data and historical patterns, but can approximate the "latent" aspects by:
  - reading past human forecasts,
  - adjusting to persona-specific biases
- Hybrid approach: AI + human signals can exceed pure human or purely data-driven ML forecasts
- Demonstrates the viability of AI-augmented macroeconomic surveys, potentially powerful for policy or research: "virtual forecasting lab"

Forecasting Failure Mode

- In measurement, the central risk is noisy or biased labels
- In forecasting, the central risk is **lookahead**
- LLMs may have seen future text, revised data, solved benchmarks, or leaked labels
- A model that looks good in backtests may be grading its memory

Large Language Models: An Applied Econometric Framework

Jens Ludwig, **Sendhil Mullainathan** & **Ashesh Rambachan**

Why a framework?

Think of an LLM as a very talented but mysterious intern: it can read text and give you predictions/labels you want, but you don't really know what it studied or remembers..

- LLMs are powerful *and* brittle.
- We need "contracts" for when economists can safely use that intern's work in prediction or estimation without breaking econometric guarantees.

Why a framework?

Think of an LLM as a very talented but mysterious intern: it can read text and give you predictions/labels you want, but you don't really know what it studied or remembers..

- Just like in OLS, if your assumptions don't hold, your inference breaks..

Two Research Tasks in Economics with Text

- **Prediction:** Use text data to forecast an associated numerical or categorical outcome.
  Examples include:
  - Predicting stock returns from earnings call transcripts.
  - Predicting whether a bill will pass based on its full text or legislative debate.

  The goal is to minimize out-of-sample prediction error, often using held-out data. Accuracy is judged by comparing predictions to real outcomes.

Two Research Tasks in Economics with Text

- **Estimation:** Use text to **measure a latent or abstract concept** that enters a downstream econometric analysis. Examples include:
  - Constructing a "policy uncertainty" index from central bank speeches to estimate its effect on investment.
  - Using summaries of bills to quantify ideological leanings and estimate effects on voting behavior.

  The goal is to produce a valid, interpretable regressor or label that allows for meaningful causal or structural inference.

Key Condition for Prediction to Be Valid

- **No Training Leakage:** The LLM must not have seen your evaluation data (text or labels) during its training.
  Otherwise, you're not measuring generalization, you're grading its memory.
- *Why this matters:* If the model has memorized parts of your test set (e.g., past headlines, legal cases, or transcripts), its performance will appear artificially high.
- To test its true out-of-sample performance, your inputs must be **new and unseen** to the model.
- Best practices:
  - Use open-source LLMs with a known training cutoff date.
  - Ensure your test data (or its metadata) didn't circulate publicly before that date.

Why leakage is common

- Closed models don't disclose data; often trained on web corpora & benchmarks.
- Memorization documented in CS; risk extends to econ datasets.
- Prompting cannot "un-train" a model; it only affects generation, not training.

Evidence of leakage in econ contexts

- **Congressional bills:** Model can complete bill descriptions verbatim; predicts passage with implausibly high accuracy.
- **Finance headlines:** Model reproduces exact headlines; "look-ahead" bias also observed in related work.

Estimation (1/2): Motivation and Measurement

- We want to estimate an economic relationship involving a text-based concept V.

  Examples of V: policy sentiment, hawkishness, ideology, topic intensity, credibility.

- V captures a latent or abstract concept and usually requires human interpretation.

- A gold-standard version of V can be obtained through:
  - Expert readings of the text
  - Manual coding by trained annotators
  - Surveys or focus group analysis

  Problem: These methods are accurate, but costly, slow, and not scalable to thousands of documents.

Estimation (2/2): Temptation and Solution

- **Temptation:** Use an LLM to automatically generate proxy labels $\hat{V}$ and plug $\hat{V}$ directly into the regression.
- This is appealing because LLM labels are cheap, fast, and scalable.
- **But beware:** $\hat{V}$ is a noisy approximation of the true concept V.
- If LLM errors are systematic, they can:
  - Bias the estimated relationship
  - Change coefficient magnitudes or signs
  - Lead to misleading inference
- **Solution:** Use the LLM for scale, but use a small gold-standard sample to calibrate and correct its errors.

Why Estimation Is Fragile

- Plug-in estimation is valid only if LLM measurement errors are not systematically related to the variables in the regression.
- Accuracy alone does not guarantee valid downstream inference. Even seemingly nonsystematic LLM measurement error can bias estimates, depending on how the measured concept enters the model.
- An LLM can be "pretty good" on average, but still distort the economic relationship we care about.
- Example: The LLM may systematically understate hawkishness when inflation is high.
- Then a regression using $\hat{V}$ may understate how strongly hawkishness responds to inflation.

Simple Debiasing: Intuition

- The large LLM-labeled sample tells us the relationship (i.e., b/w hawkishness and in) at scale.
- The small human-labeled sample tells us how that relationship is distorted.
- We use the human labels to estimate the LLM's systematic error.
- Then we add that correction back to the large-sample estimate.
- **Intuition:** The LLM gives us cheap measurement at scale. The human labels tell us how to calibrate it.

Comparing Estimation Approaches

- **Gold-standard only approach:**
  - Uses only the small set of human-coded labels V
  - Accurate, but often high-variance because n is small
- **Naive LLM approach:**
  - Uses $\hat{V}$ for the full sample
  - Low cost and low variance
  - But potentially biased if LLM errors are systematic
- **Debiased LLM approach:**
  - Uses $\hat{V}$ on the large sample
  - Uses human labels on a small validation sample
  - Corrects the large-sample estimate using the validation sample

When Is Debiased Estimation More Precise?

- **Works best when:**
  - LLM labels are reasonably informative
  - The primary sample is large
  - The validation sample is representative
  - LLM errors are structured enough to estimate and correct
- **Intuition:** The LLM acts as a noisy but cheap labeling machine.
- We let it scale up measurement, then use a smaller set of gold labels to correct its mistakes.
- **Analogy:** Like scanning thousands of pages with a blurry scanner. If the blur is consistent, you only need to inspect a few pages by hand to correct the full scan.

If you have no validation data

Three (weak) options:

1. Assume zero error in $\hat{V}$ (not credible)
2. Model LLM error structurally (heroic)
3. Redefine target as "the LLM's label" (changes the question)

Conclusion: Collect some validation labels.

Other uses: Simulating human subjects

- Treat like **estimation**: V = average human response.
- Use LLM for many designs, but **validate** on subset of real participants.
- Useful when there are **many** designs (e.g., large choice sets) – work as "sample amplifiers".

Replicability in LLM-based estimation

Include in replication package:

1. **Benchmark data** (gold labels)
2. **Prompts + open-source model snapshot** (fixed weights)
3. **Final data & code** for debiasing and estimation

Closed models risk non-replicable outputs..

Practical checklist

- Define prediction vs estimation **up front**
- For prediction: enforce **no leakage**
- For estimation: collect **validation labels**; report **bias-corrected** results

- Document prompts and model versions
- Don't rely on accuracy alone; don't use closed models for key inferences

Chronologically Consistent LLMs: The Problem

- In a historical prediction exercise, the model should only use the information set available at time t
- With standard LLMs, this condition can fail even when the *prompt* contains only historical information:
  - the model weights may have been trained on text from the future
  - embeddings can therefore encode information unavailable at time t
- He, Lv, Manela, and Wu address this by training separate **vintage models**: ChronoBERT and ChronoGPT models whose training data stop at a known historical cutoff

Source: He, Lv, Manela, and Wu, "Chronologically Consistent Large Language Models," 2025.

Chronologically Consistent LLMs: Does Leakage Matter?

- They use the appropriate historical model vintage to encode financial news and predict next-day stock returns
- This creates a direct comparison:

  What could the model have predicted then? vs. **What can a modern LLM predict looking backward?**

- Surprisingly, the chronologically consistent models perform about as well as the much larger Llama 3.1 model in their portfolio application
- **Takeaway:** lookahead bias is modest in this application, but its magnitude is *model- and application-specific*
- So chronological consistency gives us a way to **measure leakage rather than simply assume it away**

Source: He, Lv, Manela, and Wu, "Chronologically Consistent Large Language Models," 2025.

Forecasting Is More Than the Mean

- Point forecasts ask whether a model can match professional forecaster consensus
- Scenario forecasts ask whether a model can condition on counterfactual paths
- Structural scenarios ask whether the model knows why the counterfactual happens
- Density forecasts ask whether the model can quantify uncertainty, not just choose a number

What Does the Evidence Say So Far?

The evidence on LLM forecasting is promising, but mixed.

- **Macroeconomics:** LLMs can sometimes match or outperform professional forecasts
- **Forecasting tournaments:** LLM systems are becoming competitive with human crowds
- **Ensembles and retrieval:** combining models or adding current information can improve performance

Two issues remain:

- Was the information genuinely available at the forecast date?
- Is the model's reported uncertainty actually calibrated? A model can be accurate but poorly calibrated. For instance, it might get 85% of questions right overall but routinely claim 99% confidence. It is good at answering, but overconfident.

Source: Faria-e-Castro and Leibovici (2023); Halawi et al. (2024); Schoenegger et al. (2024).

Questions?

Day 3 wrap-up