Machine Learning, Artificial Intelligence, and Natural Language Processing

Marcelo C. Medeiros

Department of Economics
The University of Illinois at Urbana-Champaign

Central message

- AI is changing economic research because it expands what economists can **measure**, **forecast**, and **simulate**.
- But AI does not replace identification, economic theory, or careful validation.

A practical organizing principle

Treat AI output as either a **prediction**, a **measurement**, or a **simulation**. Each has different econometric requirements.

Roadmap

I. Text as data → LLMs as measurement

[central bank communication, firm disclosure]

II. AI-based expectations

[querying models for beliefs; case study: Wu, Xi & Xie (2026)]

III. Simulated economic agents

[homo silicus, silicon samples, and their limits]

IV. Forecasting

[inflation, returns, and the memorization problem]

V. Discovery and causal inference

[hypothesis generation, IV search]

VI. Econometrics of AI-generated variables

[the part that should worry us most]

VII. Practice, reproducibility, profession

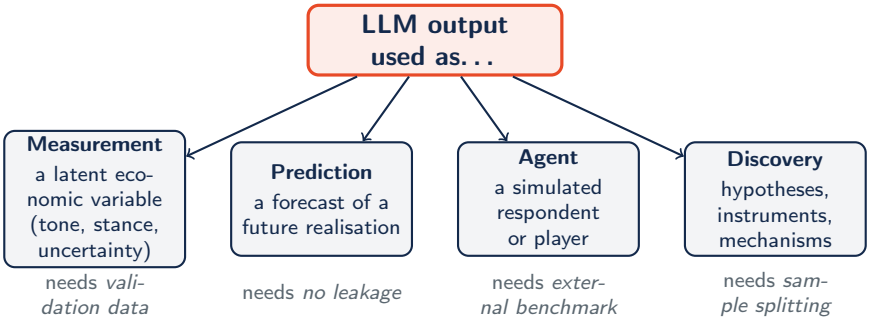
Three waves of "AI" in economics

1. **Statistical learning (c. 2014–2018).** Prediction as a first-class object; regularization, cross-validation, prediction policy problems.
   [Varian (2014); Mullainathan & Spiess (2017); Kleinberg, Ludwig, Mullainathan & Obermeyer, AER P&P 2015]
2. **Causal machine learning (c. 2016–2022).** Orthogonalization, heterogeneous effects, high-dimensional controls with valid inference.
   [Belloni–Chernozhukov–Hansen; Chernozhukov et al. (DML); Athey & Imbens; Wager & Athey]
3. **Foundation models / LLMs (2022–).** Unstructured data at scale, *and* a qualitatively new object: a model that emits plausible human responses.
   [Korinek, JEL 2023; Dell, JEL 2025; Ludwig, Mullainathan & Rambachan 2025]

The discontinuity

Waves 1–2 changed how we *estimate*. Wave 3 changes what we can *measure* — and tempts us to change what we treat as *data*.

A taxonomy for the rest of the talk



[Organising logic follows Ludwig, Mullainathan & Rambachan (2025), "LLMs: An Applied Econometric Framework"]

What is an "AI agent"? Two vocabularies

Agent, in economics

- A decision-maker characterised by a *preference ordering*, an *information set*, and a *constraint set*.
- Behaviour is derived: it is the solution to an optimisation problem, closed by an equilibrium concept.
- The model is transparent by construction. We can state exactly why the agent did what it did.

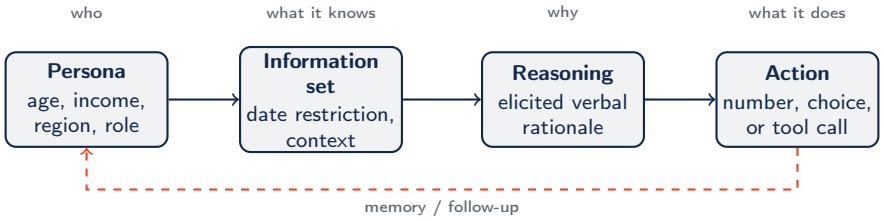
Agent, in AI

- A language model placed in a *loop*: it is given a role, receives an input, emits an action, and observes the result.
- Behaviour is *generated*, not derived — next-token prediction conditioned on a role description and a context.
- There is no objective function, no budget constraint, and no equilibrium. Consistency is an empirical property, not a theorem.

Working definition for this talk

An AI agent is an LLM instantiated with (i) a *persona* — a role and attribute vector, (ii) a *restricted information set*, (iii) *memory* across turns, and (iv) an *action space* (a number, a choice, a text, or a tool call). It is a sampling device from a conditional distribution over responses — not an optimizer.

Anatomy of an AI agent, and the four design margins



- **Persona** controls the cross-sectional distribution; drawn from a real survey, it inherits that survey's demographics.
- **Information set** is the only lever that separates "forecast" from "recall".
- **Reasoning** is a free by-product with no human-survey analogue — and, as we will see, a new source of economic evidence.
- **Memory** turns a cross-section into a panel: the same agent can be re-interviewed at any horizon.

I. Text as Data: LLMs as Measurement Instruments

II. AI-Based Expectations

III. Simulated Economic Agents

IV. Forecasting

V. Discovery and Causal Inference

VI. The Econometrics of AI-Generated Variables

VII. Research Practice and the Profession

The pre-history: text as high-dimensional data

- Economics has treated text as data for two decades. The econometric content was always the same: a very high-dimensional X mapped to a low-dimensional economic object.
- **Dictionary methods.** Loughran–McDonald financial sentiment lists; Baker, Bloom & Davis (QJE 2016) Economic Policy Uncertainty.
- **Unsupervised topic models.** LDA (Blei et al. 2003); Hansen, McMahon & Prat (QJE 2018) on FOMC transparency and deliberation.
- **Supervised learning.** Gentzkow, Kelly & Taddy (JEL 2019) survey; penalised regression on document-term matrices.
- **Contextual encoders.** BERT (2019) and finance/central-bank variants (FinBERT; CB-LMs, BIS WP 1215).
- **Generative LLMs.** Few-shot or zero-shot classification, extraction, and structured coding without task-specific training data.

What actually changed with generative models

Three distinct gains

1. **Cost of annotation collapses.** Constructs that required trained RAs ("is this sentence hawkish?") can be coded at scale.

2. **Constructs become expressible in natural language.** You can measure "the speaker is signalling a data-dependent pause" without a labelled corpus.

3. **Context and negation are handled.** Dictionary methods fail on "inflation is *no longer* elevated"; contextual models generally do not.

What did *not* change

The output is still an estimate $\hat{\theta}_i$ of a latent θ_i, with error. Everything downstream is a generated-regressor problem. We return to this in Part VI.

Central bank communication: why economists care

- Communication is a policy instrument. Words move the yield curve independently of the policy rate.

  [Kuttner (2001); Gürkaynak, Sack & Swanson (2005); Nakamura & Steinsson (2018)]

- The identification problem: an announcement bundles a policy shock and an information shock about the central bank's own outlook.

  [Jarociński & Karadi (AEJ:Macro 2020); Bauer & Swanson (2023)]

- High-frequency surprises are a *residual*: informative but noisy and contaminated by the Fed's private information.
- **Text offers a different route**: directly reconstruct the information set and the stance, rather than cleaning the surprise ex post.

Can a model read the Fed as well as an expert?

- Hansen & Kazinnik (2024) show GPT-4 classifies FOMC communication stance at a level comparable to expert human coders — the benchmark result that opened the area.
- BIS work (Gambacorta, Kwon, Park, Patelli & Zhu, WP 1215) builds **CB-LMs**, central-bank domain-adapted language models, and finds they outperform foundational models on masked-word prediction in central-bank idiom and on stance classification of FOMC statements — beating much larger general-purpose generative LLMs on that specific task.

Reading for an econometrician

Domain adaptation still wins where the label space is narrow, and the idiom is specialized. "Bigger model" is not automatically "better measurement". Benchmark against a fine-tuned small model before paying for frontier inference.

Scale: what becomes feasible

- IMF WP 25/109 classifies central bank communication on four dimensions — **topic, communication stance, sentiment, audience** — with a fine-tuned LLM, sentence by sentence.
- Corpus: 74,882 documents from 169 central banks, 1884–2025, multilingual.
- Findings include a structural break with the adoption of inflation targeting: backward-looking exchange-rate discussion gives way to forward-looking statements on inflation, rates and conditions.
- They construct a *directional communication index* capturing signals about future policy-rate changes and unconventional measures.

Why this matters methodologically

This is a panel that could not exist without automated coding: 169 cross-sectional units, 140 years, multiple languages. The identification opportunities are the real prize.

From text to a monetary policy shock

- **Narrative identification at scale.** Fernandez-Fuertes (2025) builds a multi-agent LLM pipeline that reads Beige Books and Minutes released *before* each FOMC meeting, forms conditional expectations, and differences the realised decision against them.
- Reported results: less noisy surprises than market-based measures; impulse responses in which contractionary shocks generate persistent disinflation; profitable yield-curve strategies.
- Related: LLM-derived linguistic indices of policy stance built from the *procedural structure* of communication, constructed independently of contemporaneous market movements — attractive for cross-country work where disclosure practices are staggered.

Beyond central banks: firms, disclosure, and markets

- **Corporate policy from conversation.** Jha, Qian, Weber & Yang, "ChatGPT and Corporate Policies" — LLM-summarised earnings-call content predicts investment and financing decisions.
- **Disclosure quality.** Kim, Muhn & Nikolaev on LLM-based analysis of financial statements and on "bloated disclosure" — measuring redundancy and informativeness at scale.
- **News and returns.** Lopez-Lira & Tang: headline-conditioned LLM sentiment predicts next-day returns in-sample. This is exactly the design most exposed to the memorization critique (Part IV).
- **Labour and organisation.** Vafa et al. measure career histories from CVs; Hansen, Lambert, Bloom, Davis, Sadun & Taska classify remote-work provisions in job postings at scale.

Part I: what to take away

1. LLMs are best understood here as cheap, high-throughput annotators for constructs that were previously RA-limited.
2. The economic payoff is not the classifier; it is the *panel* the classifier makes feasible, and the identification strategies that panel supports.
3. Domain-adapted smaller models remain competitive — and cheaper, more reproducible, and open-weight.
4. Every measurement is an estimate. Whether the downstream regression is valid is a separate question that the classification accuracy statistic does not answer.

I. Text as Data: LLMs as Measurement Instruments

II. AI-Based Expectations

III. Simulated Economic Agents

IV. Forecasting

V. Discovery and Causal Inference

VI. The Econometrics of AI-Generated Variables

VII. Research Practice and the Profession

The expectations agenda, briefly

- Modern macro-finance has largely abandoned full-information rational expectations (FIRE) as a description of *measured* beliefs.
- Robust empirical regularities:
  - **Consensus under-reaction** to news; individual over-reaction.
    [Coibion & Gorodnichenko (AER 2015); Bordalo, Gennaioli, Ma & Shleifer (AER 2020)]
  - **Extrapolative return expectations**, negatively correlated with objective expected returns.
    [Greenwood & Shleifer (RFS 2014)]
  - **Belief distortions** as a source of business-cycle and asset-price dynamics.
    [Bianchi, Ludvigson & Ma (AER 2022)]
- The binding constraint has always been data: surveys are short, coarse, infrequent, and stop where the panel starts.

Bybee (2023): a survey of a machine

Design. Query an LLM with Wall Street Journal articles, month by month, 1984–2021, and elicit expectations of financial and macroeconomic variables. Treat the output as a synthetic survey series.

- The resulting expectations closely match the SPF, the AAII individual-investor survey, and the Duke CFO survey.
- They reproduce the *deviations* from FIRE found in those surveys: macro expectations display the under-reaction found in consensus SPF forecasts.
- Return expectations are extrapolative, disconnected from objective measures of expected returns, and negatively correlated with future realized returns — the Greenwood–Shleifer pattern.

[Bybee, "Surveying Generative AI's Economic Expectations", arXiv:2305.02823]

The generalization test — and why it is the whole paper

- The natural objection: the model has read the SPF. It is retrieving, not forecasting.
- Bybee's response: repeat the exercise on articles outside the model's training period. The correlation with existing survey measures *persists*.
- Interpretation: the model is not reproducing memorized series; it is mapping a news information set into a belief in a way that mimics human respondents.

What this gives us if it holds up

A *laboratory* for expectation formation: hold the information set fixed, vary the persona, the salience, the framing, the horizon — and read off the belief. That is an experiment we cannot run on the SPF.

Personas and heterogeneous belief formation

- Extension: assign demographic or experiential personas and compare with household surveys.
- Example design: mimic the information set and demographics of the Bank of England's Inflation Attitudes Survey, exploiting a model training cut-off of September 2021 as a quasi-experiment — the model has no knowledge of the subsequent UK inflation surge.
- Reported result: the model tracks aggregate survey projections and official statistics at short horizons, and replicates disaggregated regularities in household inflation perceptions by income and housing tenure.
- Related: persona designs isolating the "experience" channel — forecasters with professional memory of the 1970s versus forecasters shaped by the Great Moderation — holding data and instructions fixed.

Two ways to read these results

Optimistic reading

- The LLM is a compressed representation of how humans map text into beliefs.
- It reproduces *biases*, not just conditional means — which is the harder test.
- It extends belief series backwards and to populations that were never surveyed.
- It permits counterfactual information sets.

Sceptical reading

- "Matches the SPF" has low power: many procedures match a smooth persistent series.
- The training corpus contains commentary *about* survey biases; reproducing them may be stylistic mimicry.
- Persona sensitivity: results can move with prompt wording in ways no human survey analogue does.
- Which model, which date, which temperature — and does the finding survive the next release?

Discriminating test

Report performance against a *non-trivial* benchmark (a fitted AR/noisy-information model), not only correlation with the survey. And pre-register the persona grid.

Wu, Xi & Xie (2026): the design

Goal. Not "can an LLM answer a survey?" but "can an LLM survey recover treatment effects from a randomised control trial?"

1. **Date restriction.** A prompt freezes the agent's knowledge at the survey date. Validated by an *event-awareness test* on six salient events (9/11, Iraq, Lehman, two elections, COVID): essentially zero recognition before, near-universal after.
2. **Internal consistency.** The same 200 personas — drawn from the NY Fed SCE — are held fixed across arms and waves, with treatment wording identical throughout.

- Benchmark: Weber et al. (*Econometrica* 2025) — a ten-wave household RCT, 2018–2023, treatments: past inflation (T1), the Fed's 2% target (T2), the FOMC forecast (T3). Implementation: GPT-4.1 (June 2024 cut-off), replicated with GPT-5.

Validation: does it recover the human treatment effects?

- **Priors line up.** Regressing LLM median priors on SCE medians gives a slope of 1.05. Against the Michigan Survey over the long sample, the correlation is 0.84 — despite the LLM eliciting a full probability distribution and the MSC a point forecast.
- **The central result replicates.** Weber et al.'s finding is that responsiveness to information falls as inflation rises. The LLM survey reproduces it.

Correlation of scaled slope with inflation	T1	T2	T3	Pooled
LLM survey	0.92	0.73	0.85	0.79
Human survey [Weber et al. 2025]	0.69	0.87	0.13	0.57

- **Cost.** Roughly $40 in total, against a $150,000–$1,000,000 estimate for the human benchmark — a factor of up to 25,000.

What the framework then does that no human survey can

1. **Retrospective coverage.** Extend ten waves (2018–23) to more than fifty waves back to 1990, quarterly during the Great Recession and the post-COVID surge, annually otherwise. The state dependence survives, though attenuated: correlations of 0.61 (T1) and 0.76 (T3), but only 0.18 for the Fed-target treatment — plausibly because the 2% target is constant and the agents know it.
2. **Elicited reasoning.** Agents state *why* before answering. A mixed human–LLM coding scheme yields three channels: mean reversion, attention (individual vs. aggregate), and business-cycle conditions.
3. **Dynamic treatment effects.** Monthly follow-ups for twelve months: the effect is about half its initial size at three months, statistically insignificant by six, and gone within a year.
4. **Clean identification.** Treat with the CPI print for the survey month — information not yet released. Impossible in a real-time human survey.

I. Text as Data: LLMs as Measurement Instruments

II. AI-Based Expectations

III. Simulated Economic Agents

IV. Forecasting

V. Discovery and Causal Inference

VI. The Econometrics of AI-Generated Variables

VII. Research Practice and the Profession

Homo silicus

- Horton (NBER WP 31122; ACM EC 2024): LLMs are *implicit computational models of humans*. Endow them with preferences, information and constraints, then run the scenario.
- Replicates four classic experiments with qualitatively similar results — Kahneman–Knetsch–Thaler fairness, status-quo bias, a minimum-wage / employer-preference design.
- Claimed uses: piloting before costly fielding, extending designs to conditions never run, hypothesis search.
- Horton's memorization defence: the pattern is consistent across five models and thousands of simulations, and payoffs are varied *beyond* the original design, so a memorised script would not suffice.

The supporting literature

- **Turing Experiments.** Aher, Arriaga & Kalai (ICML 2023) formalise the protocol and replicate the Ultimatum Game, garden-path sentences, the Milgram study, and wisdom-of-crowds effects.
- **Silicon sampling.** Argyle, Busby, Fulda, Gubler, Rytting & Wingate (Political Analysis 2023): conditioning on demographics reproduces group-level survey response distributions with striking fidelity ("algorithmic fidelity").
- **Generative agents.** Park et al. (2023): multi-agent, memory-based societies exhibiting believable individual and social behaviour.
- **Predicting experiments.** Ashokkumar et al. report correlations of $r \approx 0.85$ between LLM-simulated and actual results across a large archive of social science experiments — and $r \approx 0.90$ on unpublished studies that could not have been in the training data.
- **Automated social science.** Manning, Zhu & Horton: generate hypotheses via structural causal models, then test them in silico.

Where it breaks — and it does break

Bisbee, Clinton, Dorff, Kenkel & Larson (Political Analysis 2024)

Synthetic ANES respondents: means are close, but the distributions are biased and overconfident. Crucially, *regression results estimated on human and synthetic samples often differ* — coefficients relating characteristics to opinions can shrink or flip sign.

- **Homogenization / variance compression.** Simulated subgroups are too internally uniform; minority opinions within a subgroup are suppressed. Marginal means survive; second moments and tails do not.
- **Structural inconsistency.** Accuracy at one level of aggregation does not carry to intersectional subgroups.
- **Task dependence.** Alignment is good on neutral or factual items and much worse on socially sensitive ones.
- **Asymmetry across roles.** In ultimatum-game replications, proposer-side behavior can be matched while responder-side behavior is not, under any tested configuration.

[Wu, Xi & Xie estimate a randomized *difference*, so persona-level biases net out.]

When is a silicon experiment credible? A checklist

1. **Is the estimand a mean or a distribution?** Means, sometimes. Distributions, dispersion, tails: presume failure until proven otherwise.
2. **Is there a held-out human benchmark?** Simulation without a validation sample is a conjecture, not evidence.
3. **Is the design outside the training corpus?** Prefer novel conditions and payoff grids.
4. **Is the result stable across models and prompts?** Report the prompt grid and across-model dispersion, as you would a specification curve.
5. **Is the mechanism economic or linguistic?** A treatment that is a wording change may be measuring the prompt, not the preference.
6. **Is the role of the agent symmetric?** Test both sides of a strategic interaction.

[Charness, Jabarian & List discuss the complementary — rather than substitutive — role of LLMs in experimental design]

Part III: what to take away

- The strongest use is ex ante: piloting, power calculation, design search, hypothesis generation — where being approximately right is enough and the human experiment still follows.
- The weakest use is substitutive: replacing a survey or an experiment and running inference on the synthetic sample as if it were data.
- The intermediate use is the interesting one: LLM agents as a *model of bounded rationality* to be disciplined by data, in the same way we discipline a behavioural model — not as a data source.

Open question worth a dissertation

Under what conditions does an LLM agent population satisfy the exchangeability or transportability assumptions needed for its moments to identify human population moments? Work on prediction-powered and design-based inference (Part VI) is the natural formal home.

I. Text as Data: LLMs as Measurement Instruments

II. AI-Based Expectations

III. Simulated Economic Agents

IV. Forecasting

V. Discovery and Causal Inference

VI. The Econometrics of AI-Generated Variables

VII. Research Practice and the Profession

Two very different claims

Claim A — the LLM as a *feature extractor*

Text is converted into a numeric signal; a conventional forecasting model uses that signal. The LLM never sees the target. *Econometrically:* a generated-regressor problem. Tractable.

Claim B — the LLM as a *forecaster*

The model is asked directly for $\mathbb{E}_t[y_{t+h}]$. Its parameters encode the entire internet up to some cut-off. *Econometrically:* an identification problem. If the evaluation window precedes the training cut-off, "forecast" and "recall" are observationally equivalent.

Most of the excitement is about B. Most of the defensible evidence is about A.

The headline result: inflation

Faria-e-Castro & Leibovici (FRB St. Louis WP 2023-015; *Review* 106(12), 2024)

- Use PaLM to produce distributions of *conditional* inflation forecasts at multiple horizons, 2019–2023 (US CPI, year-over-year).
- Benchmark: the Philadelphia Fed's Survey of Professional Forecasters.
- Result: lower MSE overall, in most individual years, and at almost all horizons.
- A substantive finding, not just a horse-race: LLM forecasts exhibit *slower reversion to the 2% anchor* — i.e. they behave less like a forecaster who imposes the target and more like one extrapolating recent realizations.
- The authors choose PaLM over GPT-4 precisely because its corpus is refreshed, whereas GPT-4's information set was truncated at September 2021 for their purposes.

The authors' own caveat is the important part

- These are described as in-sample conditional forecasts. The evaluation window overlaps the corpus.
- So the honest statement is: *conditional on the prompt's information set, the model produces a distribution whose realised loss is lower than the SPF's over 2019–2023* — with no guarantee that the model was not conditioning on knowledge of the outcome.
- This is not a criticism of the paper, which is explicit about it. It is the reason the next four slides exist.

Related work with the same structure

Nowcasting inflation with LLM-extracted news sentiment; evaluations against FRED-MD benchmarks; synthetic forecaster personas that often outperform human experts. All share the identification problem.

The memorization problem

Lopez-Lira, Tang & Zhu (2025), "The Memorization Problem: Can We Trust LLMs' Economic Forecasts?"

- Before its cut-off, GPT-4o can recall specific S&P 500 index values with perfect precision on certain dates, unemployment rates to a tenth of a percentage point, and precise quarterly GDP figures.
- Recall is selective and seemingly random across data types and dates — the model reconstructs the broad path of indices with occasional large errors.
- Consequence: even accurate pre-cut-off "forecasts" cannot be attributed to predictive ability. The alternative hypothesis is reasoning backwards from a memorised outcome.

The econometric statement

Selective, non-random recall means the contamination is *not* a classical measurement error. It is an omitted-variable problem in which the omitted variable is the realised outcome.

Vintages: a subtler and very economics-specific failure

Crane, Karra & Soto evaluate LLM recall of historical macro variables and release dates:

- Precise knowledge of some recent statistics; performance degrades going further back.
- Two recall errors matter especially for macroeconometrics:
  1. Smoothing across vintages — mixing first prints with subsequent revisions.
  2. Mixing up release dates, so the model's implied information set is not the one available in real time.

Why this is fatal for real-time exercises

The entire real-time forecasting literature (Croushore & Stark; ALFRED vintages) exists because first prints differ from revised data. A model that silently returns revised values while claiming to answer as of date t has broken the exercise before the loss function is computed.

The constructive fix: chronologically consistent models

He, Lv, Manela & Wu (2025), "Chronologically Consistent Large Language Models"

- Train **ChronoBERT** and **ChronoGPT** on timestamped text only: the vintage-t model has seen nothing after t. Leakage is ruled out *by construction*, not by prompt instruction.
- Result 1: chronological consistency does not require sacrificing capability — these models match or surpass BERT and GPT-2 on language comprehension.
- Result 2 (the surprise): in a news-based stock return prediction application, look-ahead bias appears *modest*. The chronologically consistent model performs comparably to Llama 3.1.
- Interpretation offered: understanding the real-time news flow is sufficient for short-horizon return prediction; no disruption of the time continuum is required.
- Follow-up work adds instruction-tuned chronologically consistent models (e.g. a 1999-vintage instruct model, giving two decades of genuine out-of-sample).

Reconciling the two findings

- Look-ahead bias is model- and application-specific.
- Aggregated, low-frequency series (index levels, CPI, GDP) are heavily memorised: they are printed in every corpus. Bias is large.
- High-frequency, cross-sectional, individual-security outcomes are memorised much less: too many of them, too little textual salience. Bias is small.
- Downstream predictive models can also *adapt* to a weaker language model, absorbing some of the loss.
- Practical corollary: the more your target looks like a famous macro aggregate, the more sceptical you should be of a direct LLM forecast.
- Smaller models memorise less — capacity-constrained recall is a feature when you need it.

A protocol for credible LLM forecast evaluation

1. **Prefer chronologically consistent or open-weight models with documented training windows.** If you cannot date the corpus, you cannot date the information set. Prompt-based date restriction (Part II) is a weaker substitute, but at least a testable one.
2. **Evaluate strictly post-cut-off** where feasible. Pre-cut-off performance is descriptive, not predictive.
3. **Run a recall audit.** Ask the model directly for the realized values in your evaluation window. Quantify what it knows before asking what it predicts.
4. **Report a memorization-adjusted subsample** — exclude observations where recall is precise.
5. **Benchmark against something that can win.** AR(p), random walk, the SPF, a shrinkage factor model. Beating a naive benchmark is not news.
6. **Report the loss differential with a proper test** (Diebold–Mariano, Giacomini–White) and account for the fact that model choice was itself data-dependent.
7. **Pin the model version, date, temperature, and seed.** These are now part of the empirical specification.

I. Text as Data: LLMs as Measurement Instruments

II. AI-Based Expectations

III. Simulated Economic Agents

IV. Forecasting

V. Discovery and Causal Inference

VI. The Econometrics of AI-Generated Variables

VII. Research Practice and the Profession

Machine learning as a tool for hypothesis *generation*

Ludwig & Mullainathan, QJE 139(2), 2024, 751–827

- Hypothesis *testing* is highly formalised; hypothesis *generation* is not. They propose a procedure that uses algorithms to notice patterns people would not.
- Application: judicial detention decisions. A deep model predicting detention finds that the defendant's mugshot alone explains one-quarter to nearly one-half of the predictable variation in detention.
- The methodological core: black-box predictors are not interpretable, so they build a pipeline in which generative models produce *counterfactual* faces and human subjects interact with the algorithm to extract an interpretable hypothesis.
- Discipline: the hypothesis is generated on one sample and tested on another. Generation is exploratory; testing remains confirmatory.

Adjacent uses of AI in the discovery stage

- **Instrument search.** Han (2024), "Mining Causality: AI-Assisted Search for Instrumental Variables" — LLM-assisted generation of candidate exclusion arguments, to be evaluated by the researcher.
- **Mapping the literature's causal claims.** Garg & Fetzer (2025), "Causal Claims in Economics": extract directed causal links from tens of thousands of papers, record method, sample size, significance, and data availability, and map free-text variables to JEL codes.
- **Outcomes learned from text.** LLM-based tools that form candidate themes distinguishing treatment from control documents and *statistically validate them on held-out documents*.
- **Structure-aware simulation.** Manning, Zhu & Horton: structural causal models to generate and screen hypotheses in silico before fielding.

The discipline problem

The garden of forking paths, industrialized

An LLM can propose 500 plausible mechanisms, 200 candidate instruments, and 50 heterogeneity dimensions before lunch. Without a pre-committed testing protocol, the effective number of comparisons is unbounded and unrecorded.

- **Sample splitting is non-negotiable.** Generation set, testing set, and — if the procedure is iterative — a third holdout.
- **Pre-registration should cover the search procedure**, not just the final specification: which model, which prompt family, how many candidates, what selection rule.
- **Report the funnel.** How many hypotheses were generated, how many screened, how many tested.
- Where the search is genuinely adaptive, use inference that accounts for it (selective inference; sample-splitting-based post-selection methods).

I. Text as Data: LLMs as Measurement Instruments

II. AI-Based Expectations

III. Simulated Economic Agents

IV. Forecasting

V. Discovery and Causal Inference

VI. The Econometrics of AI-Generated Variables

VII. Research Practice and the Profession

The situation we are actually in

Almost every applied paper in Parts I–V follows the same two steps:

1. Generate $\hat{\theta}_i$ — an estimate of a latent economic variable θ_i — with an AI/ML procedure applied to text, images, or survey-like prompts.
2. Plug $\hat{\theta}_i$ into a downstream econometric model as if it were data.

Familiar examples of step 1

Policy uncertainty from newspapers (Baker–Bloom–Davis); CEO behaviour from diaries (Bandiera–Prat–Hansen–Sadun); remote work from job postings (Hansen et al.); labour-market experience from CVs (Vafa et al.); patient health from surveys (Einav et al.); firm characteristics from holdings (Gabaix et al.).

The question is not whether the classifier is accurate. It is whether $\hat{\beta}$ is consistent and whether the reported standard error means anything.

Why plug-in fails

LLM errors are not classical. They are systematically related to features of the document and, through those features, to the outcome.

- Errors correlate with document length, language, jargon density, era, and firm size.
- Errors correlate with the *class prevalence* the model expects — models regress toward the modal label.
- In the silicon-sample case, errors correlate with subgroup membership: worse on socially sensitive items, worse in the tails.

Consequence

Attenuation is no longer the worst case: the bias can go in *either* direction, and can flip signs, which is exactly what Bisbee et al. document for regression coefficients estimated on synthetic respondents.

Battaglia, Christensen, Hansen & Sacher (2024/25)

"Inference for Regression with Variables Generated by AI or Machine Learning"

- Show *theoretically and empirically* that naively treating AI/ML-generated variables as data yields biased estimates and invalid inference.
- Two proposed remedies:
  1. An explicit bias correction with bias-corrected confidence intervals.
  2. Joint estimation (joint MLE) of the regression parameters and the latent variables — i.e. do not condition on $\hat{\theta}$; integrate over it.
- Applications: label imputation, dimensionality reduction, and index construction via classification and aggregation. Across these, the common approach generates substantial bias; both corrections perform well.

[arXiv:2402.15585; CEPR DP19115; Cowles Discussion Paper 2421]

Their sharpest observation

Most corrections in the statistics and political science literature assume a **validation subsample** in which $(Y_i, \theta_i, \hat{\theta}_i)$ are all observed.

But then...

...you could have estimated the model on the validation sample alone. The AI-generated variable is only buying efficiency, not identification.

And — decisively for the applications economists care about — a validation sample is *not available* when θ_i is genuinely latent: uncertainty, sentiment, hawkishness, managerial attention. There is no gold-standard measurement of the tone of a Beige Book. This is why joint estimation, which puts a model on the measurement process rather than requiring ground truth, is the more general answer.

Ludwig, Mullainathan & Rambachan (2025): the framework

Two canonical uses, two different conditions.

Prediction tasks

Using LLM output to predict some economic outcome (missing-data imputation, feature generation) is valid under one condition: no "leakage" between the LLM's training data and the researcher's sample.

Operational guarantee: use *open-source models with documented training data and published weights*.

Estimation tasks

Using LLM output to automate *measurement* of an economic concept requires the researcher to collect at least some validation data. Without it, the errors of the automation cannot be assessed or accounted for.

With these steps, LLM outputs can be used with the familiar econometric guarantees. Their two applications — finance and political economy — find the requirements *stringent*, and that violating them is consequential.

Where this connects to what we already know

AI-era problem	Classical antecedent
Plug-in $\hat{\theta}$ from a classifier	Errors-in-variables; misclassified binary regressors [Aigner 1973]
Two-step with estimated first stage	Generated regressors [Pagan, IER 1984; Murphy & Topel 1985]
Validation subsample corrections	Measurement error with auxiliary data [Chen, Hong & Tamer, ReStud 2005; Lee & Sepanski 1995]
ML predictions as covariates	[Fong & Tyler, Political Analysis 2021]
Cheap synthetic labels + few real ones	Prediction-powered inference [Angelopoulos et al., *Science* 2023; Zrnic & Candès]; design-based supervised learning [Egami et al.]
Surrogate outcomes from a model	Surrogate index [Athey, Chetty, Imbens & Kang]

Nothing here is new econometrics. What is new is the volume of papers that skip it.

Designing the validation sample

If you must hand-label, do it properly

- **Stratify** on the covariates that drive both the classifier error and the outcome (length, sector, era, language). Simple random sampling wastes labels where the model is already reliable.
- **Power the validation sample for the correction, not for the accuracy statistic.** The relevant precision is that of the estimated misclassification matrix, which enters the bias correction nonlinearly.
- **Use multiple human coders** and report inter-coder agreement. The "gold standard" is itself an estimate: if humans agree only 88% of the time, no sharper benchmark exists.
- **Label blind to the model's output**, or you create a correlated-error problem inside the validation set.
- **Report both the naive and corrected estimates.** The gap is informative and reviewers will want it.

A minimal reporting standard for AI-generated variables

1. Model name, version string, access date, temperature, seed, and whether weights are open.
2. The full prompt, verbatim, in an appendix. It is the instrument.
3. Validation sample: size, sampling design, coder protocol, agreement rates, confusion matrix.
4. Bias-corrected estimates alongside naive plug-in.
5. Sensitivity to model choice — at minimum one alternative model, ideally one open-weight.
6. An explicit leakage argument: what is the training cut-off, what is the evaluation window, and why do they not overlap?
7. Code and intermediate outputs $(\hat{\theta}_i)$, not merely the final dataset — API outputs are not reproducible by re-running the prompt.

The research process itself

Korinek (JEL 61(4), 2023) catalogues use cases across the research workflow — ideation and feedback, writing, background research, coding, data analysis, mathematical derivations — with periodic updates as capabilities change.

- Empirically, adoption is measurable: recent work estimates LLM penetration in economics writing from distinctive linguistic markers, against a backdrop of heterogeneous journal disclosure requirements.
- Uncontroversial gains: code translation and debugging, literature triage, first-pass referee-report structuring, translation, LaTeX and figure work.
- Contested: idea generation, and anything where fluent wrongness is costly — derivations, citations, and summaries of papers the reader will not check.

Reproducibility is the binding constraint

The problem

A closed API endpoint is a moving target. The model behind a given name changes; deprecations remove it entirely; temperature and system prompts are not always under user control. An empirical result conditioned on "GPT-4, June 2024" may be literally unreproducible by 2026.

- **Mitigations:** open-weight models with documented corpora; archive raw model outputs as data; version-pin; report across-run dispersion from repeated sampling.
- **For journals:** treat the model as part of the data-generating process in the data availability policy. The intermediate outputs must be deposited.
- **For referees:** ask for the prompt, the validation design, and the leakage argument. In practice these three questions separate careful papers from the rest quickly.

Where I think the frontier actually is

1. **Chronologically consistent model families as research infrastructure.** A public series of vintage-dated models would do for text-based macro-finance what ALFRED did for real-time data.
2. **Identification results for simulated agents.** Conditions under which LLM-population moments identify human-population moments; transportability; exchangeability of tasks.
3. **Joint estimation as the default.** Measurement models estimated jointly with structural parameters, rather than two-step plug-in.
4. **Measurement of concepts that were previously unmeasurable** — 140 years of central bank communication, the universe of job postings, the full text of the patent record — combined with credible identification. This is where the citations will be in five years.
5. **Honest negative results.** We need more papers reporting where LLM measurement failed and why.

Summary

- **Measurement** is the mature application. It works, it scales, and the payoff is the panel it creates.
- **Expectations** is the most intellectually interesting: a laboratory for belief formation, if generalisation beyond the training window holds up.
- **Agents** are useful ex ante and dangerous ex post. Means may transfer; dispersion does not.
- **Forecasting** is the area where published results should be discounted most heavily, until leakage is ruled out by construction rather than by prompt.
- **The econometrics** is classical: generated regressors, errors-in-variables, auxiliary-data corrections. The remedies exist. They are simply not being applied.

The right posture is neither enthusiasm nor dismissal, but the ordinary demand for an identification argument.

Thank you!