Expanding Economic Data:
The Power of Data Science, Machine Learning, and AI

Michael McMahon[1]

ESCoE Conference Keynote, May 2026

[1]University of Oxford and CEPR

About me

- Michael McMahon
  - My focus is macroeconomics and monetary economics

$\Rightarrow$ Accidentally fell into Data Science via NLP, Text Analysis and Narrative Economics

- Economics:
  - ERC-funded research focuses on *the effects of CB communication*
- Data Science:
  - DSTL-ATI funded research on *Narrative Detection & Tracking with Multimodal Language and Graph Models*

AI for Economics Researchers

• Using AI in Your Research

- **Code AI as an assistant:** review, test, and own all generated code
- **Validate first:** build a labelled gold standard and benchmark before scaling
- **Document prompts:** treat prompts as part of your methodology (version-control them)
- **Check for hallucinations:** verify any factual claim the LLM makes in your pipeline
- **Reproducibility:** set seeds; save model versions; report model name and date

⋆ Studying AI and Its Impacts

- Identification is still paramount: AI creates variation but endogeneity problems remain
- Exploit **quasi-experiments**: rollout timing, capability thresholds, industry exposure
- Track **complementary investments**: organisational change, training, etc
- Watch for **general equilibrium** effects: partial equilibrium may mislead
- Distinguish **adoption** from **impact**: usage $\neq$ productivity gain

I will focus on AI as a tool in our research.

Key Messages

My view on AI for Economics Research

AI is a powerful set of tools that can augment economic research...

Key Messages

My view on AI for Economics Research

AI is a powerful set of tools that can augment economic research...

but it is not a substitute for economic reasoning and careful empirics.

Key Messages

My view on AI for Economics Research

AI is a powerful set of tools that can augment economic research...

but it is not a substitute for economic reasoning and careful empirics.

- Embrace it
- Leverage it
- Deploy it to improve YOUR research

Key Messages

My view on AI for Economics Research

AI IS A POWERFUL SET OF TOOLS THAT CAN AUGMENT ECONOMIC RESEARCH...

BUT IT IS NOT A SUBSTITUTE FOR ECONOMIC REASONING AND CAREFUL EMPIRICS.

- Embrace it
- Leverage it
- Deploy it to improve YOUR research

If you're going to do it, do it right!

Econ needs to be a bit more like NLP—and beware of domain-familiarity effects with LLMs!

Outline

1. What is AI and how it affects economics research?
2. Benchmarking for Economic Measurement
3. LLMs for Narrative Analysis

What Is Artificial Intelligence?

What Is Artificial Intelligence?

A Working Definition

Artificial Intelligence refers to computational systems that perform tasks which, if performed by a human, would be said to require intelligence — such as perceiving, reasoning, learning, communicating, and acting purposefully.

Key features of the definition:

- **Broad** — covers many techniques and applications
- **Task-oriented** — defined by capability, not mechanism
- **Evolving** — the boundary shifts as technology advances
- **Pragmatic** — no requirement for "consciousness"

Historical Roots

The term was coined at the **1956 Dartmouth Conference**. Early AI used explicit rules; modern AI predominantly learns from data.

The "AI Effect"

Once a problem is solved, it is often no longer called AI — OCR, spam filtering, and chess engines were all once frontier AI.

Narrow AI vs. Artificial General Intelligence (AGI)

• Narrow AI

Optimised for **one class of tasks**

- Image classification
- Language translation
- Fraud detection
- Playing Go or chess

✓ **This is the AI that exists today**

⋆ AGI

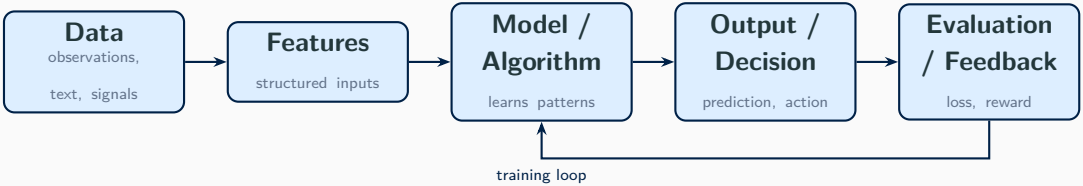
Human ability **across any cognitive task**

- Transfer knowledge across domains
- Reason under true novelty
- Plan long multi-step sequences
- Acquire new skills autonomously

○ **Does not yet exist; timeline disputed**

- For economics research, **narrow AI** is the relevant category
- Though modern Large Language Models blur the distinction.

Key Components of AI Systems



Data

The raw material. Can be text, numbers, images, audio. Quality and quantity both matter enormously.

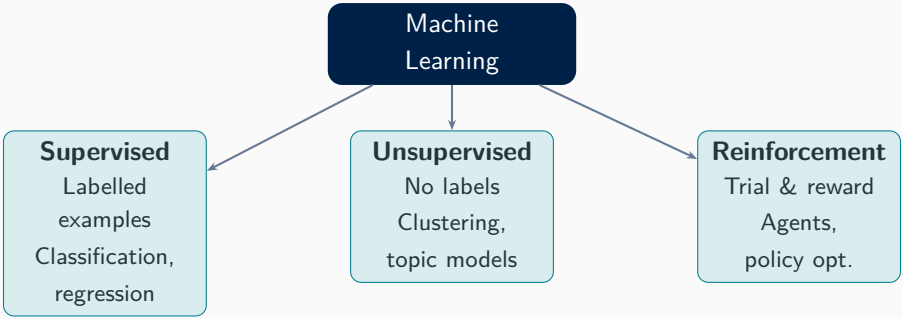
Model

A parameterised function that maps inputs to outputs. Parameters are learned to minimise prediction error.

Evaluation

A loss or reward signal that tells the model how well it is doing. Defines what we are optimising for.

Types of Machine Learning



Supervised examples:
Predicting GDP growth, classifying news sentiment, estimating treatment effects

Unsupervised examples:
Topic modelling FOMC minutes, clustering firms by disclosure language

RL examples:
Optimal monetary policy under uncertainty, algorithmic trading strategies

Deep Learning: The Engine of Modern AI

What makes it "deep"?

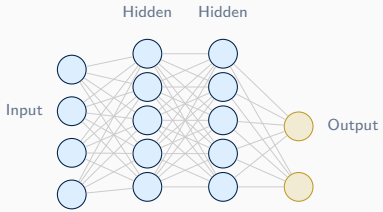
Neural networks with many layers of interconnected units, each learning increasingly abstract representations of the data.

Why did it take off?

- **Scale:** large labelled datasets (internet, digitisation)
- **Compute:** GPUs enabling parallel matrix operations
- **Algorithms:** backpropagation + better architectures

Key architectures:

- **CNN** — images, spatial data
- **RNN / LSTM** — sequences, time series
- **Transformer** — text, audio, multimodal



Foundation Models and Large Language Models (LLMs)

What is a Foundation Model?

A model trained at large scale on broad data that can be **adapted** (fine-tuned or prompted) to a wide range of downstream tasks — without retraining from scratch.

LLMs specifically:

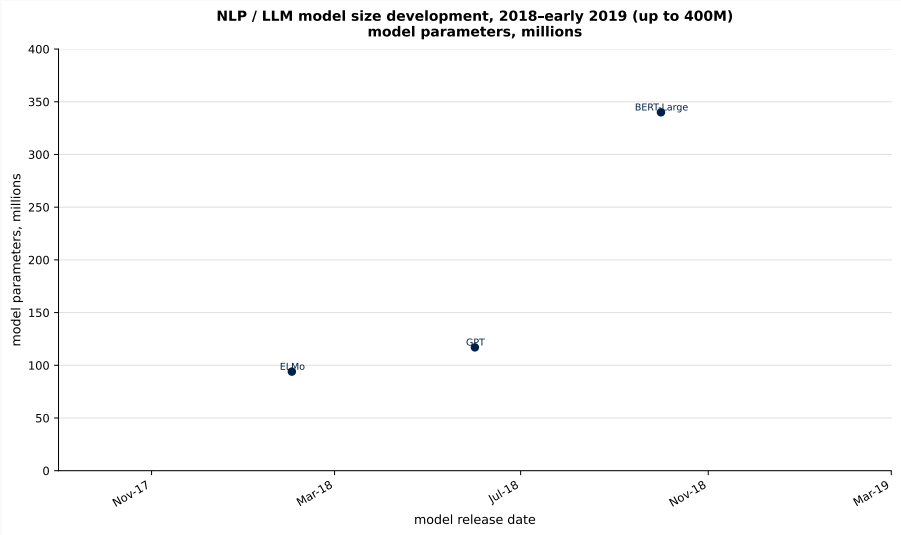
- Trained to predict the next token in a sequence
- Emergent abilities: reasoning, code, translation, Q&A
- Follow instructions via **RLHF** (Reinforcement Learning from Human Feedback)
- Examples: GPT-4, Claude, Gemini, Llama, Mistral

Why this matters for economics:

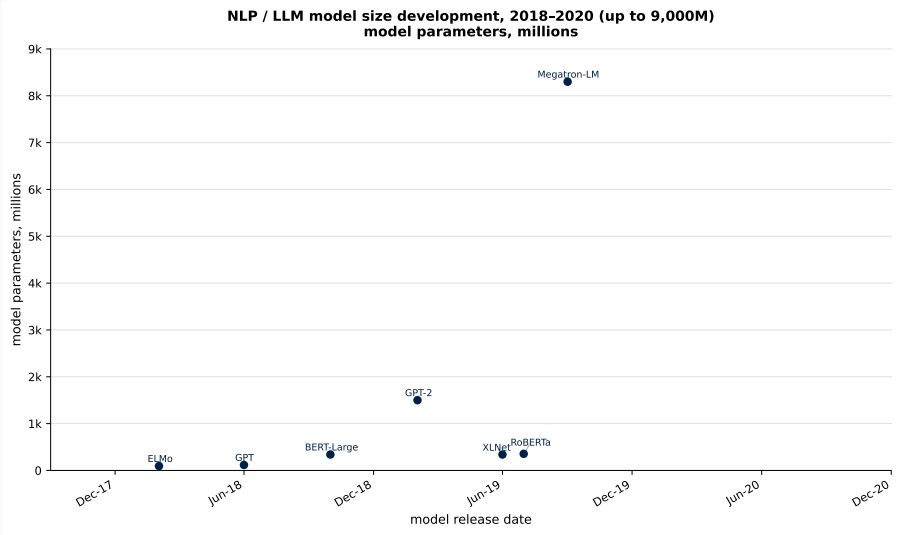
- Process and classify **unstructured text** at scale
- Automate coding, data wrangling, literature review
- Generate hypotheses and summarise evidence
- Interface for interacting with quantitative tools
- Object of study: AI as a **general-purpose technology**

The Transformer architecture (Vaswani et al., 2017) underpins virtually all modern LLMs via the **self-attention** mechanism.

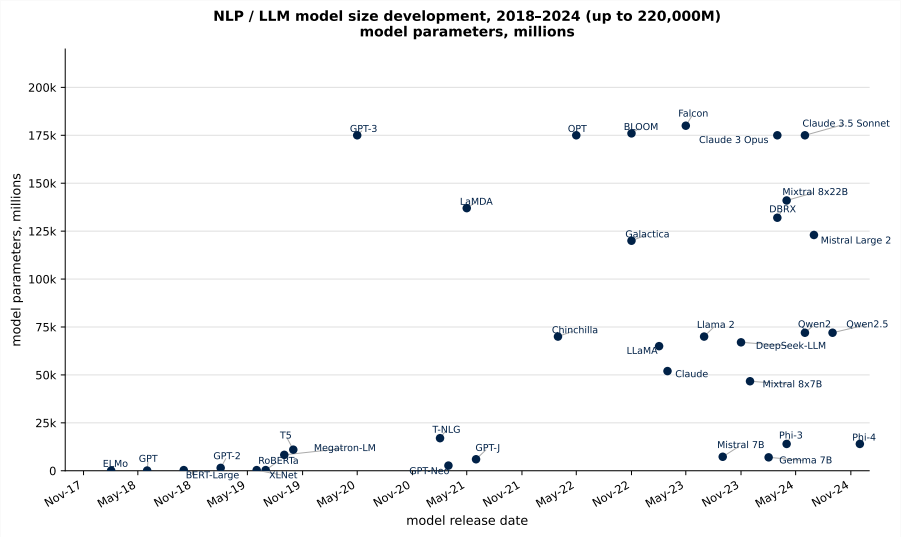
Scale matters



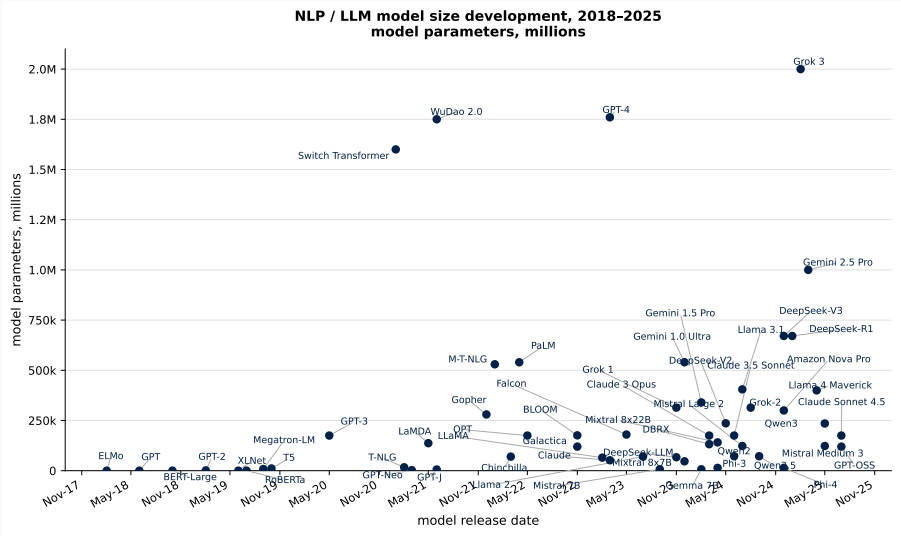
Scale matters



Scale matters



Scale matters



AI as Economics Research Tools

AI as a Research Tool: A Taxonomy

Stage	AI Capability	Example
Data collection	Web scraping, OCR, speech-to-text	Digitising historical records, firm filings
Data labelling	Classification, sentiment, NER	Coding FOMC transcripts, news sentiment
Feature engineering	Embeddings, topic models	Textual similarity, thematic indices
Estimation	Prediction, causal ML	Double/debiased ML, high-dimensional stats
Coding & workflow	Code generation, debugging	Python/Stata/R from natural language
Writing & review	Drafting, summarising, translation	Literature review, referee reports

REMINDER: AI tools augment, but do not replace, economic reasoning and causal identification.

Natural Language Processing (NLP) in Economics

Why text data matters:

- Vast quantities of economic-relevant text exist: central bank communications, earnings calls, news, policy documents
- Text encodes **beliefs, expectations, and intentions** that numeric data may not capture

Key NLP tools:

- **Dictionaries:** LM, Harvard IV-4
- **Topic models:** LDA, structural topic models
- **Pretrained models:** BERT, FinBERT, FLANG-BERT
- **LLMs via prompting:** GPT-4, Claude

Prominent applications

- Measuring policy **uncertainty** (Baker, Bloom & Davis)
- Classifying central bank **tone** and forward guidance
- Analysing firm-level **earnings call** language
- Parsing congressional speech for **political economy**

Benchmarking matters

Always evaluate on labelled economics data before deploying at scale.

Causal Inference Meets Machine Learning

Traditional ML optimises **prediction** accuracy.
Economics cares about **causal effects**.

The synergy:

- ML handles **high-dimensional** controls and nuisance functions
- Econometrics provides **identification** and inference framework

Key methods:

- **Double/Debiased ML** (Chernozhukov et al., 2018)
- **Causal Forests** / Heterogeneous treatment effects (Wager & Athey)
- **LASSO-based IV** selection
- **Synthetic Control** enhancements

Key principle

A prediction algorithm that fits data well does **not** estimate a causal parameter. Identification assumptions must still be defended on economic grounds.

Double ML — intuition

Partial out confounders using ML, then run a "clean" regression on residuals. Prediction power of ML + valid inference from cross-fitting.

Limitations and Critical Considerations

△ Technical Limitations

- **Hallucination:** confident generation of false facts
- **Distribution shift:** poor out-of-sample performance
- **Spurious correlations:** may not generalise causally
- **Opaqueness:** "black box" limits interpretability
- **Context length:** limits on long-document reasoning

⊕ Social & Ethical Concerns

- **Bias:** reflects and can amplify training data biases
- **Privacy:** data used may be sensitive or protected
- **Reproducibility:** stochastic outputs complicate replication
- **Academic integrity:** attribution and verification norms evolving
- **Concentration:** high compute costs favour incumbents

REMINDER: **Validate!** Off-the-shelf accuracy does not transfer automatically across domains.

Some useful resources

1. Chris Blattman: `claudeblattman.com`
2. Markus' Academy with Paul Goldsmith-Pinkham: `https://youtube.com/playlist?list=PLPKR-Xs1slgTqMU3E1UJszSK4PYRFF-Py&si=6vkQica8KY59iAwy`
3. Antonio Melé: `https://meleantonio.github.io/awesome-econ-ai-stuff/`

Benchmarking within Economics Research

Different focus of the research

Economics use of Data Science

AI/ML/NLP typically is a measurement device that improves the inputs that we can use. Often, a new measure is shown to have some external validation, but there is not the effort to find "best fit" in the measurement phase.

- Methods used are often not complex
- General preference, at least in applied monetary economics, for transparency
- Contribution is rarely the data science!

Different focus of NLP research

Data Science Research

- The focus is on showing how, in a downstream task, the algorithm improves on our ability to model / understand the language.
- Train and Test use of the data set is fundamental.
- There are explicit measures for assessing fit.

Bringing the Data Science Approach to Economics

Idea of Ahrens, Gorduza and McMahon (2026)

Compare the performance of different NLP approaches across a series of downstream tasks using typical economics or financial datasets.

1. What is a typical economics/finance dataset?
2. Evaluation metrics

Overall Take Away

Econ needs to be a bit more like Data Science!

Classification and the Confusion Matrix

Confusion Matrix:

		Actual	
		0	**1**
Predicted	**0**	tn	fn
	1	fp	tp

- Define the following predictive outcomes from the downstream task:
  - tp $\equiv$ true positives
  - fp $\equiv$ false positives
  - fn $\equiv$ false negatives
  - tn $\equiv$ true negatives
- Define:

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \qquad \text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

Confusion Matrix Example I

- Actual data is 30 negative and 70 positive:
- Is the following classifier any good?

  Confusion Matrix:

		Actual	
		0	**1**
Predicted	**0**	20	30
	1	10	40

- Calculate:

$$\text{precision} = \frac{40}{40+10} = \frac{40}{50} = 0.8$$

$$\text{recall} = \frac{40}{40+30} = \frac{40}{70} = 0.57$$

Confusion Matrix Example II

- What abour this 2nd classifier?
  Confusion Matrix:

		Actual	
		0	**1**
Predicted	**0**	0	0
	1	30	70

- Calculate:

$$\text{precision} = \frac{70}{70+30} = \frac{70}{100} = 0.7$$

$$\text{recall} = \frac{70}{70+0} = \frac{70}{70} = 1$$

Confusion Matrix Example III

- What about this 3rd classifier?
  Confusion Matrix:

		Actual	
		0	**1**
Predicted	**0**	25	5
	1	5	65

- Calculate:

$$\text{precision} = \frac{65}{65+5} = \frac{65}{70} = 0.93$$

$$\text{recall} = \frac{65}{65+5} = \frac{65}{70} = 0.93$$

A Popular Evaluation Metric

- The F1 score is the harmonised mean over precision and recall:

$$\text{F1} = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$$

- The F1 score is widely used; especially useful in unbalanced datasets.
  - Multilabel classification: use (unweighted) mean over the per-class F1 scores.
- Across our example classifiers:
  1. Classifier 1: 0.667
  2. Classifier 2: 0.8235
  3. Classifier 3: 0.9286

Idea of the paper

Bringing the Data Science Approach to Economic Data Science

Compare the performance of different NLP approaches across a series of downstream tasks using typical economics or financial datasets.

1. The unique features of typical economics/finance datasets
   - The datasets that we use
2. NLP approaches we consider
3. Results

Typical Data

Some key features of text sources in the economics and finance domain:

1. *Small*:
   - Often contain only hundreds or thousands of data points compared to millions or billions of data points in 'classic' NLP datasets.
2. *Long text sequences*
   - Mean sequence lengths can often be as large as 3,000 while maximum sequences can get as high as 13,000 words.
3. *Relevant tabular metadata*
   - This can be vital to gauge the context in which a statement is to be interpreted.
   - Think about a suggestion that inflation next year will increase by 2pp...

Data we use

1. The Financial Phrase Bank (FPB) dataset
   - Malo et al (2014)
   - Contains sentences from financial news that are annotated as reflecting:
     - Positive sentiment
     - Neutral sentiment
     - Negative sentiment
   - We use only sentences that have 100% annotator agreement.

Data we use

1. The Financial Phrase Bank (FPB) dataset
2. The Twitter Financial News (TFN) dataset
    - Contains annotated finance-related tweets classified as:
        - Bearish
        - Neutral
        - Bullish

Data we use

1. The Financial Phrase Bank (FPB) dataset
2. The Twitter Financial News (TFN) dataset
3. The FOMC Bluebook Alternatives (FBA) dataset
    - Contains FOMC statement alternatives and labelled with their implied fed funds rate (FFR).
    - The original labels are discrete. We categorise the label as:
        - Negative if the FOMC decision leads to a decrease in the FFR
        - Neutral if the FFR is unchanged
        - Positive if the FFR has been increased

Data we use

1. The Financial Phrase Bank (FPB) dataset
2. The Twitter Financial News (TFN) dataset
3. The FOMC Bluebook Alternatives (FBA) dataset
4. The FOMC Greenbook dataset
   - contains the Greenbook paragraphs about inflation (consumer price index) labelled with the associated one-quarter-ahead Greenbook forecast
   - We define the forecast as:
     - Increasing
     - Unchanged
     - Decreasing
   - The dataset also contains tabular metadata which consists of the one-quarter-ahead forecasts on CPI, GDP growth, and unemployment of the previous FOMC meeting.

Data we use

1. The Financial Phrase Bank (FPB) dataset
2. The Twitter Financial News (TFN) dataset
3. The FOMC Bluebook Alternatives (FBA) dataset
4. The FOMC Greenbook dataset

dataset	type	total	train	val	test	neg	neut	pos	mean len	max len	min len
Fin Phrase Bank	text	2,264	60%	15%	25%	13%	61%	25%	122	315	9
Fin. Twitter	text	11,931	64%	16%	20%	15%	20%	65%	86	227	2
Bluebooks	text	418	64%	16%	20%	6%	75%	17%	2,716	5,934	666
Greenbooks	text+tab	144	64%	16%	20%	48%	6%	47%	3,940	13,063	292

Models: Dictionary Models I

- Representative dictionary: We use the Loughran & McDonald (JoF, 2011) financial dictionary (LMdict)
- LMdict provides hand-crafted word lists along seven categories:
  1. negative
  2. positive
  3. uncertain
  4. litigious
  5. strong modal
  6. weak modal
  7. constraining.
- Typical represented as % of text with each category.
- Often only some categories are used.

Models: Dictionary Models II

1. **LMdict-naive**
   - Only uses the positive (%*positive*) and negative (%*negative*) categories.
   - Doesn't learn any parameter weights in a data-driven way.
   - A sequence is classified as negative, neutral, or positive, according to naive thresholds:

$$y = \begin{cases} 0, & \frac{\%negative}{\%negative+\%positive} \geq 0.75 \\ 1, & 0.75 < \frac{\%negative}{\%negative+\%positive} > 0.25 \\ 2, & \frac{\%negative}{\%negative+\%positive} \leq 0.25 \end{cases} \quad (1)$$

2. **LMdict-lin**
3. **LMdict-nonlin**

Models: Dictionary Models II

1. **LMdict-naive**
2. **LMdict-lin**
   - Use the seven word list features from the LMdict: *% negative*, *% positive*, *% uncertainty*, *% litigious*, *% strong modal*, *% weak modal*, *% constraining*.
   - Add the additional features *# of words*, *# of digits*, *# of numbers*
   - Feed these features into a logistic regression to learn the parameter weights for each feature given the training dataset.
3. **LMdict-nonlin**

Models: Dictionary Models II

1. **LMdict-naive**
2. **LMdict-lin**
3. **LMdict-nonlin**
   - Use the same feature set as in LMdict-lin, but instead of a logistic regression, we fit a model zoo of non-linear machine learning methods on the respective training dataset.
   - Use the AutoGluon (Erickson et al, 2020) machine learning model zoo.

NLP Approaches: Word Count Models

- Extract the word counts from the input text sequences and represent them in a document-term-matrix (DTM).
- Word counts are weighted by their tf-idf score.
- The tf-idf scaled word counts then serve as features in a linear and a non-linear classification model:
  - **WordCount-lin**
    Use a logistic regression model that uses elastic net regularization. The ratio between LASSO and ridge regularization is a hyperparameter that is being tuned via cross-validation in the model training phase.
  - **WordCount-nonlin**
    Use the same feature set but fit a model zoo of non-linear machine learning methods on the respective training dataset. We use the AutoGluon machine learning model zoo.

NLP Approaches: Topic Models

- Fit unsupervised latent Dirichlet allocation (LDA) models.
- Estimate LDA models with $K = [5, 10, 50, 100, 250, 500, 1000]$ topics.
- We then use the estimated topic parameters as features in a logistic regression.

NLP Approaches: Transformer Models

- **BERT-base**
  The standard BERT-base model which has not been further pretrained for financial domain applications.
- **FinBERT**
  A BERT model that has been further pre-trained on a financial corpus containing a subset of the Thomson Reuters Text Research Collection (TRC2), and then been fine-tuned on a subset of the Financial Phrase Bank dataset.
- **FLANG BERT**
  FLANG-BERT has been pre-trained on general English corpora from Wikipedia and BookCorpus as well as financial corpora from SEC EDGAR, Reuters, Bloomberg, Seeking Alpha, and Investopedia. Furthermore, the model makes use of preferential token masking for financial terms.

NLP Approaches: Multimodal Transformer Models

- **Tab+:** We use the TabularPredictor model architecture from AutoGluon to fuse tabular data with text data. This yields the same models as described above, but with a multimodal fusion capability. Such models are labelled with a "Tab+" prefix.
- **Multimodal Transformer**: We add one additional model, which we name Multimodal Transformer. This model uses the MultimodalPredictor model architecture instead of TabularPredictor. The difference is that this approach directly fuses multiple neural network models for the different modalities.

NLP Approaches: Large Language Models

- **GPT-4o** and **GPT-4o-mini**: State-of-the-art instruction-tuned LLMs evaluated in a *zero-shot* setting—no task-specific fine-tuning on our benchmark datasets.
- Reflects how economists typically deploy LLMs: off-the-shelf, without retraining.
- To assess sensitivity to **prompt design**, we test 10 systematically varied prompt formulations per dataset:
  - Minimal direct instruction
  - Expert financial analyst persona
  - Chain-of-thought reasoning
  - Conversational / question-answer style
- We report the **best** and **worst** prompt, bracketing the range achievable through prompt engineering alone.

LLM Prompt Examples I

Prompt 1 — Direct / Strict (original)

```
You are GPT-4.  Your task is to read the provided text input and select the single
output option that best fits the input.  The only valid responses are the options
provided below.  Do not include any additional text, explanation, or formatting--respond
with exactly one of the options.

Text Input:

{phrase}

Labelling strategy:

{labelling_strategy}

Only chose one out of the following output options, your result must be a single
integer:

{outputs_nlabels}
```

LLM Prompt Examples II

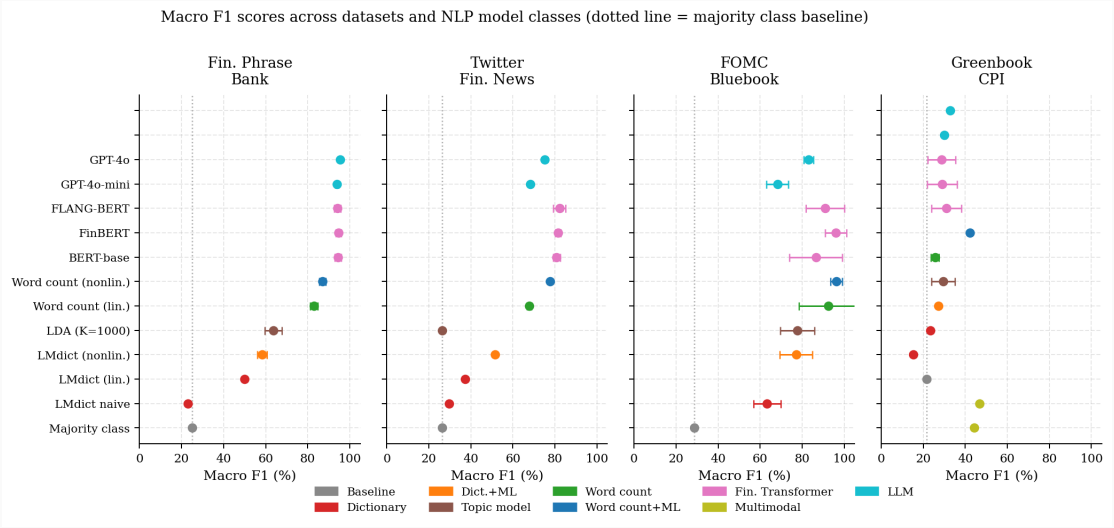
Prompt 2 — Expert Financial Analyst Persona

```
You are an expert financial analyst with decades of experience in market sentiment
analysis. Analyze the following text carefully and classify it according to the
provided criteria.

Text to analyze:

{phrase}

Classification criteria:

{labelling_strategy}

Available categories: {outputs_nlabels}

Provide only the single numeric label that best represents your classification.

Do not include any explanation or additional text.
```

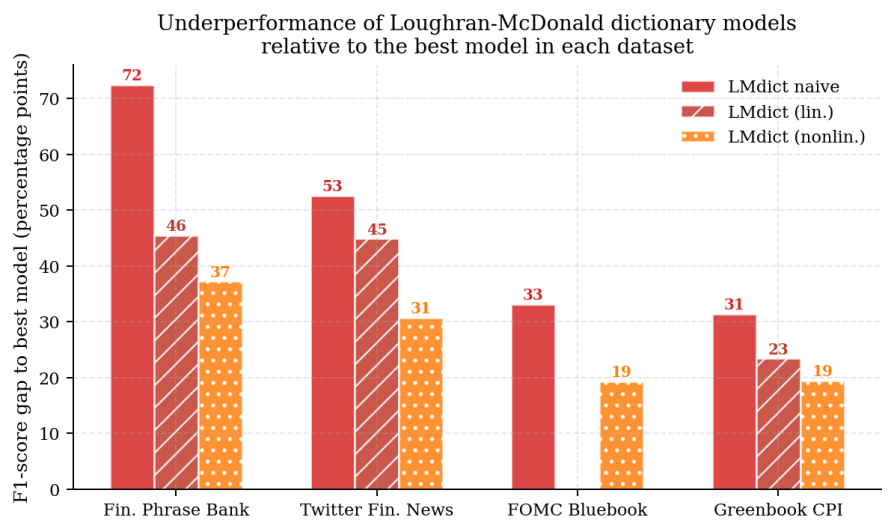
Results

1. Financial transformer models perform best on text-only classification datasets.
2. Financial transformer models struggle on multimodal economics and finance datasets
   - This highlights more need for concerted research efforts on fusing multimodal data and pre-training/fine-tuning models for tasks in these domains.
3. Relatively simple word count models do very well across all datasets.
4. Unsupervised topic models show lack in performance across all datasets.
5. The Loughran-McDonald dictionary underperforms across all evaluated datasets.
6. GPT models do best where they are more likely familiar with the type of language before.
   - LLMs perform strongly on familiar datasets but poorly on specialised economics data.
7. There is variation in LLM performance across prompts.

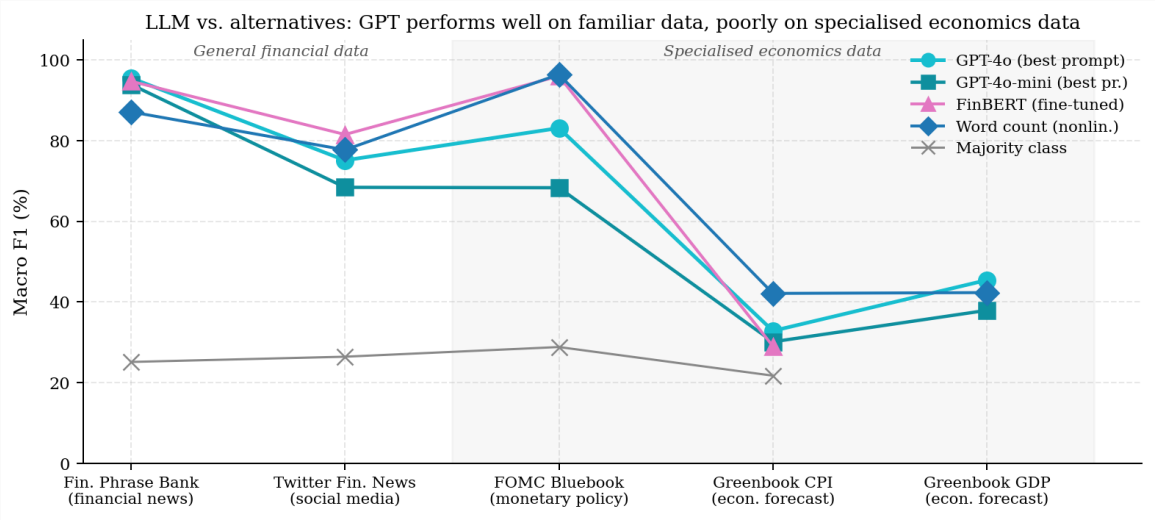
Results 1–4: Overview Across All Datasets and Model Classes



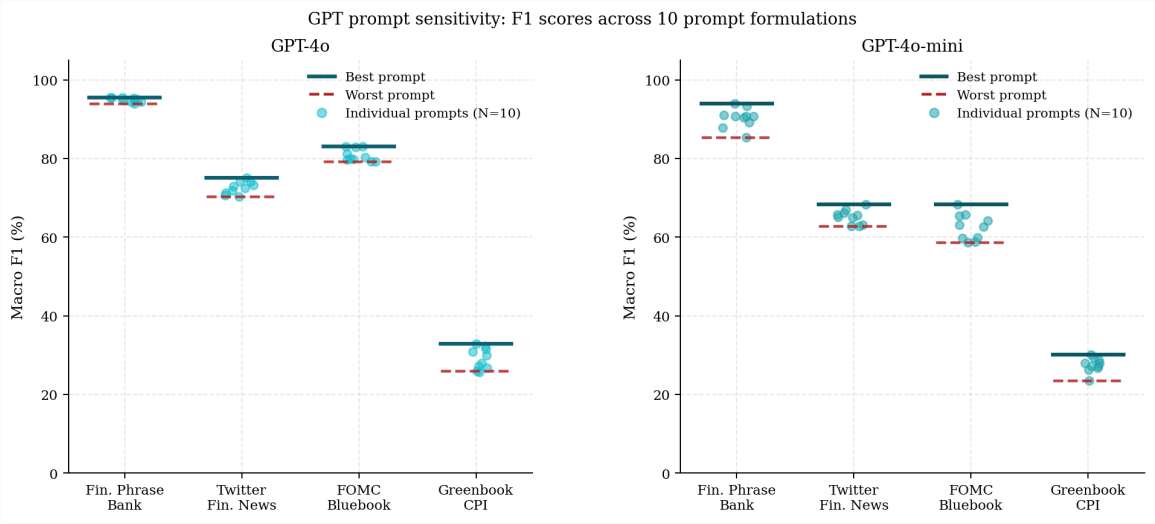
Loughran-McDonald: Underperformance Relative to Best Model



LLM Performance vs. Alternatives



GPT Prompt Sensitivity Across Datasets



Narrative Example: Cieslak et al (2026)

Monetary Policy and Uncertainty

- Policymakers have long viewed uncertainty as central to decision-making
- *"Uncertainty is not just an important feature of the monetary policy landscape; it is the defining characteristic of that landscape... As a consequence, the conduct of monetary policy in the United States has come to involve, at its core, crucial elements of risk management."* (Greenspan 2004)
- *"The Governing Council bases its monetary policy decisions, including the evaluation of the proportionality of its decisions and potential side effects, on an integrated assessment of all relevant factors. In particular, it takes into account not only the most likely path for inflation and the economy but also surrounding risks and uncertainty"* (ECB Strategy Statement 2025)

Research Questions

- Do policymakers respond to risk/uncertainty?
- If so, why?

Research Questions

- Do policymakers respond to risk/uncertainty?
- If so, why?

$\Longrightarrow$ How to measure risk/uncertainty?

FOMC Transcripts and Meeting Structure

- FOMC transcripts are publicly available from 1976 onwards, published with a five-year lag.
  - Verbatim, attributed, (nearly) complete record of policy deliberations.
- Main sample covers October 1982—December 2019, 299 meetings in total.

FOMC Transcripts and Meeting Structure

FOMC meetings follow a regular structure in much of our sample. Core parts:

Discussion of Economic/Financial Conditions (FOMC1)

- Staff presentation on economic conditions
- Q&A on staff presentation
- FOMC member presentations on economic conditions

Discussion of Policy Response (FOMC2)

- Staff presentation on policy alternatives
- Q&A on policy alternatives
- FOMC members sequentially state and justify preferred alternative

Measuring Risk Perceptions at the Individual Level

- Feed each FOMC member's text (from FOMC1) into an LLM (Claude Sonnet 4) and ask for a risk assessment in each of two dimensions: inflation and growth.
- Four categories: Upside, Balanced, Downside, Not Discussed
- Coded view $BoR_{it}^{v} \in \{U, B, D, N\}$ for member i, variable $v \in \{I, G\}$, meeting t.
- Query focuses on risk perceptions *around* individual forecasts, not disagreement with staff.
- LLM reasoning must include direct quotations from the transcript.

Example Quotations (Inflation)

"We can't dismiss the possibility that compensation growth will drift upward, raising core inflation"
Yellen, May 1996

"I would also judge the risk to the inflation outlook as being biased to the downside"
Lockhart, Sep 2010

"The risks to the inflation outlook remain balanced"
Bostic, May 2019

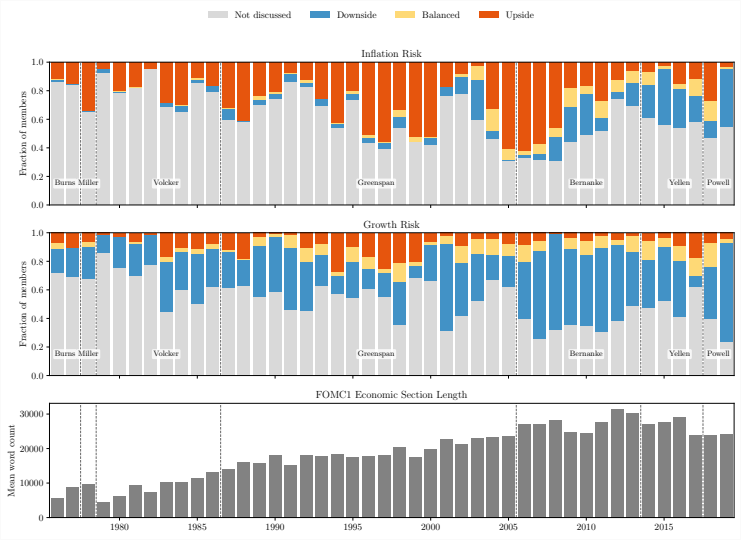
Example Quotations (Growth)

"I'm beginning to think our greatest problem may be an excessive increase in activity"
Partee, Feb 1985

"The risk to satisfactory economic performance... will remain skewed very much to the downside"
Kohn, Mar 2008

"Considerable risks still in place to growth, including gasoline prices, Europe and the deleveraging of European banks, continued stresses and tight credit in the U.S. banking system..."
Bernanke, Mar 2012

Distribution of Speaker-Level Classifications over Time



Meeting-Level Balance-of-Risks Measure

Define counts of FOMC members classified as downside/upside:

$$R^v_{D,t} = \sum_i \mathbb{1}\left(BoR^v_{it} = D\right), \qquad R^v_{U,t} = \sum_i \mathbb{1}\left(BoR^v_{it} = U\right)$$

Aggregate upside & downside measures for variable v:

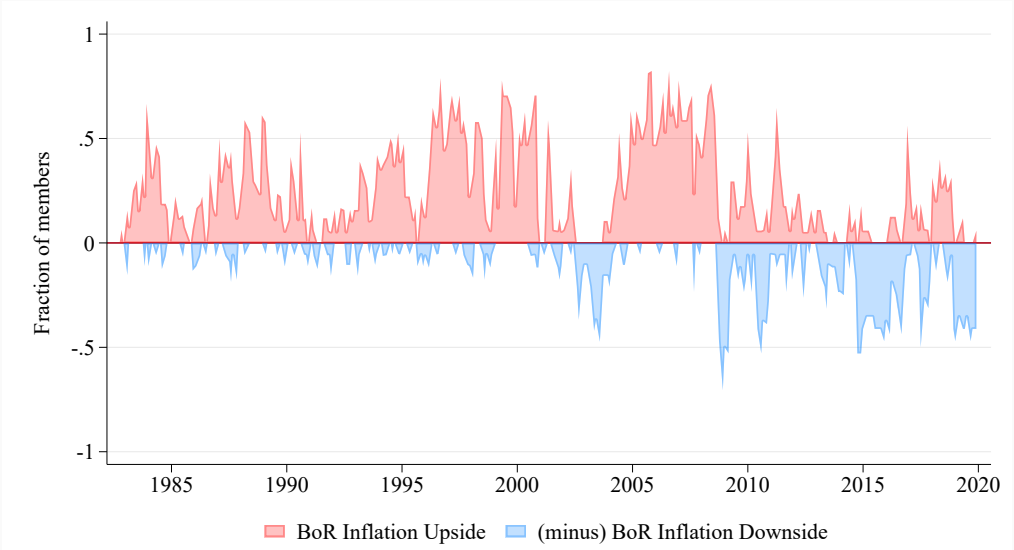
$$BoR(U)^v_t = \frac{R^v_{U,t}}{N_t} \qquad BoR(D)^v_t = \frac{R^v_{D,t}}{N_t}$$

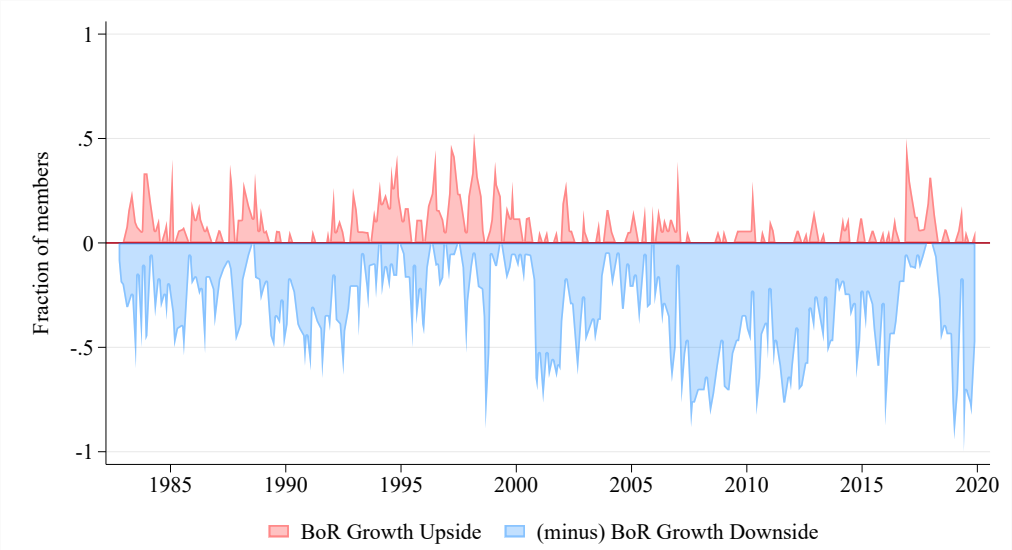
Aggregate skew measure for variable v:

$$BoR^v_t = \frac{R^v_{U,t} - R^v_{D,t}}{N_t}$$

where N_t is the number of FOMC members at meeting t.

Captures skewness in the distribution of individual risk perceptions in any given meeting.


Fraction of members
1
.5
0
-.5
-1
1985
1990
1995
2000
2005
2010
2015
2020
BoR Inflation Upside
(minus) BoR Inflation Downside


1
.5
0
-.5
-1
Fraction of members
1985
1990
1995
2000
2005
2010
2015
2020
BoR Growth Upside
(minus) BoR Growth Downside

Validation I: Dictionary Methods

- We also build dictionaries to measuring members' uncertainty/risk perceptions.
- Policymakers' Uncertainty (PMU) indices for inflation/disinflation and growth, respectively, tabulate the frequency of risk/uncertainty-related terms in sentences with topical keywords.
- Sentiment indices tabulate directional terms in sentences with topical keywords.
- Advantage is simplicity, transparency, replicability.
- Disadvantage is less sophisticated understanding of context.

Validation I: BoR Skew and Dictionary Measures

	BoR Infl Upside	BoR Infl Downside	BoR Infl Skew	BoR Grow Upside	BoR Grow Downside	BoR Grow Skew
	(1)	(2)	(3)	(4)	(5)	(6)
PMU: Inflation	0.146***	-0.0288***	0.175***			
	(0.0147)	(0.00833)	(0.0207)			
PMU: Deflation	-0.0395***	0.0888***	-0.124***			
	(0.0102)	(0.00792)	(0.0152)			
Sent: Inflation	0.0408***	-0.0172*	0.0601***			
	(0.0151)	(0.00952)	(0.0219)			
PMU: Growth				-0.0182***	0.0986***	-0.116***
				(0.00620)	(0.0140)	(0.0173)
Sent: Growth				0.0384***	-0.119***	0.155***
				(0.00650)	(0.0104)	(0.0145)
Observations	245	245	258	245	245	258
R^2	0.609	0.500	0.606	0.151	0.488	0.428

Validation II: Summary of Economic Projections

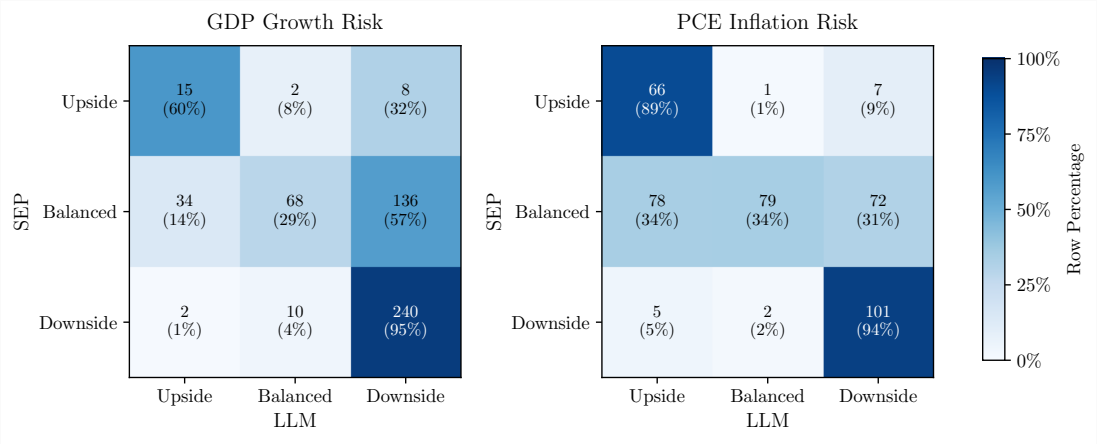
From Oct 2007, and four times a year from 2008 onwards, FOMC members respond to a survey about economic conditions collated in the Summary of Economic Projections.

Includes question "Please indicate your judgment of the risk weighting around your projections" with three responses:

- Weighted to downside
- Broadly balanced
- Weighted to upside

Additional questions on member-specific point estimates of macro variables and optimal policy used as controls in regression analysis.

Comparison to SEP

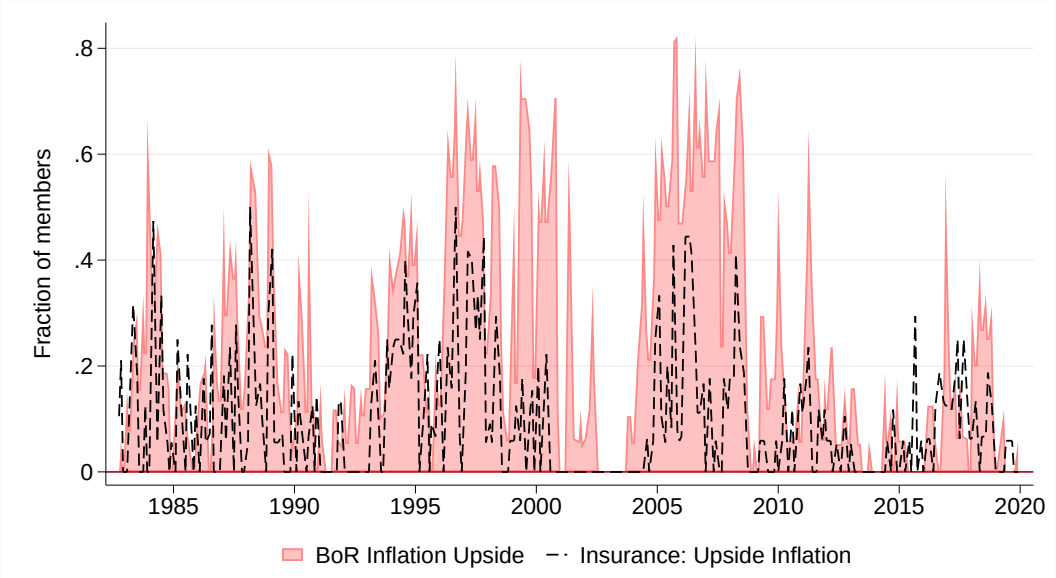


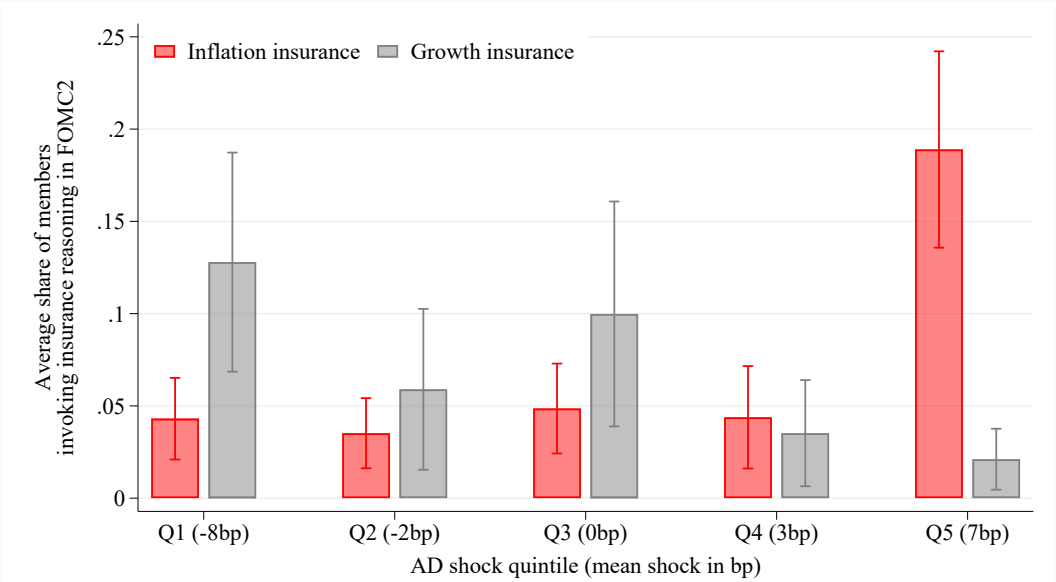
Examples of Mismatch

Kocherlakota Apr 2012, Inflation, LLM: Upside, SEP: Balanced
"In this scenario, we might well see rapid increases in inflation and inflation expectations when labor markets seem weak by historical standards... This does not strike me as a risk that we should ignore."

Yellen Dec 2013, Growth, LLM: Balanced, SEP: Downside
"I now see appreciably diminished downside risks to the outlook for the labor market and more balanced risks to growth than in September."

Bullard Sep 2016, Inflation, LLM: Downside, SEP: Upside
"I do remain concerned that market-based measures of inflation expectations remain too low to be consistent with the 2 percent inflation forecast. This is a risk to the outlook that I have described."


Fraction of members
.8
.6
.4
.2
0
1985
1990
1995
2000
2005
2010
2015
2020
BoR Inflation Upside
Insurance: Upside Inflation


Inflation insurance
Growth insurance
Average share of members invoking insurance reasoning in FOMC2
.25
.2
.15
.1
.05
0
Q1 (-8bp)
Q2 (-2bp)
Q3 (0bp)
Q4 (3bp)
Q5 (7bp)
AD shock quintile (mean shock in bp)

Conclusion

Key Message

My view on AI for Economics Research

AI IS A POWERFUL SET OF TOOLS THAT CAN AUGMENT ECONOMIC RESEARCH...

BUT IT IS NOT A SUBSTITUTE FOR ECONOMIC REASONING AND CAREFUL EMPIRICS.

- Embrace it
- Leverage it
- Deploy it to improve YOUR research

If you're going to do it, do it right!

Econ needs to be a bit more like NLP—and beware of domain-familiarity effects with LLMs!

End

Questions?

Further Reading

Foundations & overview

- Jesus Fernandez-Villaverde - various papers and lectures
- Vaswani et al. (2017) — "Attention Is All You Need"

AI in economics / NLP methods

- Bryan, Kevin A. 2026. "The Economic Impacts of Artificial Intelligence: A Multidisciplinary, Multi-book Review." Journal of Economic Literature 64 (1): 281–300.
- Gentzkow, Kelly & Taddy (2019) — "Text as Data," *JEL*
- Athey & Imbens (2019) — "Machine Learning Methods Economists Should Know," *AER P&P*
- Chernozhukov et al. (2018) — "Double/Debiased Machine Learning," *Econometrics Journal*