Discussion of Bhagwat, Cookson, Dim, and Niessner (2026)

"The Market's Mirror: Revealing Investor Disagreement with LLMs"

Discussant: Sangmin Simon Oh (Columbia Business School)

SFS Cavalcade NA 2026

Recap

Question: Can LLMs reveal how different investors interpret the same firm news, and does that disagreement show up in trading?

Recap

Question: Can LLMs reveal how different investors interpret the same firm news, and does that disagreement show up in trading?

Methodology:

- Prompt LLMs as 216 personas built from FINRA demographics + political orientations
- Elicit buy/hold/sell responses and short rationales for Ravenpack headlines, with disagreement measured as dispersion across personas

Recap

Question: Can LLMs reveal how different investors interpret the same firm news, and does that disagreement show up in trading?

Methodology:

- Prompt LLMs as 216 personas built from FINRA demographics + political orientations
- Elicit buy/hold/sell responses and short rationales for Ravenpack headlines, with disagreement measured as dispersion across personas

Main Findings

- Disagreement varies across news type (as a function of share of fundamentals)
- Income and politics generate large response dispersion
- LLM disagreement predicts same-day and next-day abnormal volume

Recap

Question: Can LLMs reveal how different investors interpret the same firm news, and does that disagreement show up in trading?

Methodology:

- Prompt LLMs as 216 personas built from FINRA demographics + political orientations
- Elicit buy/hold/sell responses and short rationales for Ravenpack headlines, with disagreement measured as dispersion across personas

Main Findings

- Disagreement varies across news type (as a function of share of fundamentals)
- Income and politics generate large response dispersion
- LLM disagreement predicts same-day and next-day abnormal volume

Ambitious exercise to use LLM personas for a scalable application

- **A very thought-provoking exercise for future asset pricing research**

Recap

Question: Can LLMs reveal how different investors interpret the same firm news, and does that disagreement show up in trading?

Methodology:

- Prompt LLMs as 216 personas built from FINRA demographics + political orientations
- Elicit buy/hold/sell responses and short rationales for Ravenpack headlines, with disagreement measured as dispersion across personas

Main Findings

- Disagreement varies across news type (as a function of share of fundamentals)
- Income and politics generate large response dispersion
- LLM disagreement predicts same-day and next-day abnormal volume

Ambitious exercise to use LLM personas for a scalable application

- **A very thought-provoking exercise for future asset pricing research**

Plan for Discussion

1. Using AI/ML in Finance Research
2. AI Joint Hypothesis Problem
3. Connecting to Literature on Investor Disagreement

Comment 1. Using AI/ML in Finance Research

4 Types of Papers on Using AI/ML in Finance

1. **Information extraction**, in which existing AI/ML tools can be used to extract information from sources that are traditionally not easy to parse

The Costs of Housing Regulation: Evidence from Generative Regulatory Measurement

Alexander W. Bartik, Arpit Gupta, and Daniel Milo*

August 19, 2025

Abstract

We introduce "generative regulatory measurement," using Large Language Models to interpret administrative documents with 96% accuracy in binary classification and 0.87 correlation for continuous questions. Our analysis of U.S. zoning regulations reveals four facts: (1) Housing regulations are multidimensional with two main principal components. (2) The first principal component represents *value capture* in high housing demand areas. (3) The second principal component associates with *exclusionary zoning*, increasing housing costs and socioeconomic exclusion. (4) Zoning follows a monocentric pattern with regional variations and is especially strict in Northeast suburbs. We develop a model of municipal regulatory choice consistent with these facts.

4 Types of Papers on Using AI/ML in Finance

1. **Information extraction**, in which existing AI/ML tools can be used to extract information from sources that are traditionally not easy to parse
2. **Modelling complex relationships**, where non-linearities can be flexibly integrated

Empirical Asset Pricing via Machine Learning*

Shihao Gu
Booth School of Business, University of Chicago

Bryan Kelly
Yale University, AQR Capital Management, and NBER

Dacheng Xiu
Booth School of Business, University of Chicago

We perform a comparative analysis of machine learning methods for the canonical problem of empirical asset pricing: measuring asset risk premiums. We demonstrate large economic gains to investors using machine learning forecasts, in some cases doubling the performance of leading regression-based strategies from the literature. We identify the best-performing methods (trees and neural networks) and trace their predictive gains to allowing nonlinear predictor interactions missed by other methods. All methods agree on the same set of dominant predictive signals, a set that includes variations on momentum, liquidity, and volatility. (*JEL* C52, C55, C58, G0, G1, G17)

4 Types of Papers on Using AI/ML in Finance

1. **Information extraction**, in which existing AI/ML tools can be used to extract information from sources that are traditionally not easy to parse
2. **Modelling complex relationships**, where non-linearities can be flexibly integrated
3. **Intellectual transfer**, where core ideas from AI/ML are borrowed and applied to economics and finance

Asset Embeddings*

Xavier Gabaix† Ralph S.J. Koijen‡ Robert J. Richmond§ Motohiro Yogo¶

August 1, 2025

Abstract

Traditional representations of firms use accounting and financial market data, but investors use richer information sets. Theoretically, portfolio holdings contain all relevant information for asset prices, recoverable under empirically realistic conditions. Building on recent advances in machine learning and artificial intelligence, we develop asset embeddings that leverage portfolio holdings to represent firms, similar to word embeddings leveraging document structure. We evaluate different methods of estimating asset embeddings on three new benchmarks. We also develop investor embeddings to represent investors and their strategies. We economically interpret asset (investor) embeddings by applying large language models to firm- (investor-)level text data.

JEL: C53, G12, G23

Keywords: Artificial intelligence, Machine learning, Asset pricing, Recommender systems, Transformer models, Benchmarks

4 Types of Papers on Using AI/ML in Finance

1. **Information extraction**, in which existing AI/ML tools can be used to extract information from sources that are traditionally not easy to parse
2. **Modelling complex relationships**, where non-linearities can be flexibly integrated
3. **Intellectual transfer**, where core ideas from AI/ML are borrowed and applied to economics and finance
4. **Benchmarking human decisions**, where AI/ML-generated forecasts can be considered as benchmarks to which human judgments can then be evaluated

Predictably Unequal? The Effects of Machine Learning on Credit Markets

ANDREAS FUSTER, PAUL GOLDSMITH-PINKHAM, TARUN RAMADORAI, and ANSGAR WALTHER

ABSTRACT

Innovations in statistical technology in functions including credit-screening have raised concerns about distributional impacts across categories such as race. Theoretically, distributional effects of better statistical technology can come from greater flexibility to uncover structural relationships or from triangulation of otherwise excluded characteristics. Using data on U.S. mortgages, we predict default using traditional and machine learning models. We find that Black and Hispanic borrowers are disproportionately less likely to gain from the introduction of machine learning. In a simple equilibrium credit market model, machine learning increases disparity in rates between and within groups, with these changes attributable primarily to greater flexibility.

4 Types of Papers on Using AI/ML in Finance

1. **Information extraction**, in which existing AI/ML tools can be used to extract information from sources that are traditionally not easy to parse
2. **Modelling complex relationships**, where non-linearities can be flexibly integrated
3. **Intellectual transfer**, where core ideas from AI/ML are borrowed and applied to economics and finance
4. **Benchmarking human decisions**, where AI/ML-generated forecasts can be considered as benchmarks to which human judgments can then be evaluated

Predictably Unequal? The Effects of Machine Learning on Credit Markets

ANDREAS FUSTER, PAUL GOLDSMITH-PINKHAM, TARUN RAMADORAI, and ANSGAR WALTHER

ABSTRACT

Innovations in statistical technology in functions including credit-screening have raised concerns about distributional impacts across categories such as race. Theoretically, distributional effects of better statistical technology can come from greater flexibility to uncover structural relationships or from triangulation of otherwise excluded characteristics. Using data on U.S. mortgages, we predict default using traditional and machine learning models. We find that Black and Hispanic borrowers are disproportionately less likely to gain from the introduction of machine learning. In a simple equilibrium credit market model, machine learning increases disparity in rates between and within groups, with these changes attributable primarily to greater flexibility.

LLM as Human Benchmarks

Many papers have used ML forecasts as “rational” or “fair” benchmarks to which human judgments can then be evaluated.

LLM as Human Benchmarks

Many papers have used ML forecasts as “rational” or “fair” benchmarks to which human judgments can then be evaluated.

This paper: Flips the benchmark logic

LLM as Human Benchmarks

Many papers have used ML forecasts as “rational” or “fair” benchmarks to which human judgments can then be evaluated.

This paper: Flips the benchmark logic

Traditional ML Benchmark

- ML forecast ≈ Rational, unbiased
- Human – ML = bias, mistake, or discrimination

LLM as Human Benchmarks

Many papers have used ML forecasts as “rational” or “fair” benchmarks to which human judgments can then be evaluated.

This paper: Flips the benchmark logic

Traditional ML Benchmark

- ML forecast ≈ Rational, unbiased
- Human – ML = bias, mistake, or discrimination

This Paper

- LLM persona ≈ Investor type
- Dispersion across LLM personas = disagreement

LLM as Human Benchmarks

Many papers have used ML forecasts as “rational” or “fair” benchmarks to which human judgments can then be evaluated.

This paper: Flips the benchmark logic

Traditional ML Benchmark

- ML forecast ≈ Rational, unbiased
- Human – ML = bias, mistake, or discrimination

This Paper

- LLM persona ≈ Investor type
- Dispersion across LLM personas = disagreement

This is a big deal!

- Gives us a scalable “survey” when real surveys are too slow or too sparse
- Allows us to hold the informational signal fixed, which is infeasible in most cases
- Different from current use cases of AI/ML in studying human behavior

LLM as Human Benchmarks

Many papers have used ML forecasts as "rational" or "fair" benchmarks to which human judgments can then be evaluated.

This paper: Flips the benchmark logic

Traditional ML Benchmark

- ML forecast ≈ Rational, unbiased
- Human – ML = bias, mistake, or discrimination

This Paper

- LLM persona ≈ Investor type
- Dispersion across LLM personas = disagreement

This is a big deal!

- Gives us a scalable "survey" when real surveys are too slow or too sparse
- Allows us to hold the informational signal fixed, which is infeasible in most cases
- Different from current use cases of AI/ML in studying human behavior

Suggestion #1a. Emphasize the contribution to the broader literature on AI/ML in finance

- Dispersion across personas as an estimate of a latent population object that no survey can produce at scale

LLM to Overcome Existing Limitations

Authors motivate the paper highlighting the capabilities of LLMs to overcome traditional limitations of survey instruments:

LLM to Overcome Existing Limitations

Authors motivate the paper highlighting the capabilities of LLMs to overcome traditional limitations of survey instruments:

> Personal characteristics shape financial behavior and portfolio choice.[1] It is therefore natural to expect heterogeneity in how investors interpret firm news. This differential interpretation can be an important driver of disagreement and trading (Miller, 1977; Kandel and Pearson, 1995; Hong and Stein, 1999). But measuring which demographic dimensions drive disagreement about *specific* news at high frequency is difficult: surveys are infrequent and rarely map cleanly to article-level belief updates. In this paper, we deploy a local LLM at scale to address these questions.

An **excellent** use case of using LLMs and authors have the scope to do much more.

LLM to Overcome Existing Limitations

Authors motivate the paper highlighting the capabilities of LLMs to overcome traditional limitations of survey instruments:

> Personal characteristics shape financial behavior and portfolio choice.[1] It is therefore natural to expect heterogeneity in how investors interpret firm news. This differential interpretation can be an important driver of disagreement and trading (Miller, 1977; Kandel and Pearson, 1995; Hong and Stein, 1999). But measuring which demographic dimensions drive disagreement about *specific* news at high frequency is difficult: surveys are infrequent and rarely map cleanly to article-level belief updates. In this paper, we deploy a local LLM at scale to address these questions.

An **excellent** use case of using LLMs and authors have the scope to do much more.

Suggestion #1b. Explore and highlight the potential LLMs in high-frequency analysis

- Earnings announcements
- FOMC announcements
- Macro vs. firm-specific news
- Episodes: COVID-19? Meme stock surges?

Example: FOMC

Real-time Price Discovery via Verbal Communication:

Method and Application to Fedspeak

Roberto Gómez-Cram Marco Grotteria*

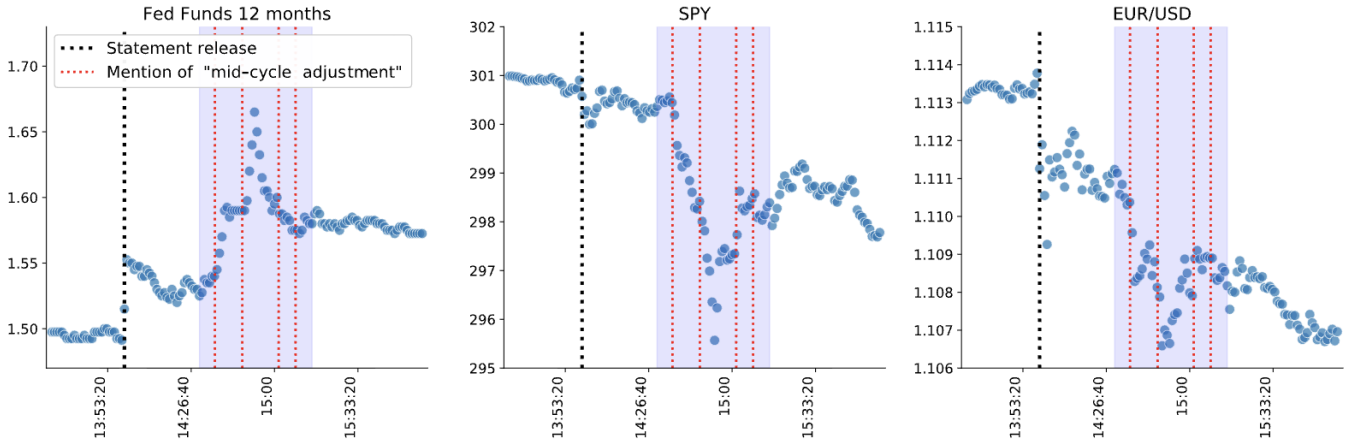

Fig. 1. *Notes*: The figure shows the intraday evolution of the implied rate from the 12-month federal funds futures, the SPY price level, and the EUR/USD exchange rate on July 31, 2019. The black dashed vertical line highlights the time the FOMC statement was released (14:00). The shaded area denotes the FOMC press conference. The conference started at 14:30 and lasted for about 45 minutes. The red dotted lines highlight the times in which the Chairman mentioned "mid-cycle adjustment to policy."

The authors' methodology can produce a real-time belief updating as the text arrives!

Comment 2. AI Joint Hypothesis Problem

AI Joint Hypothesis Problem

Large Language **Model** (LLM) is a **model**, so the usual caution with models apply.

AI Joint Hypothesis Problem

Large Language **Model** (LLM) is a **model**, so the usual caution with models apply.

One form of caution is what I'd call the **AI joint hypothesis problem**.

> LLM output quality is jointly a function of (i) prompt quality and (ii) LLM quality. You can't tell which is failing from output alone.

AI Joint Hypothesis Problem

Large Language **Model** (LLM) is a **model**, so the usual caution with models apply.

One form of caution is what I'd call the **AI joint hypothesis problem**.

> LLM output quality is jointly a function of (i) prompt quality and (ii) LLM quality. You can't tell which is failing from output alone.

In the context of this paper:

> The cross-persona variation jointly tests (i) human behavior and (ii) that the LLM is a faithful mirror of the persona it's labeled with.

The paper implicitly assumes (ii) and proceeds to study (i).

Abstract

Large language models (LLMs) can emulate human perspectives. Leveraging this idea, we study how investor disagreement emerges in response to firm news. We endow an LLM with 216 representative investor personas and elicit buy, hold, or sell responses to S&P 500 firm news headlines from 2022-2025. Dispersion in responses yields article-level disagreement for 1.25 million headlines and sheds light on its sources. Disagreement is highest for socially-oriented news and lowest for fundamentals. Persona responses reflect non-pecuniary rationales and align with human-survey benchmarks. Disagreement predicts elevated same- and next-day abnormal trading volume, especially retail, and results persist beyond the LLM's training cutoff.

How good is an LLM at approximating human behavior?

Complaint #1: LLMs produce too little variation, unstable responses across prompts/time, and regression coefficients that often differ from real survey data:

Synthetic Replacements for Human Survey Data? The Perils of Large Language Models

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel and Jennifer M. Larson

Political Science Department, Vanderbilt University, Nashville, TN, USA

Corresponding author: James Bisbee; Email: james.h.bisbee@vanderbilt.edu

(Received 2 May 2023; revised 18 January 2024; accepted 20 January 2024; published online 17 May 2024)

Abstract

Large language models (LLMs) offer new research possibilities for social scientists, but their potential as "synthetic data" is still largely unknown. In this paper, we investigate how accurately the popular LLM ChatGPT can recover public opinion, prompting the LLM to adopt different "personas" and then provide feeling thermometer scores for 11 sociopolitical groups. The average scores generated by ChatGPT correspond closely to the averages in our baseline survey, the 2016–2020 American National Election Study (ANES). Nevertheless, sampling by ChatGPT is not reliable for statistical inference: there is less variation in responses than in the real surveys, and regression coefficients often differ significantly from equivalent estimates obtained using ANES data. We also document how the distribution of synthetic responses varies with minor changes in prompt wording, and we show how the same prompt yields significantly different results over a 3-month period. Altogether, our findings raise serious concerns about the quality, reliability, and reproducibility of synthetic survey data generated by LLMs.

How good is an LLM at approximating human behavior?

Complaint #2: Prompting “be a member of group X” often turns lived experience into a stereotype or average representation, especially when identity is central to the task.

Large language models that replace human participants can harmfully misportray and flatten identity groups

Angelina Wang[1], Jamie Morgenstern[2], John P. Dickerson[3,4]

[1]Computer Science, Stanford University, Palo Alto, CA, USA.
[2]Computer Science & Engineering, University of Washington, Seattle, WA, USA.
[3]Computer Science, University of Maryland, College Park, MD, USA.
[4]Arthur, New York City, NY, USA.

Abstract

Large language models (LLMs) are increasing in capability and popularity, propelling their application in new domains—including as replacements for human participants in computational social science, user testing, annotation tasks, and more. In many settings, researchers seek to distribute their surveys to a sample of participants that are representative of the underlying human population of interest. This means in order to be a suitable replacement, LLMs will need to be able to capture the influence of positionality (i.e., relevance of social identities like gender and race). However, we show that there are two inherent limitations in the way current LLMs are trained that prevent this. We argue analytically for why LLMs are likely to both *misportray* and *flatten* the representations of demographic groups, then empirically show this on 4 LLMs through a series of human studies with 3200 participants across 16 demographic identities. We also discuss a third limitation about how identity prompts can essentialize identities. Throughout, we connect each limitation to a pernicious history of epistemic injustice against the value of lived experiences that explains why replacement is harmful for marginalized demographic groups. Overall, we urge caution in use cases where LLMs are intended to replace human participants whose identities are relevant to the task at hand. At the same time, in cases where the benefits of LLM replacement are determined to outweigh the harms (e.g., the goal is to supplement rather than fully replace, engaging human participants may cause them harm), we provide inference-time techniques that we empirically demonstrate do reduce, but do not remove, these harms.

How good is an LLM at approximating human behavior?

Complaint #3: LLMs often fill in missing context (e.g. coffee costs $100 may lead the LLM to infer additional information about its quality)

The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective

George Gui and Olivier Toubia*

November 22, 2025

Abstract

Large Language Models (LLMs) have shown impressive potential to simulate human behavior. We identify a fundamental challenge in using them to simulate experiments: when LLM-simulated subjects are blind to the experimental design (as is standard practice with human subjects), variations in treatment systematically affect unspecified variables that should remain constant, violating the unconfoundedness assumption. Using demand estimation as a context and an actual experiment with 40 different products as a benchmark, we show this can lead to implausible results. While confounding may in principle be addressed by controlling for covariates, this can compromise ecological validity in the context of LLM simulations: controlled covariates become artificially salient in the simulated decision process. We show formally that confoundness stems from ambiguous prompting strategies. Therefore, it can be addressed by developing unambiguous prompting strategies through unblinding, i.e., revealing the experiment design in LLM simulations. Our empirical results show that this strategy consistently enhances model performance across all tested models, including both out-of-box reasoning and non-reasoning models. We also show that it is a technique that complements fine-tuning: while fine-tuning can improve simulation performance, an unambiguous prompting strategy makes the predictions robust to the inclusion of irrelevant data in the fine-tuning process.

How good is an LLM at approximating human behavior?

Complaint #4: Model gives correct-looking answers on curated tests but lacks a coherent concept representation that would support reliable extrapolation to nearby tasks.

Potemkin Understanding in Large Language Models

Marina Mancoridis [1] **Keyon Vafa** [2] **Bec Weeks** [3] **Sendhil Mullainathan** [1]

Abstract

Large language models (LLMs) are regularly evaluated using benchmark datasets. But what justifies making inferences about an LLM's capabilities based on its answers to a curated set of questions? This paper first introduces a formal framework to address this question. The key is to note that the benchmarks used to test LLMs—such as AP exams—are also those used to test people. However, this raises an implication: these benchmarks are only valid tests if LLMs misunderstand concepts in ways that mirror human misunderstandings. Otherwise, success on benchmarks only demonstrates **potemkin understanding:** the illusion of understanding driven by answers irreconcilable with how any human would interpret a concept. We present two procedures for quantifying the existence of potemkins: one using a specially designed benchmark in three domains, the other using a general procedure that provides a lower-bound on their prevalence. We find that potemkins are ubiquitous across models, tasks, and domains. We also find that these failures reflect not just incorrect understanding, but deeper internal incoherence in concept representations.

Validations

Different tests validate different links in the chain:

Validations

Different tests validate different links in the chain:

Implementation

- Model Swap: Compares Llama to GPT-4.1 mini on matched headlines
- Headline vs. Article: Compares headline-only prompts to full-article prompts

Validations

Different tests validate different links in the chain:

Implementation

- Model Swap: Compares Llama to GPT-4.1 mini on matched headlines
- Headline vs. Article: Compares headline-only prompts to full-article prompts

Disagreement Measure

- Demographic Mirror: Checks whether persona responses have intuitive demographic gradients (i.e. does the measure behave like a disagreement construct?)

Validations

Different tests validate different links in the chain:

Implementation

- Model Swap: Compares Llama to GPT-4.1 mini on matched headlines
- Headline vs. Article: Compares headline-only prompts to full-article prompts

Disagreement Measure

- Demographic Mirror: Checks whether persona responses have intuitive demographic gradients (i.e. does the measure behave like a disagreement construct?)

Human Approximation

- Human Moral Survey: Benchmark LLM personas against human survey
- Digital Twins Survey: Examine if LLMs can reproduce human behavioral-bias survey responses better than random guessing

Validations

Different tests validate different links in the chain:

Implementation

- Model Swap: Compares Llama to GPT-4.1 mini on matched headlines
- Headline vs. Article: Compares headline-only prompts to full-article prompts

Disagreement Measure

- Demographic Mirror: Checks whether persona responses have intuitive demographic gradients (i.e. does the measure behave like a disagreement construct?)

Human Approximation

- Human Moral Survey: Benchmark LLM personas against human survey
- Digital Twins Survey: Examine if LLMs can reproduce human behavioral-bias survey responses better than random guessing

I don’t think we are fully there yet in establishing that LLM has the right internal mapping from demographics × news → belief updates.

Validations

Different tests validate different links in the chain:

Implementation

- Model Swap: Compares Llama to GPT-4.1 mini on matched headlines
- Headline vs. Article: Compares headline-only prompts to full-article prompts

Disagreement Measure

- Demographic Mirror: Checks whether persona responses have intuitive demographic gradients (i.e. does the measure behave like a disagreement construct?)

Human Approximation

- Human Moral Survey: Benchmark LLM personas against human survey
- Digital Twins Survey: Examine if LLMs can reproduce human behavioral-bias survey responses better than random guessing

I don't think we are fully there yet in establishing that LLM has the right internal mapping from demographics × news → belief updates.

An Analogy: A structural model can match some moments in the data and still produce unreliable counterfactuals.

What can we do?

What can we do?

For authors:

Suggestion #2a. Clarify the role of different validation tests in the paper

- LLMs are a new frontier and it's good for the literature to be very precise
- Question: What does it mean to be concerned that "model memorized demographic differences from its training data" in this context?

What can we do?

For authors:

Suggestion #2a. Clarify the role of different validation tests in the paper

- LLMs are a new frontier and it's good for the literature to be very precise
- Question: What does it mean to be concerned that "model memorized demographic differences from its training data" in this context?

For everyone:

Suggestion #2b. Can we define the latent primitives of the LLM as a "human simulator"?

What can we do?

For authors:

Suggestion #2a. Clarify the role of different validation tests in the paper

- LLMs are a new frontier and it's good for the literature to be very precise
- Question: What does it mean to be concerned that "model memorized demographic differences from its training data" in this context?

For everyone:

Suggestion #2b. Can we define the latent primitives of the LLM as a "human simulator"?

- In many structural models, we estimate deep parameters that are not directly observed

Demand models	Price sensitivity, substitution patterns
Portfolio Choice	Risk aversion, beliefs, participations costs, attention costs
Firm Investment	Adjustment costs, financing frictions
Labor search	Job offer rates, search costs, reservation wages

What can we do?

For authors:

Suggestion #2a. Clarify the role of different validation tests in the paper

- LLMs are a new frontier and it's good for the literature to be very precise
- Question: What does it mean to be concerned that "model memorized demographic differences from its training data" in this context?

For everyone:

Suggestion #2b. Can we define the latent primitives of the LLM as a "human simulator"?

- In many structural models, we estimate deep parameters that are not directly observed

Demand models	Price sensitivity, substitution patterns
Portfolio Choice	Risk aversion, beliefs, participations costs, attention costs
Firm Investment	Adjustment costs, financing frictions
Labor search	Job offer rates, search costs, reservation wages

- It would be useful to agree on LLM-specific primitives:

What can we do?

For authors:

Suggestion #2a. Clarify the role of different validation tests in the paper

- LLMs are a new frontier and it's good for the literature to be very precise
- Question: What does it mean to be concerned that "model memorized demographic differences from its training data" in this context?

For everyone:

Suggestion #2b. Can we define the latent primitives of the LLM as a "human simulator"?

- In many structural models, we estimate deep parameters that are not directly observed

Demand models	Price sensitivity, substitution patterns
Portfolio Choice	Risk aversion, beliefs, participations costs, attention costs
Firm Investment	Adjustment costs, financing frictions
Labor search	Job offer rates, search costs, reservation wages

- It would be useful to agree on LLM-specific primitives:
  - **Baseline prior**: The model's default answer absent persona information
  - **Prompt sensitivity**: How much responses change when wording changes
  - **Stereotype intensity**: How strongly demographic labels trigger learned group associations

Comment 3. Connecting to Literature on Investor Disagreement

Models of Investor Disagreement

Heterogeneous Beliefs (Harrison and Kreps, 1978; Scheinkman and Xiong, 2003)

- Investors disagree about expected payoffs

⋮

Models of Investor Disagreement

Heterogeneous Beliefs (Harrison and Kreps, 1978; Scheinkman and Xiong, 2003)

- Investors disagree about expected payoffs

Risk Aversion (Chan and Kogan, 2002; Gârleanu and Panageas 2015)

- Investors agree on expected returns but respond differently because of heterogeneous preferences

⋮

Models of Investor Disagreement

Heterogeneous Beliefs (Harrison and Kreps, 1978; Scheinkman and Xiong, 2003)

- Investors disagree about expected payoffs

Risk Aversion (Chan and Kogan, 2002; Gârleanu and Panageas 2015)

- Investors agree on expected returns but respond differently because of heterogeneous preferences

Mandates / Constraints (Chien, Cole, and Lustig, 2012; Koijen and Yogo, 2019)

- An index fund and a hedge fund may agree on a stock's prospects but trade differently because of tracking error constraints

⋮

Models of Investor Disagreement

Heterogeneous Beliefs (Harrison and Kreps, 1978; Scheinkman and Xiong, 2003)

- Investors disagree about expected payoffs

Risk Aversion (Chan and Kogan, 2002; Gârleanu and Panageas 2015)

- Investors agree on expected returns but respond differently because of heterogeneous preferences

Mandates / Constraints (Chien, Cole, and Lustig, 2012; Koijen and Yogo, 2019)

- An index fund and a hedge fund may agree on a stock's prospects but trade differently because of tracking error constraints

Limited Participation (Mankiw and Zeldes, 1989; Vissing-Jorgensen, 2002)

- Some investors don't hold the risky asset at all, so non-participants have $u_{it} = 0$ while participants have active demand shifts

⋮

Models of Investor Disagreement

Heterogeneous Beliefs (Harrison and Kreps, 1978; Scheinkman and Xiong, 2003)

- Investors disagree about expected payoffs

Risk Aversion (Chan and Kogan, 2002; Gârleanu and Panageas 2015)

- Investors agree on expected returns but respond differently because of heterogeneous preferences

Mandates / Constraints (Chien, Cole, and Lustig, 2012; Koijen and Yogo, 2019)

- An index fund and a hedge fund may agree on a stock's prospects but trade differently because of tracking error constraints

Limited Participation (Mankiw and Zeldes, 1989; Vissing-Jorgensen, 2002)

- Some investors don't hold the risky asset at all, so non-participants have $u_{it} = 0$ while participants have active demand shifts

Heterogeneous Income Risk (Constantinides and Duffie, 1996; Heaton and Lucas, 2000)

- Investors face idiosyncratic uninsurable income shocks that force portfolio adjustments unrelated to views about the stock

⋮

Models of Investor Disagreement

Heterogeneous Beliefs (Harrison and Kreps, 1978; Scheinkman and Xiong, 2003)

- Investors disagree about expected payoffs

Risk Aversion (Chan and Kogan, 2002; Gârleanu and Panageas 2015)

- Investors agree on expected returns but respond differently because of heterogeneous preferences

Mandates / Constraints (Chien, Cole, and Lustig, 2012; Koijen and Yogo, 2019)

- An index fund and a hedge fund may agree on a stock's prospects but trade differently because of tracking error constraints

Limited Participation (Mankiw and Zeldes, 1989; Vissing-Jorgensen, 2002)

- Some investors don't hold the risky asset at all, so non-participants have $u_{it} = 0$ while participants have active demand shifts

Heterogeneous Income Risk (Constantinides and Duffie, 1996; Heaton and Lucas, 2000)

- Investors face idiosyncratic uninsurable income shocks that force portfolio adjustments unrelated to views about the stock

⋮

Connecting to Asset Pricing Literature

LLMs probably most useful for disentangling beliefs vs. risk aversion.

Connecting to Asset Pricing Literature

LLMs probably most useful for disentangling beliefs vs. risk aversion.

Suggestion #3a. Separate beliefs vs. risk aversion

- Q1. “Where do you think the stock will be in 12 months?”
- Q2. “Given your belief, should you BUY/HOLD/SELL?”

Connecting to Asset Pricing Literature

LLMs probably most useful for disentangling beliefs vs. risk aversion.

Suggestion #3a. Separate beliefs vs. risk aversion

- Q1. “Where do you think the stock will be in 12 months?”
- Q2. “Given your belief, should you BUY/HOLD/SELL?”

Suggestion #3b. Institutional Personas

- Much of trading in the U.S. is still driven by institutional investors
- A small companion analysis with institutional personas (index funds, value funds, growth funds) would be useful

Connecting to Asset Pricing Literature

LLMs probably most useful for disentangling beliefs vs. risk aversion.

Suggestion #3a. Separate beliefs vs. risk aversion

- Q1. “Where do you think the stock will be in 12 months?”
- Q2. “Given your belief, should you BUY/HOLD/SELL?”

Suggestion #3b. Institutional Personas

- Much of trading in the U.S. is still driven by institutional investors
- A small companion analysis with institutional personas (index funds, value funds, growth funds) would be useful

Suggestion #3c. Tests of Return Predictability

- Miller (1997): high-disagreement stocks are overpriced and earn lower future returns
- Can we sort firm-days into quintiles of abnormal LLM disagreement?

Connecting to Asset Pricing Literature

LLMs probably most useful for disentangling beliefs vs. risk aversion.

Suggestion #3a. Separate beliefs vs. risk aversion

- Q1. “Where do you think the stock will be in 12 months?”
- Q2. “Given your belief, should you BUY/HOLD/SELL?”

Suggestion #3b. Institutional Personas

- Much of trading in the U.S. is still driven by institutional investors
- A small companion analysis with institutional personas (index funds, value funds, growth funds) would be useful

Suggestion #3c. Tests of Return Predictability

- Miller (1997): high-disagreement stocks are overpriced and earn lower future returns
- Can we sort firm-days into quintiles of abnormal LLM disagreement?

Suggestion #3d. Wealth-reweighting as a robustness check

- For asset prices, we care about wealth-weighted averages rather than population-weighted averages.

Final Thoughts

- Ambitious paper to use LLMs as a scalable instrument for measuring how different investors interpret the same news with article-level resolution

Final Thoughts

- Ambitious paper to use LLMs as a scalable instrument for measuring how different investors interpret the same news with article-level resolution

- **Punchline:** Synthetic disagreement generated from LLMs can measure how different investor types interpret the same firm news

Final Thoughts

- Ambitious paper to use LLMs as a scalable instrument for measuring how different investors interpret the same news with article-level resolution

- **Punchline:** Synthetic disagreement generated from LLMs can measure how different investor types interpret the same firm news

- **A few suggestions for future iterations:**
  - Push the measure into settings where article-level frequency is decisive
  - Acknowledge the AI joint hypothesis problem and clarify the role of validation
  - Speak more to the asset pricing literature on disagreement

Final Thoughts

- Ambitious paper to use LLMs as a scalable instrument for measuring how different investors interpret the same news with article-level resolution

- **Punchline**: Synthetic disagreement generated from LLMs can measure how different investor types interpret the same firm news

- **A few suggestions for future iterations:**
  - Push the measure into settings where article-level frequency is decisive
  - Acknowledge the AI joint hypothesis problem and clarify the role of validation
  - Speak more to the asset pricing literature on disagreement

- **A few questions prompted by the paper for the future:**
  - What other reduced-form objects in finance does this machinery unlock?
  - As frontier models evolve, does persona dispersion converge or diverge?

Final Thoughts

- Ambitious paper to use LLMs as a scalable instrument for measuring how different investors interpret the same news with article-level resolution

- **Punchline:** Synthetic disagreement generated from LLMs can measure how different investor types interpret the same firm news

- **A few suggestions for future iterations:**
  - Push the measure into settings where article-level frequency is decisive
  - Acknowledge the AI joint hypothesis problem and clarify the role of validation
  - Speak more to the asset pricing literature on disagreement

- **A few questions prompted by the paper for the future:**
  - What other reduced-form objects in finance does this machinery unlock?
  - As frontier models evolve, does persona dispersion converge or diverge?

- **Very much looking forward to the next version!**