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

Day #4: Synthetic Agents, Simulations, and Counterfactuals

Sophia Kazinnik

Digital Economy Lab
Stanford University

August 20, 2026

Today's Plan

- **First block**: LLMs as economic agents
- **Second block**: Applications: bank runs and agentic systems
- **Question**: When is simulation useful enough?

Simulation as a Laboratory

- LLM simulations are cheap, fast, and flexible
- They can help us model scenarios that are hard, slow, or risky to field with humans
- (But the model is still a measurement instrument)
- The workflow: **theory** $\rightarrow$ **human validation** $\rightarrow$ **scaled simulation** $\rightarrow$ **external test**
- The research opportunity here is not replacing data, but expanding the design space

Large Language Models as Simulated Economic Agents: What Can We Learn from *Homo Silicus*?*

John J. Horton
MIT & NBER

January 19, 2023

Abstract

Newly-developed large language models (LLM)—because of how they are trained and designed—are implicit computational models of humans—a *homo silicus*. LLMs can be used like economists use *homo economicus*: they can be given endowments, information, preferences, and so on, and then their behavior can be explored in scenarios via simulation. Experiments using this approach, derived from Charness and Rabin (2002), Kahneman, Knetsch and Thaler (1986), and Samuelson and Zeckhauser (1988) show qualitatively similar results to the original, but it is also easy to try variations for fresh insights. LLMs could allow researchers to pilot studies via simulation first, searching for novel social science insights to test in the real world.

The Homo Silicus Argument

Economists already use simplified humans (*homo economicus*). Homo silicus is a new kind of simplification.

- Homo sapiens = an actual human
- Homo economicus = a mathematical model of a human
- Homo silicus = an LLM used as a computational model of a human

Homo silicus, like *homo economicus*, can be wrong in many ways, but are still useful

The Homo Silicus Argument

Instead of solving a model analytically, you instantiate an agent and observe repeated choices.
How does it work? And does it work?

- Training and alignment encode regularities from human written language and decisions.
- Researchers can vary information, preferences, status quo, and incentives.
- Some classic behavioral regularities appear in LLM simulations, though not uniformly.
- Failures can reveal where the model's behavioral representation departs from humans.

Category	Company	
Synthetic audiences	**Aaru**	**$1B headline valuation:** TechCrunch reported a Series A led by Redpoint in Dec. 2025, with the round above $50M. Aaru also signed a strategic partnership with Interpublic to use predictive audience simulations in marketing.
Grounded twins	**Simile**	**$2B valuation:** In July 2026, Simile raised a $200M Series B at a $2B valuation, only five months after its $100M Series A. The company was founded by researchers including Joon Sung Park, Michael Bernstein, and Percy Liang.
Grounded twins	**Ipsos × Stanford PASCL**	**Incumbent + academic validation:** Ipsos announced the Stanford partnership in Aug. 2025 to build and rigorously validate digital-twin panels using its KnowledgePanel infrastructure.
Platform moves	**Qualtrics**	**Synthetic panels inside the survey stack:** Qualtrics introduced synthetic panels into its online-panel workflow, powered by a proprietary model trained on thousands of responses. The public preview was announced for Q4 2025.
Research infrastructure	**Expected Parrot**	**Simulation + human validation:** Open-source, model-agnostic platform for designing and running research with AI agents, including surveys, experiments, and synthetic respondents, with built-in workflows for validating results against real human participants. Part of Y Combinator's Fall 2025 batch and backed by Bloomberg Beta.

How Many Simulation Runs Do We Need?

The same logic as experimental design, e.g.:

More noise $\Rightarrow$ More runs

Smaller effect to detect $\Rightarrow$ More runs

$$n = 2\left(\frac{(z_{0.975} + z_{0.80})\sigma}{\Delta}\right)^2$$

σ = variation across runs $\quad$ Δ = effect we want to detect

Persona Prompts Are Bundles

- A phrase like "you are wealthy" can move wealth, age, politics, risk tolerance, and norms together
- A clean treatment pins the whole trait vector and changes only the coordinate of interest
- Matching the mean response is not enough; check spread and relationships too
- Synthetic agents are useful for mechanisms and pilots, but human claims still need human evidence

Modeling a Bank Run

Bank Run, Interrupted: Modeling Deposit Withdrawals with Generative AI ⋆

Sophia Kazinnik[a],
First Draft: October 30th, 2023
Current Version: August 25th, 2025

[a]*Stanford University, Digital Economy Lab @ HAI*

Abstract

I study depositor behavior in panic-driven bank runs using an LLM-based survey simulation. I build a representative population of synthetic agents by assigning demographic attributes, expose them to a viral panic post, and randomize bank communication interventions. I validate model responses against human benchmarks, then generate systematic message variants that vary tone, strength, and content within the validated design space. I then include the estimated withdrawal propensities into a contagion model, which maps network nodes to depositor personas and propagates withdrawals through a single-layer proximity network. Direct, personalized bank communications with strong reassurances and explicit survival clauses substantially reduce withdrawal intent. This framework offers an efficient, inexpensive method for testing crisis messages and projecting how runs may unfold under alternative communication strategies.

Keywords: Large Language Models; Behavioral Finance; Generative Artificial Intelligence; Simulations; Policy Communication; Bank Runs; Surveys

JEL Classification: G41, G21, C63, C83, D91

Modeling a Bank Run with LLMs: Why?

- Diamond and Dybvig (1983) characterize bank runs as being driven by a "*shift in expectations*," which could arise from almost any stimulus

Modeling a Bank Run with LLMs: Why?

- Simulations using LLMs offer a scalable method to model potential human behavior during financial crises (and any hypothetical scenarios), and can be leveraged to generate a wide range of responses that mimic a diverse population

Bank Runs Today

The goal is to simulate depositor behavior during a social-media fueled, panic-driven bank run:

Modeling a Bank Run with LLMs: How?

- Create a diverse group of **synthetic agents** to explore how these agents *decide to withdraw* their bank deposits in the face of a possible bank run:
  - Expose these synthetic agents to different communication treatments to see how these interventions influence their decision
  - Prompt each synthetic agent to respond to a question related to deposit withdrawal under varying treatments/conditions
  - Gather and analyze responses to gauge the propensity to withdraw deposits under varying treatments

Prompt Inputs: Agents

Define the following demographic attributes:

- `gender`: female, male
- `income`: under $25K, $25K - $50K, $50K - $75K, $75K - $120K, over $120K

Prompt Inputs: Agents

Define the following demographic attributes:

- `education`: no high-school, high-school, college, graduate
- `age`: 18-24, 25-34, 35-54, 55-64, 65+

The combination of these attributes yields 200 unique demographic groups.

Prompt

Main Prompt

You are a [**gender**], with an annual household income of [**income**]. You have [**education**] education, and your age is in the [**age**] year range. You just heard about a run on a local bank (not yours) in your state. [**Network Effect**].

You go online and see the following tweet from a popular Twitter user:

Run on the bank!

Get your money out.

First thing on Monday.

US banks are in trouble.

FED emergency meeting.

Deposits may get locked.

Possible withdrawal limits.

When markets collapse your bank deposits that US banks use to invest may be in danger.

Cash is king. Get out now!

You are not a financial expert.

You have some uninsured deposits in your bank.

[**Treatment**]

Do you plan to withdraw your money from your bank given this information?

Prompt Inputs: Treatments

- `Bank Email`: You receive an email from your bank addressing the rumours:
  We are aware of the recent rumors circulating on social media. Please be assured that our bank is financially strong and your funds are safe with us. We pride ourselves on upholding the highest standards of financial integrity, consistently validated by rigorous internal and external audits. Your trust is paramount to us, and we are here to address any concerns you may have.
- `Bank Text`: You receive a text from your contact at the bank:
  Just a heads up: despite the online buzz, the bank is doing just fine. Always here if you want to chat.
- `President Message`: You read a news article headlined: *"President Affirms: US Banks Remain Stable and Secure"*.
- `Fed Message`: You read a news article headlined: *"Fed Chairman Confirms: Our Banking System is Strong and Stable"*.
- `Cautionary Tale`: Your regional Reserve Bank tweets out a new article ("*The Devastating Impact of Bank Runs during the Great Depression*").
- `Placebo`: You then see a news article headlined:
  "A mild and dry start to November".

Sample Responses

Do you plan to withdraw your money from your bank given this information?

- Persona 1: age 55-64, woman, \$50-75 HHI, graduate education:

  `No, I won't make a withdrawal. It's crucial to not let fear guide my financial judgments. However, I will keep an eye on the situation.`

- Persona 2: age 65+, man, under \$25K HHI, no high-school education:

  `Maybe part of it, to be safe. But I ain't a rich man so it wouldn't hurt my pocket that much even if there's trouble. Rather be calm than followin' panic.`

- Persona 3: age 18-24, man, under \$25K HHI, college education:

  `No, panic withdrawals can worsen the situation. As a founder, it's my responsibility to stay steady and not encourage economic instability. Trusting the robust Federal processes.`

Model Selection

To run the simulation, I evaluate eight LLMs, all under identical prompts drawn from the survey instrument:

- **OpenAI**: GPT-4O (2024-08-06), GPT-4O-MINI (2024-07-18), GPT-4.1 (2025-04-14), and the reasoning model O3 (2025-04-16)
- **Google**: GEMMA-2-27B-INSTRUCT
- **Alibaba**: QWEN-2.5-72B-INSTRUCT
- **Meta**: LLAMA-3.3-70B-INSTRUCT
- **DeepSeek**: DEEPSEEK-R1

Temperature Settings:

- *Non-reasoning models* (GPT-4O, GPT-4O-MINI, GPT-4.1, GEMMA, QWEN, LLAMA): $T \in \{0,\ 0.7\}$
- *Reasoning models* (O3, DEEPSEEK-R1): temperature is not applicable

Table: Withdrawal Rates (%) for Human and Synthetic Respondents

Treatment	Human	LLMs						
		DeepSeek	Gemma	GPT-4.1	GPT-4o	GPT-4o mini	o3	Llama
Baseline	47.55	37.00	96.50	97.50	65.50	33.00	64.50	48.50
Bank Email	35.21	15.50	62.00	100.00	55.50	4.00	71.00	2.00
Bank Text	27.59	12.00	57.00	97.00	45.50	0.00	47.00	3.00
Cautionary Tale	56.38	31.00	86.50	100.00	75.50	44.00	65.00	19.00
Fed Message	34.46	5.50	59.00	94.50	45.00	0.00	58.50	2.50
Network Effect	65.52	50.00	100.00	100.00	79.50	23.50	81.00	47.50
Placebo	49.65	13.50	84.00	100.00	67.50	22.00	55.00	8.00
President Message	45.52	9.00	63.50	97.00	50.50	0.50	62.50	2.50

Notes: Withdrawal rates by treatment. **Human** column shows Prolific results averaged across 1,158 respondents by treatment. **LLMs** column shows results for each model in the simulation, averaged across 1,600 synthetic respondents per model.

Defining Model Bias

Define model bias at the demographic–treatment cell level g as:

$$\Delta_g = \Pr(Y{=}1 \mid g)_{\text{LLM}} - \Pr(Y{=}1 \mid g)_{\text{Human}},$$

where g indexes combinations of *age*, *gender*, *education*, *income*, *and treatment*.

On average in this bucket, does the LLM over- or under-state the outcome vs humans, and by how much?

Bias Correction by Model

	Average Bias (Δ_g)		
Model	**Mean**	**Std**	R^2
DeepSeek	−0.327	0.554	0.282
Gemma	**0.194**	0.598	**0.363**
GPT-4.1	0.494	0.464	0.152
GPT-4o	**0.000**	0.643	**0.422**
GPT-4o mini	−0.368	0.532	0.261
GPT-o3	0.082	0.647	0.159
Llama	−0.365	0.521	0.295
Qwen	−0.224	0.596	0.358

Notes: Mean bias is $\Delta_g = \widehat{y}_g^{LLM} - y_g^{Human}$. R^2 *is from the projection* $\Delta_g = W_g^\top \theta + \varepsilon_g$. *Bold values indicate best-performing open-weight models.*

Probit Regression Results: Withdrawal Propensity

Variable	Coefficient	Std. Error	p-value	95% CI	
				Lower	Upper
Age	-0.0260^{***}	0.0045	0.0000	−0.0348	−0.0173
Male	−0.0177	0.0126	0.1609	−0.0425	0.0071
Income	0.0285^{***}	0.0046	0.0000	0.0194	0.0376
Education	0.0421^{***}	0.0054	0.0000	0.0314	0.0528
`Bank Email`	-1.5561^{***}	0.1805	0.0000	−1.9100	−1.2023
`Bank Text`	-1.6933^{***}	0.1804	0.0000	−2.0469	−1.3397
`Cautionary Tale`	-0.7180^{***}	0.1899	0.0002	−1.0902	−0.3458
`Fed Message`	-1.6423^{***}	0.1804	0.0000	−1.9959	−1.2886
`Network Effect`	5.0013^{***}	0.1576	0.0000	4.6925	5.3101
`Placebo`	-0.8224^{***}	0.1876	0.0000	−1.1901	−0.4547
`President Message`	-1.5265^{***}	0.1807	0.0000	−1.8806	−1.1725

Mental Models

Table: Example Explanations by Factor (Withdrawal = Yes)

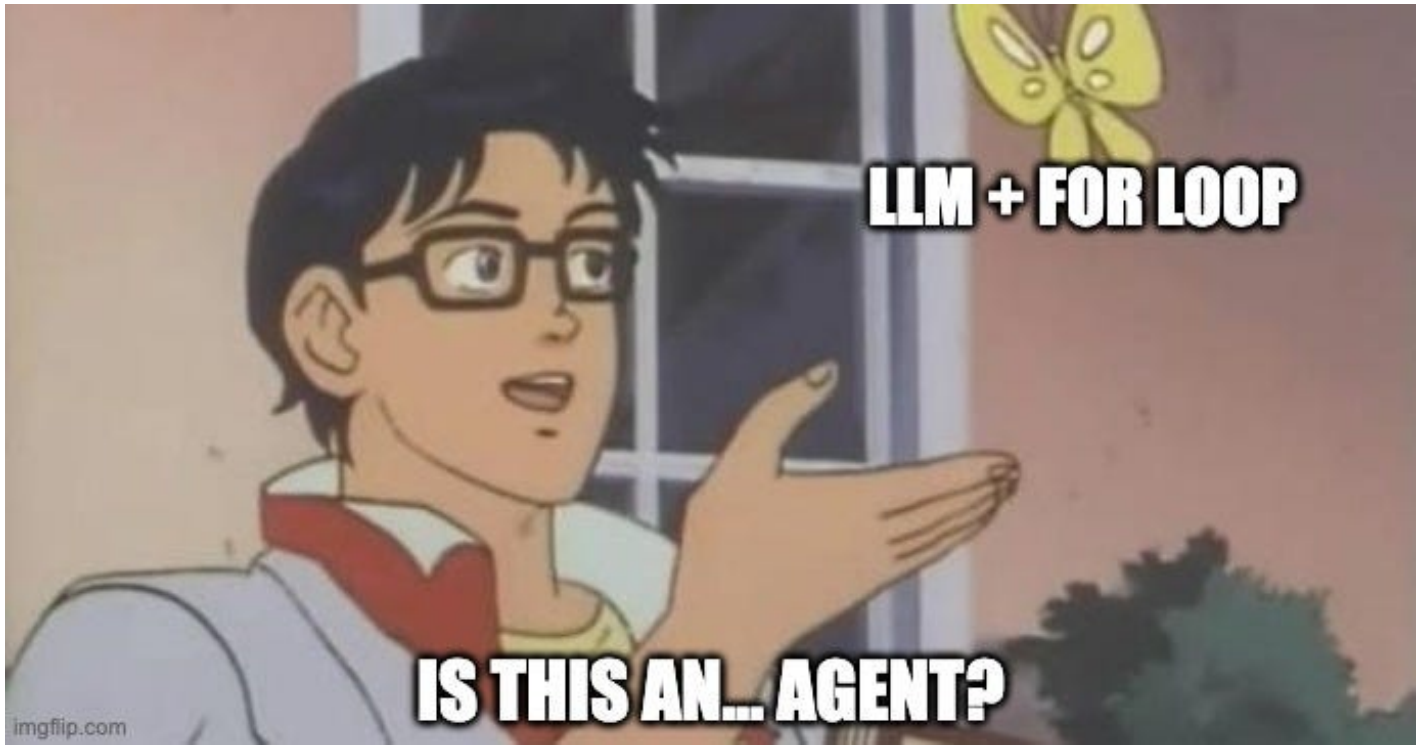
Factor	Human Examples	Gemma Examples
FDIC Insurance	It's not insured and there seem to be problems/instability with banks.	I'm worried about losing my savings, especially the uninsured portion. This tweet, even if it's alarmist, is making me nervous.
Risk and Safety	Not taking any risks with my money.	I'm close to retirement and can't risk it; I'll take my money out now.
Trust and Doubt	I'd rather have my money in my pocket than trust anybody else, even a bank's.	I can't afford to lose what little I have. That tweet scared me and I don't trust the bank's email.
Information/Evidence	News article and a tweet are no coincidence; even if I don't withdraw all my money, I'd take out a substantial amount.	That tweet scared me, even though the news says things are fine.
Financial Constraints	I live paycheck to paycheck; I can't afford to lose what little I have.	I don't understand all that talk about the markets, but I can't afford to lose my money. Better safe than sorry.
Social/Panic	If it's spreading beyond a couple of local banks, it's not a coincidence–rumors become reality when enough people panic...	Seeing everyone freaking out online about another bank is scary; I don't want to risk losing my savings.
Institutional Credibility	I can't trust this–don't you know better than to trust the government? I'd take enough money out...	I don't trust it when the bank says they're fine; too many people are saying bad things.
Personal Experience	I've had problems with my bank in the past and don't trust them on a normal day; that tweet would make me withdraw for peace of mind.	I'm young and don't have a lot of financial experience. That tweet scared me, and I don't want to risk losing my savings...

Conclusion

- Age, income, and education significantly shape withdrawal decisions during panic scenarios.
- Authoritative messages, especially from banks, the Fed, and the President, substantially reduce withdrawal intent.
- The impact of each message varies across population groups, highlighting the value of targeted messaging strategies.

This framework shows how validated LLMs can serve as scalable proxies for human behavior in high-stakes financial settings, allowing for testing of crisis communication strategies under uncertainty.

AI Agents



From *Homo Economicus* to LLM Agents

An AI agent is a software system that uses AI to pursue goals by interacting with an environment, maintaining state, making decisions, and taking actions.

- **Traditional economic models:** agents optimize explicit objectives subject to constraints; behavior is highly structured and analytically tractable.
- **Agent-based models (ABMs):** heterogeneous agents follow behavioral rules that may incorporate bounded rationality, learning, heuristics, or even zero-intelligence behavior.
- **LLM agents:** context-sensitive, language-native agents that can make decisions, communicate, plan, and interact with other agents or tools.

LLM agents can serve as *computational models of human decision makers* ("HOMO SILICUS") without requiring fully optimizing behavior. Whether their behavior is meaningfully human-like is an empirical question.

Agents for Behavioral & Computational Economics

- **Behavioral economics:** Human decisions often rely on heuristics and exhibit systematic departures from fully rational benchmarks; LLM agents can reproduce some of these patterns, though not consistently.
- **Computational economics:** LLMs can enrich ABMs by allowing agents to process unstructured information, communicate in natural language, and adapt their behavior to context.
- *Consequence*: simulations can represent richer mechanisms, including learning, social influence, communication, and norm formation.

What Is Actually Inside an AI Agent?

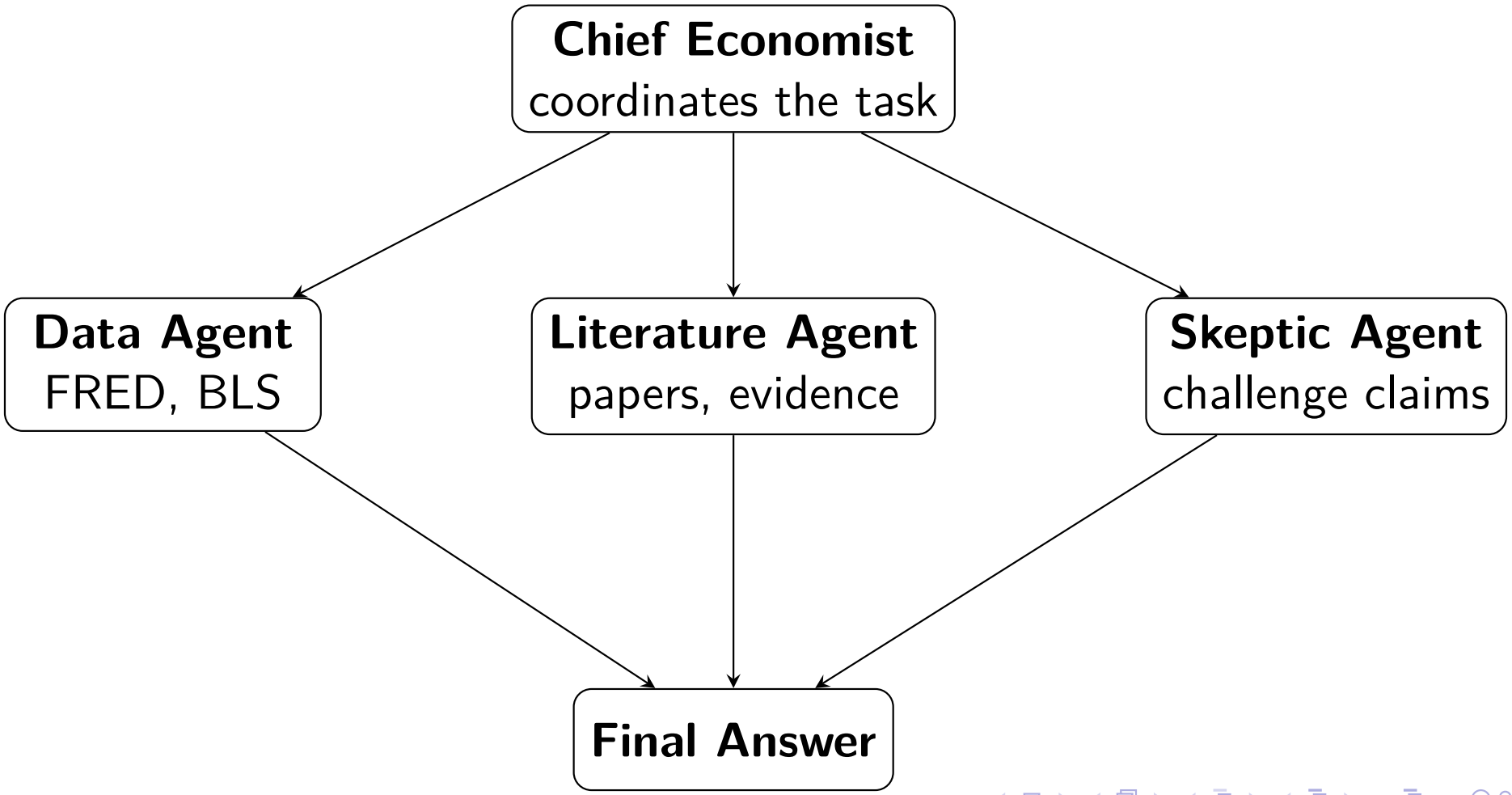
$$\boxed{\text{Agent}} = \boxed{\text{LLM}} + \boxed{\text{Instructions}} + \boxed{\text{State}} + \boxed{\text{Tools}} + \boxed{\text{Loop}}$$

- **LLM**: makes a decision
- **Instructions**: defines the role / objective
- **State**: remembers what has happened
- **Tools**: search, code, APIs, databases
- **Loop**: observe $\rightarrow$ decide $\rightarrow$ act

The "agent" is mostly **orchestration around an LLM**.

From One Agent to a Team of Agents

Task: Why did inflation rise in 2022?



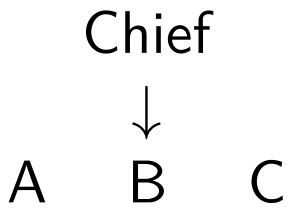
Who Writes the Loop?

Framework	Think of it as...	What it helps you do
LangChain	Agent toolkit	Models, prompts, tools, state, agent loops
LangGraph	Flowchart / state machine	Explicitly control who acts next, what state changes, and when to stop
OpenAI Agents SDK	Lightweight agent runtime	Tools, handoffs, manager agents, guardrails, tracing
AutoGen	Team / group chat	Agents communicate, take turns, select speakers, and hand off tasks
CrewAI	Organization chart	Give agents roles and tasks and specify a sequential or hierarchical process

These frameworks do not give you the economic model, but the **infrastructure for running it**.

Three Ways Agents Can Work Together

1. Manager



One agent delegates and synthesizes.

2. Handoff

$$A \rightarrow B \rightarrow C$$

Control moves from one specialist to another.

3. Discussion

$$\begin{array}{ccc} A & \leftrightarrow & B \\ \updownarrow & & \updownarrow \\ C & \leftrightarrow & D \end{array}$$

Agents deliberate or negotiate.

The important modeling choice is the **interaction structure**, not the infra.

Augmenting Survey Data

Survey Data

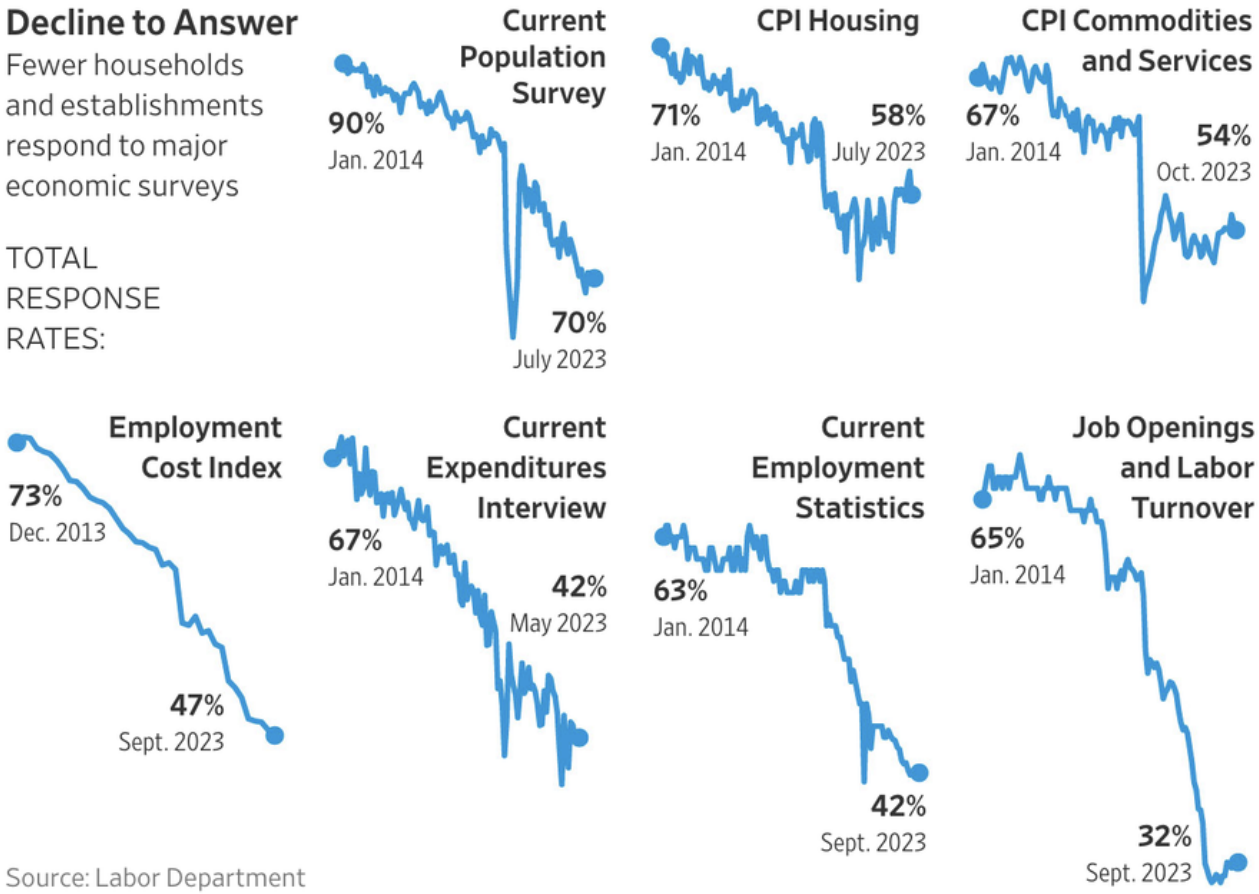
Augmenting Survey Data with Generative AI: An Application to Economic Research

By Erik Brynjolfsson, José Ramón Enríquez, Sophia Kazinnik, and David Nguyen*

We study how large language models (LLMs) can potentially augment survey-based data collected from human subjects, by focusing on two applications: (1) estimating willingness-to-accept (WTA) for giving up digital and analog goods, and (2) predicting personal income levels. We find that supplying LLMs with rich contextual data beyond demographics significantly improves predictive accuracy. Model fine-tuning and retrieval-augmented generation (RAG) further enhance performance, while changes to model temperature or prompting strategies yield only marginal improvements. Performance varies across goods studied and demographic groups. We provide a methodological blueprint for deploying LLMs as a fast, low-cost multiplier of survey coverage. This is particularly relevant in times of rapidly declining survey response rates.
JEL: C8, C9, C45, C63, C83
Keywords: Large Language Models; Generative Artificial Intelligence; Economic Measurement; Simulated Economic Agents; Survey Design

Motivation

Surveys are expensive, slow, and fewer people are responding:



Motivation

Survey data are getting worse:

- Demographic patterns of non-response (some groups are virtually unreachable)
- Dependence on online platforms (e.g., MTurk, Prolific) – "professional respondents", identity misrepresentation
- Respondents are using LLMs... (Zhang et al 2025, Veselovsky et al 2023

Motivation

Survey data are getting worse:

- Demographic patterns of non-response (some groups are virtually unreachable)
- Dependence on online platforms (e.g., MTurk, Prolific) – "professional respondents", identity misrepresentation
- Respondents are using LLMs... (Zhang et al 2025, Veselovsky et al 2023

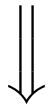
Does GenAI provide a pathway to overcome these survey limitations?

Research Question

Under what conditions can LLMs reliably replicate human survey responses?

Research Question

Under what conditions can LLMs reliably replicate human survey responses?



We evaluate LLM-generated responses on two economic survey tasks and compare them to real survey data and standard models

This Paper

Systematically vary inputs of interest:

- Type of data
- Prompt
- Foundation model
  - Workhorse and "reasoning" models; close-weights and open-weights
- Model parameters and input adjustments
  - e.g., temperature, finetuning, RAG

This Paper

Systematically vary inputs of interest:

- Type of data
- Prompt
- Foundation model
  - Workhorse and "reasoning" models; close-weights and open-weights
- Model parameters and input adjustments
  - e.g., temperature, finetuning, RAG

We benchmark LLM-estimates to:

- Statistical models
  - i.e., OLS, multinomial logistic regression
- Prediction by humans (one-shot, and test-retest)

Approach: Experimenting with LLMs

Data Sources:

- **Private:** UK WTA data from Coyle and Nguyen (2023) – Categorical outcome
- **Public:** Panel Study on Income Dynamics (PSID) – Continuous outcome

Approach: Experimenting with LLMs

Data Sources:

- **Private:** UK WTA data from Coyle and Nguyen (2023) – Categorical outcome
- **Public:** Panel Study on Income Dynamics (PSID) – Continuous outcome

Prompt Design:

- **LLM Persona:**
  - Expert advisor
  - Simulated user's own perspective
- **Prompting Strategy:**
  - Add step-by-step reasoning
  - Include only demographic information
  - Add "benchmark" context:
    - e.g., WTA: Valuations for other goods/services ($\neq p$)

Approach: Experimenting with LLMs

LLM Model Variants:

- **Model Type:**
  - OpenAI models (proprietary)
  - Llama models (open-weights)
- **Fine-tuning Approaches:**
  - Fine-tuning
  - Retrieval-Augmented Generation (RAG)
- **Hyperparameter:**
  - Temperature tuning

Prompt Example for WTA

▸ Step-by-Step

You are an expert assessing a survey of UK respondents in [year].

Consider a respondent with the following characteristics:

[demographics]

The individual provided the following responses for the amount they are willing to accept for not using each of the goods for [N] months:

- [good1] at [valuation1]
- [good2] at [valuation2]
- ...

Your task is to estimate the smallest sum of money for which the individual would be willing to go without the following good for [N] months:

- [product or service]

To answer, select one of the following ranges (all in GBP):
[pre-determined set of options: 'Don't have/Don't Use', '£1-10', '£11-20', '£21-50'...]

Baseline accuracy

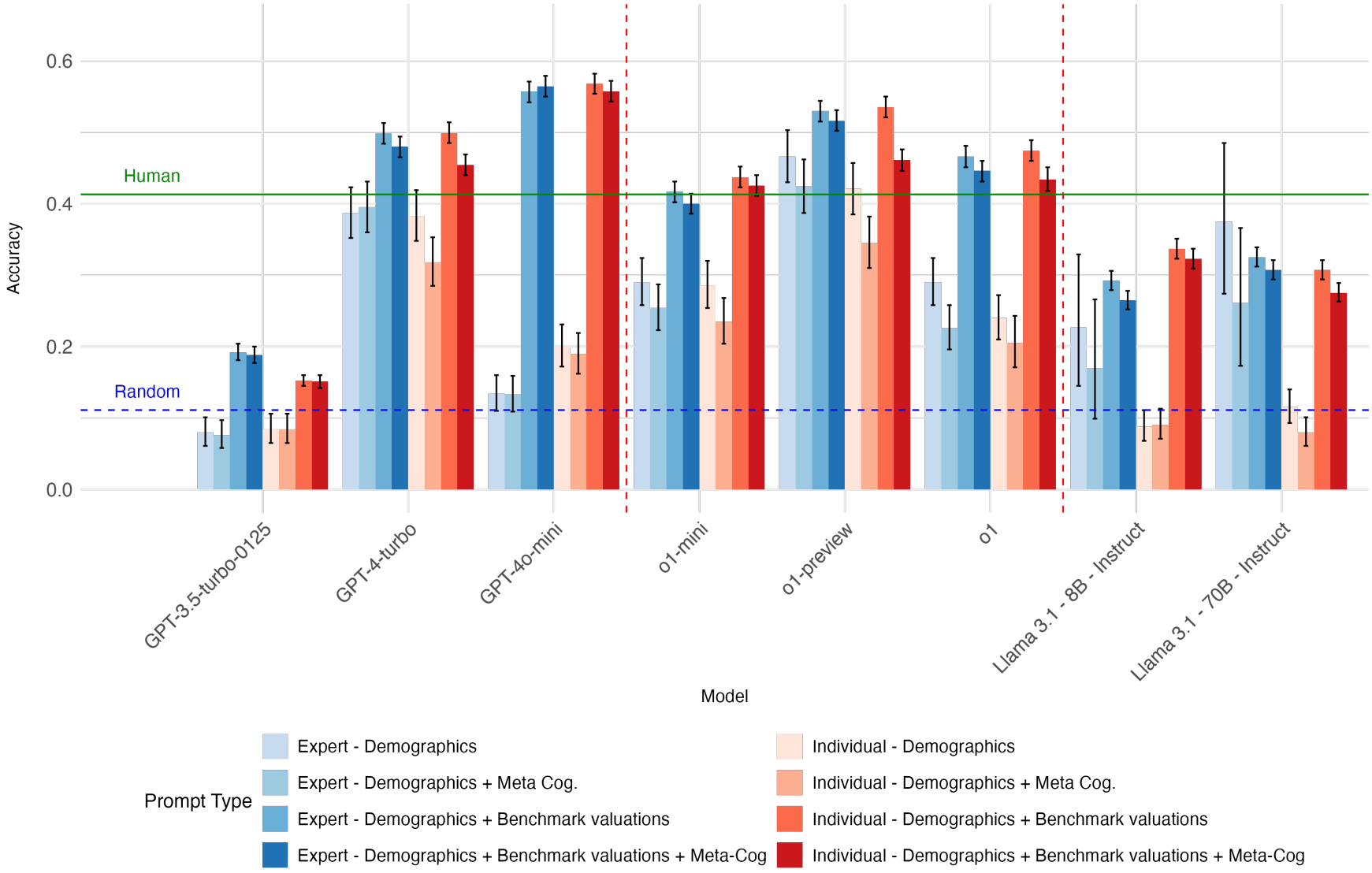
Random: Randomly choosing one of the valuation categories.

Human Predictions: Representative sample (age, gender, ethnicity) of 300 participants in the UK (2025). Each participant predicted the valuations of 4 profiles based on demographics and benchmark valuations.

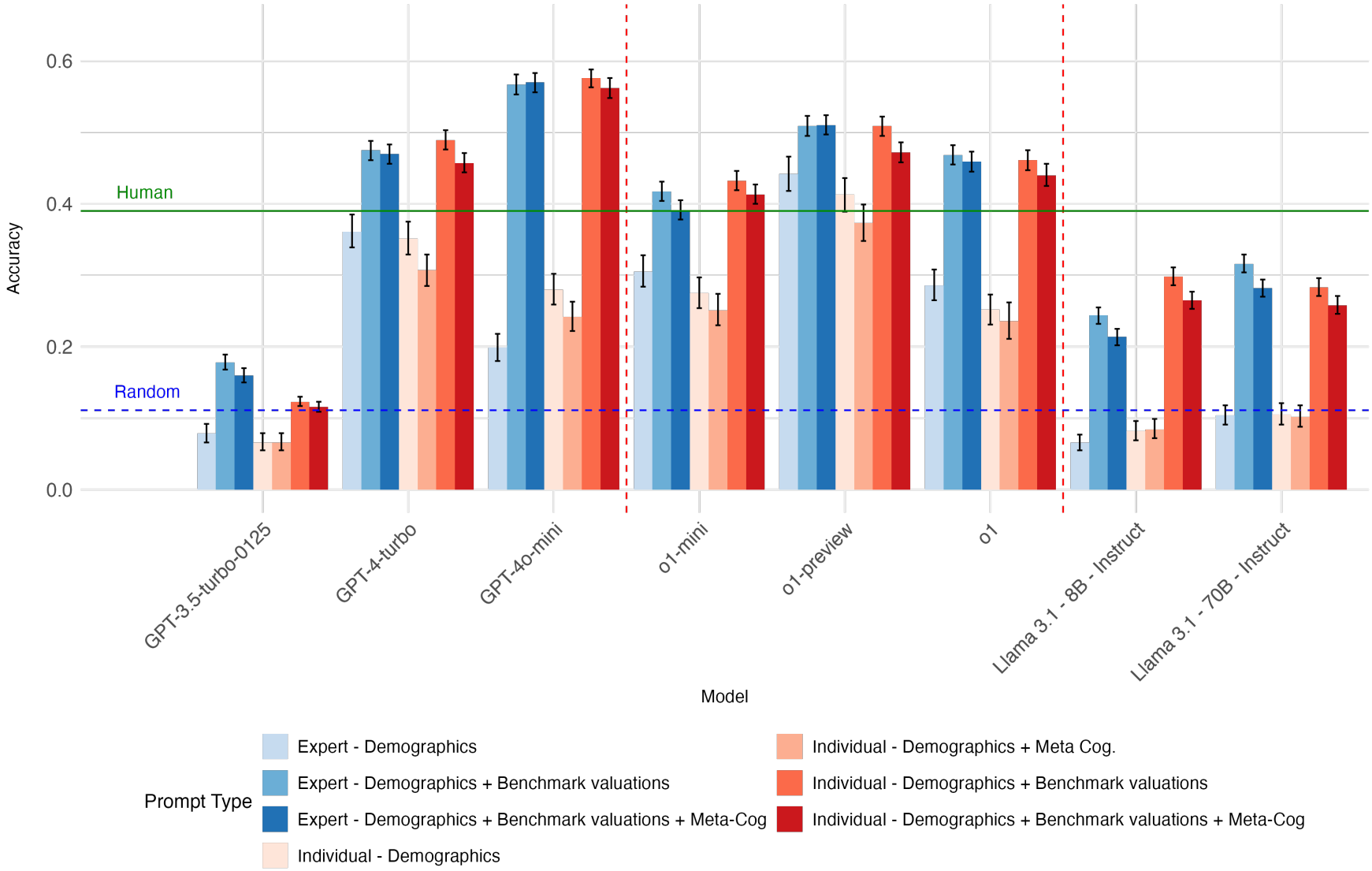
	1 month	12 months
Random	0.111	0.111
Human Predictions	0.413 (0.356 – 0.470)	0.390 (0.334 – 0.446)

Notes: 95% confidence intervals in parentheses.

Model Accuracy (1-month)



Model Accuracy (12-month)



Model Accuracy: Fine-tuning and RAG

Table: Performance on 1-month WTA Valuations (Expert Prompt)

Condition & Metric	**GPT-4o-mini**	**+ RAG**
Demographics + Benchmark Valuations		
Accuracy	0.652 (Δ 17.06%)	0.607 (Δ 8.98%)
Demographics + Benchmark Valuations + Reasoning Instructions		
Accuracy	0.642 (Δ 13.82%)	0.600 (Δ 6.38%)

A Note on RAG

Retrieval-Augmented Generation (RAG) is a technique that combines a language model with a search component.

Instead of relying solely on the model's internal knowledge, RAG retrieves relevant external documents (e.g., from a database or corpus) and incorporates them into the prompt before generating a response.

Why use RAG?

- Enhances accuracy and factual grounding
- Reduces hallucinations
- Allows the model to remain up-to-date without retraining

Overview of Findings

Accuracy increases with the value of data provided to the LLMs:

- Demographics + "benchmark" data > "benchmark" data > demographics
- Fine-tuning and RAG can give deliver a significant boost
- Prompt techniques and temperature are less relevant

Overview of Findings

Accuracy increases with the value of data provided to the LLMs:

- Demographics + "benchmark" data > "benchmark" data > demographics
- Fine-tuning and RAG can give deliver a significant boost
- Prompt techniques and temperature are less relevant

Estimates reflect significant heterogeneity:

- Simpler models can outperform "reasoning" models, depending on the task
- High accuracy for some (digital) goods: ≥ 0.90
- Valuations for young, wealthy and educated respondents might be overestimated
- High accuracy scores for some underrepresented populations (e.g., low-income individuals and those aged over 65)

Best Practices

1. Ground agents using granular benchmarks (other WTA / last-wave income).
2. Choose open weights models for stability; fine-tune when possible; RAG is context-specific.
3. Validate vs. humans: report accuracy vs. human and a test–retest ceiling.
4. Monitor heterogeneity: know where agents over/under-estimate.
5. Correct bias before using at scale.

Questions?

Day 4 wrap-up