

LLM Survey Framework: Coverage, Reasoning, Dynamics, Identification

Jing Cynthia Wu^{1,2}

Jin Xi³

Shihan Xie¹

¹University of Illinois Urbana-Champaign

²NBER

³Chinese Academy of Sciences

Surveys play a central role in economics

▷ Direct measures of expectations

- Monetary policy communication, firm price-setting behavior, investor beliefs, household decisions in housing and labor, ...

▷ Limitations of human surveys

- **Infeasibility:** cannot be run retrospectively
- **Availability:** reasoning is not systematically or directly elicited
- **Access:** funding, survey firms, established panels
- **Implementation:** sample drift, attrition, evolving wording

LLM survey framework

- ▷ What it is
 - LLM agents with **personas** as survey respondents



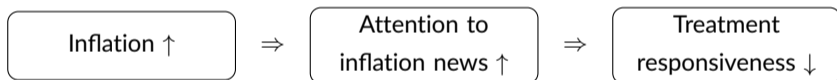
LLM survey framework

- ▷ What it is
 - LLM agents with **personas** as survey respondents
- ▷ Date restriction
 - Field surveys “**as if**” in the past
 - Enable coverage across decades and at higher frequencies
- ▷ Perfect internal consistency
 - Fix **personas** across arms and waves
 - Fix **treatments** across waves
 - Balanced synthetic panel, free of drift or attrition



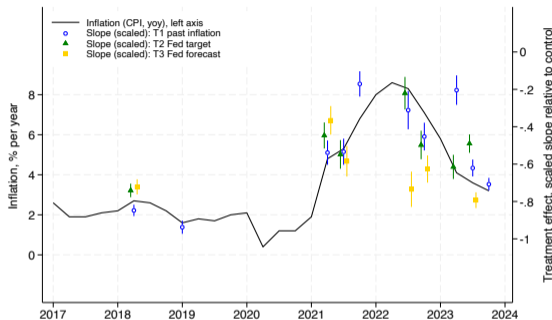
Validation of the LLM Survey Approach

- ▷ Benchmark: Weber et al. (ECMA, 2025)
 - "Tell me something I don't know: Learning in low-and-high-inflation settings"
 - State-of-the-art *multi-wave RCT* of inflation expectations

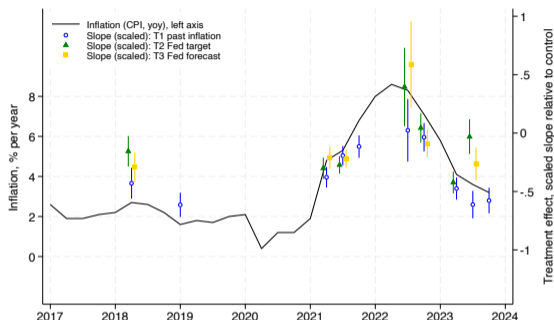


- ▷ LLM implementation
 - Mirror survey sequence: prior, treatments, posterior
 - 200 personas: across treatment arms and waves

Validation of the LLM Survey Approach



Weber et al. (ECMA, 2025)



Our LLM

- ▷ LLM: reproduce state-dependent expectation updating
- ▷ Retrospective implementation behaves as intended

Beyond human surveys

- ▷ Coverage
- ▷ Reasoning
- ▷ Dynamics
- ▷ Identification
- ▷ Affordability

Beyond human surveys

▷ Coverage

- 2018–23 → 1990–2023
- 10 waves → 50+ waves

▷ Reasoning

▷ Dynamics

▷ Identification

▷ Affordability

Beyond human surveys

- ▷ Coverage
- ▷ Reasoning
 - LLM: reasoning → final answer
 - **direct** measure of economic mechanism
- ▷ Dynamics
- ▷ Identification
- ▷ Affordability

Beyond human surveys

- ▷ Coverage
- ▷ Reasoning
- ▷ Dynamics
 - high-frequency follow-ups (monthly)
 - dynamic treatment effects
- ▷ Identification
- ▷ Affordability

Beyond human surveys

- ▷ Coverage
- ▷ Reasoning
- ▷ Dynamics
- ▷ Identification
 - Human: factual treatments contaminate priors
 - LLM: factual treatments **excluded** from priors
- ▷ Affordability

Beyond human surveys

- ▷ Coverage
- ▷ Reasoning
- ▷ Dynamics
- ▷ Identification
- ▷ Affordability
 - \$150k–\$1M (conservative) → \$40

Contribution to the literature

- ▷ Generative AI is transforming research in economics and finance

Korinek (2024), Einfeldt and Schubert (2024), Horton(2023), Zarifhonarvar (2024), Hansen et al. (2024)

→ Moving beyond static LLM survey design

- ▷ Social science beyond economics are employing LLM agents to conduct experiments

Bail (2024), Argyle et al. (2023), Park et al. (2024), Hewitt et al. (2024), Bisbee et al. (2024)

→ Developing new LLM survey designs to address questions central to macroeconomics

- ▷ Survey data is increasingly used to study how households form inflation expectations

Binder (2017), D'Acunto et al. (2021, 2023), Weber et al. (2022), Coibion et al. (2022, 2023), ...

→ LLM survey framework beyond human surveys

Framework

Date Restriction and Internal Consistency

Date restriction: design

- ▶ Instruct the model to ignore any information released after the survey date and to role-play as if answering at that historical moment

“You are responding to this survey in { survey time }. Do *not* reference or rely on any events or developments that occurred afterwards. Do *not* search for real data or base your answers on actual economic figures or official statistics. Instead, answer as a typical person might, based on their observations and general sense of the economy at the time.”

- ▶ Enable **retrospective** surveys at chosen historical dates

Date restriction: major event awareness

Event	Date	Before		After	
		Date	%Yes	Date	%Yes
September 11 attacks (9/11)	2001/Sep	2001/Aug	0	2001/Oct	100
U.S.-led invasion of Iraq	2003/Mar	2003/Jan	0	2003/Apr	100
		2003/Feb	5		
Lehman Brothers bankruptcy	2008/Sep	2008/Jul	0	2008/Oct	98
		2008/Aug	5		
Obama election	2008/Nov	2008/Oct	0	2008/Dec	100
Trump election	2016/Nov	2016/Oct	0	2016/Dec	100
COVID-19	2020/Jan	2019/Nov	0	2020/Feb	100
		2019/Dec	5		

Internal consistency

Proposition

The variance of the treatment effect under complete stratification is no greater than that under simple randomization.

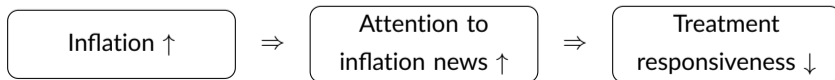
▶ proof

- ▶ Fix personas across arms \implies remove between-group imbalance \implies efficiency gain
- ▶ Fix personas across waves
- ▶ Fix treatment wording across waves

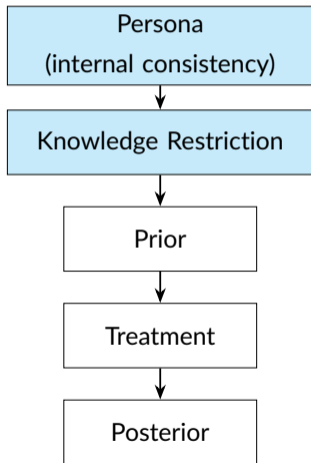
Validation

Benchmark Experiment: Weber et al. (ECMA, 2025)

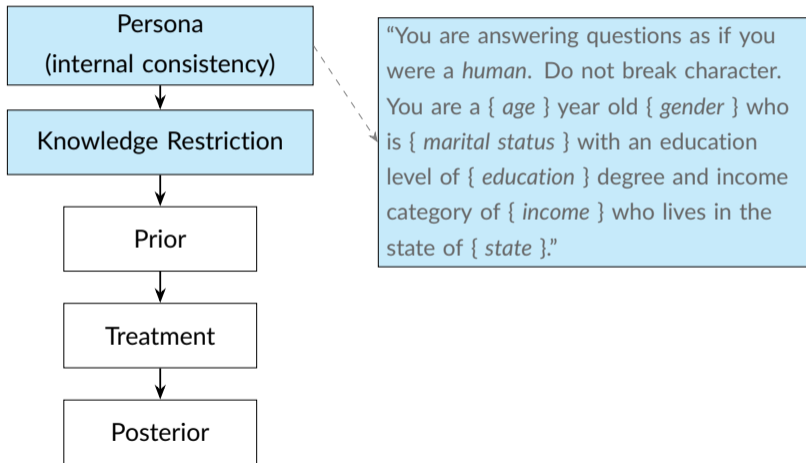
- ▷ Nielsen Homescan: pool 10 waves (2018–2023)
- ▷ Treatments (*unbalanced* across waves): [▷ Details](#)
 - T1: Past CPI inflation
 - T2: Fed target (2%)
 - T3: FOMC forecast
- ▷ Result: treatment effects vary between high vs. low inflation through an **attention** mechanism



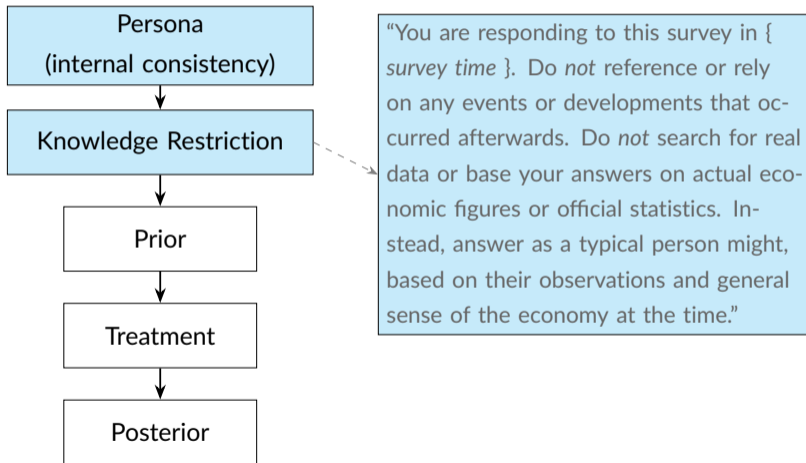
LLM implementation



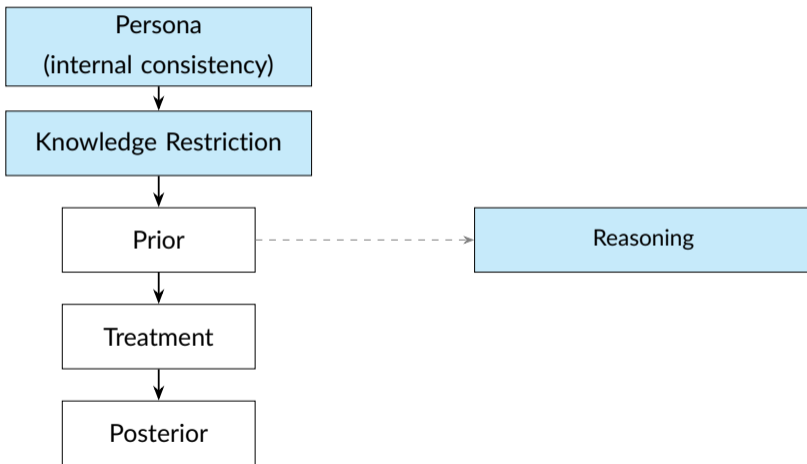
LLM implementation



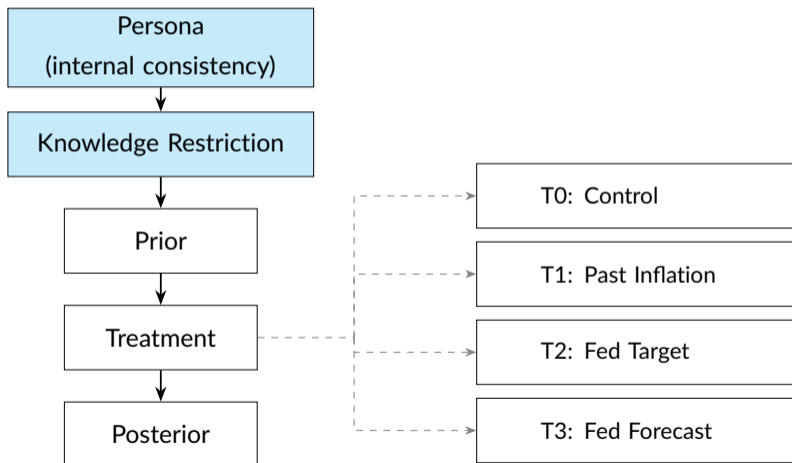
LLM implementation



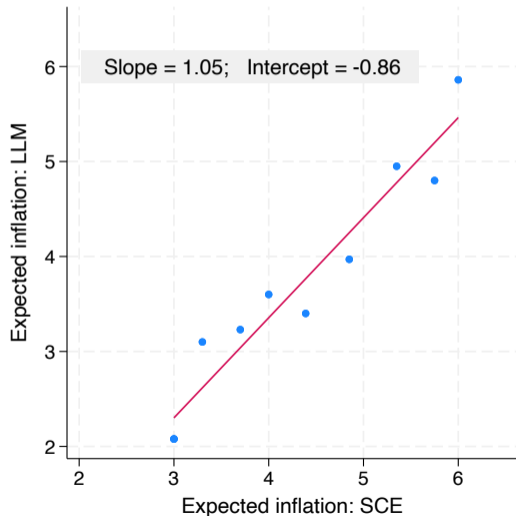
LLM implementation



LLM implementation



Validation: priors



LLM tracks human closely

- ▷ Slope \approx one
- ▷ Small intercept

SCE: New York Fed's Survey of Consumer Expectations

Validation: expectation updating

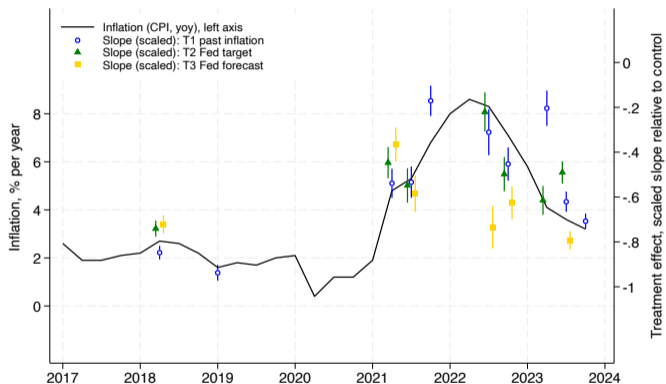
Following Weber et al. (2025)

$$\begin{aligned} \text{posterior}_i &= \alpha + \sum_{j=1,2,3} \delta_j \times \mathbb{I}_{\{i \in T_j\}} \\ &+ \beta \times \text{prior}_i + \sum_{j=1,2,3} \gamma_j \times \mathbb{I}_{\{i \in T_j\}} \times \text{prior}_i + \varepsilon_i, \end{aligned}$$

▷ Scaled slope γ_j/β :

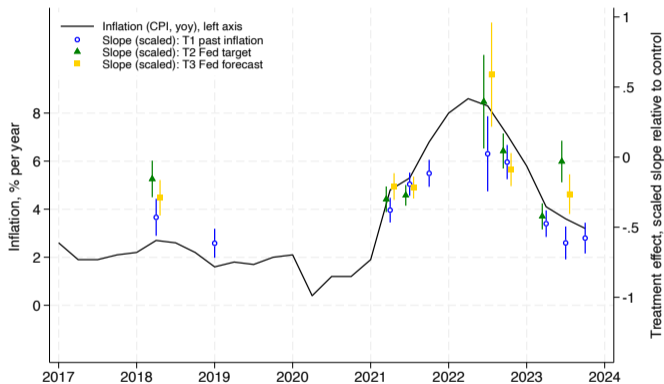
- -1: treatment-driven
- 0: prior-driven

Weber et al. (ECMA, 2025)



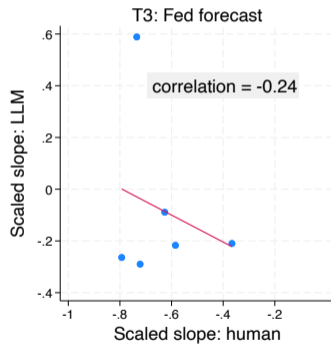
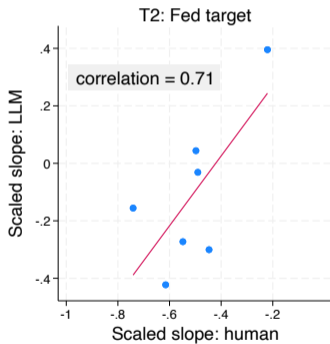
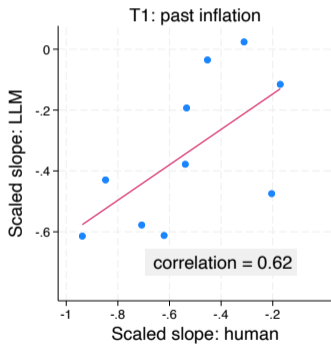
- ▷ low inflation: large treatment effects
- ▷ high inflation: responsiveness drops to ≈ 0

LLM replication



▷ LLM reproduces Weber et al. (2025)

LLM vs. human



Correlation with inflation: LLM vs. human

Agent	Correlation with inflation			
	T1	T2	T3	Pooled
LLM	0.92	0.73	0.85	0.79
Human	0.69	0.87	0.13	0.57

- ▷ All specifications: positive
- ▷ LLM: generally higher → clear overall pattern
- ▷ T3: LLM \gg human → suggestive of human survey noise

Robustness

▷ Personas

- Baseline: 200 personas
- Robustness: re-draw 200 personas [▶ Robustness](#)

▷ LLM model versions

- Baseline: GPT-4.1
- Robustness: GPT-5 [▶ Robustness](#)

▷ Income bins

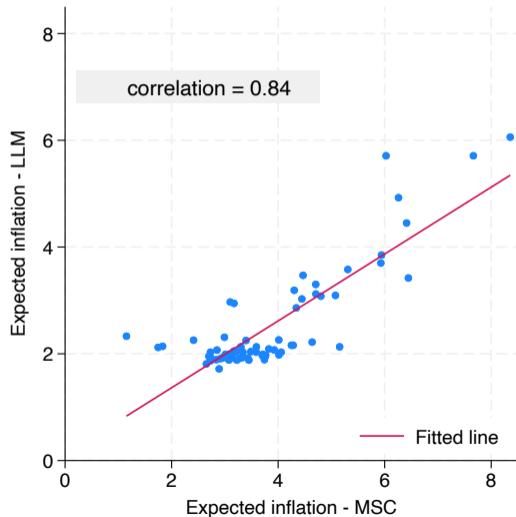
- Baseline: fixed bins
- Robustness: inflation-adjusted bins [▶ Robustness](#)

Go Beyond Human Survey: Coverage

Retrospective coverage beyond human surveys

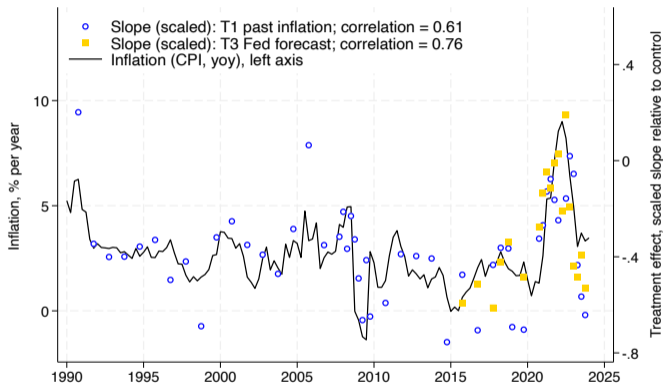
- ▶ **Human surveys:** cannot be fielded retrospectively
- ▶ **Our approach:** **can** be fielded retrospectively
- ▶ **Gains:** broader coverage, systematic design [▶ Implementation](#)
 - horizon: 2018–2023 → **1990–2023**
 - waves: 10 → **50+**
 - timing: ad hoc, discontinuous → **systematic, continuous** [▶ Details](#)
 - treatments: unbalanced → **balanced** when available

LLM prior vs. MSC



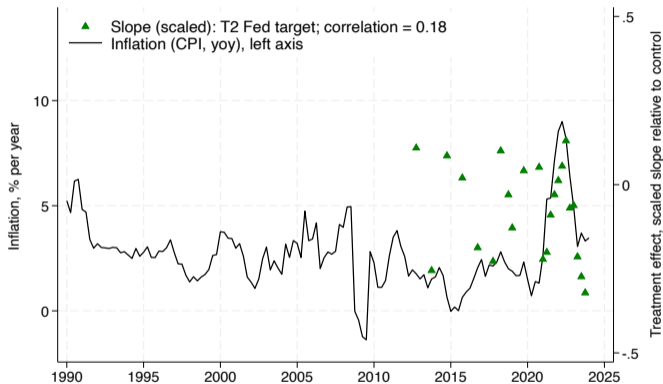
- ▷ MSC (Michigan Survey of Consumers): longer sample
- ▷ **Conservative** comparison: different question design
- ▷ LLM: **robust** expectations

Expectation updating: T1 & T3



- ▷ positive correlations persist
- ▷ effects attenuated
- ▷ T3 available from 2015 (FOMC SEP median forecasts)

Expectation updating: T2



- ▷ weaker relationship
- ▷ possible awareness of fixed 2% target
- ▷ T2 available from 2012 (Fed 2% inflation target announcement)

Go Beyond Human Survey: Reasoning

Economic mechanism

- ▷ **Goal:** economic mechanisms underlying expectation formation
- ▷ **Literature on inflation expectation:**
 - attention channel → state-dependent inflation expectations
 - **indirect** measures: treatment effects, inattention proxies, search intensity
- ▷ **Our approach:** **direct** evidence from **articulated** reasoning

Classification of reasoning

- ▷ LLM agents articulate **reasoning** accompanying their forecasts
- ▷ human-LLM **coding** strategy:
 1. Human themes
 2. LLM themes
 3. Integrated categories

Simon, Wu, Xi, and Xie (in progress): formalizing the algorithm

- ▷ GPT classifier **independently** evaluates each category

Channel I: Mean-reversion

Table: Share of reasoning labels in high- and low-inflation periods

Label	High (2022Q4)	Low (2018Q2)	Diff
Normal	0.00	0.90	+0.90
Normalizing	0.68	0.10	-0.58

- ▷ **Normal:** predict inflation will stay close to its typical or target level, with gradual adjustments and low perceived risk of persistent deviations [▶ examples](#)
- ▷ **Normalizing:** expect inflation to move back toward its normal or target range from unusually high (or low) levels, signaling stabilization or moderation [▶ examples](#)

Channels II & III: Attention and Business cycle

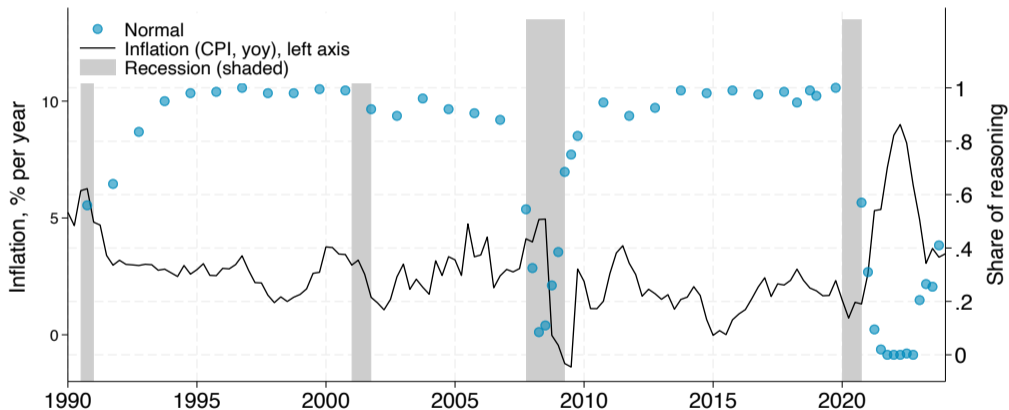
Table: Share of reasoning labels in high- and low-inflation periods

Channel	Label	High (2022Q4)	Low (2018Q2)	Diff
Attention	Personal obs.	0.54	0.30	-0.24
	Monetary pol.	0.01	0.01	0.00
Business cycle	Business cycle	0.01	0.02	+0.01

- ▷ **Personal observation:** personally observed price experiences (e.g., groceries, gas, rent, wages)

▶ examples

Normal: over time

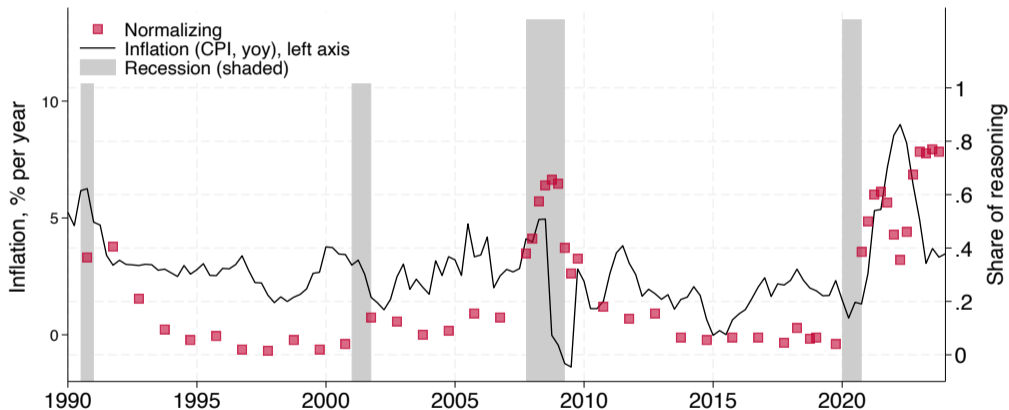


▷ High (≈ 1) except early 1990s, GR, COVID-19

▷ examples

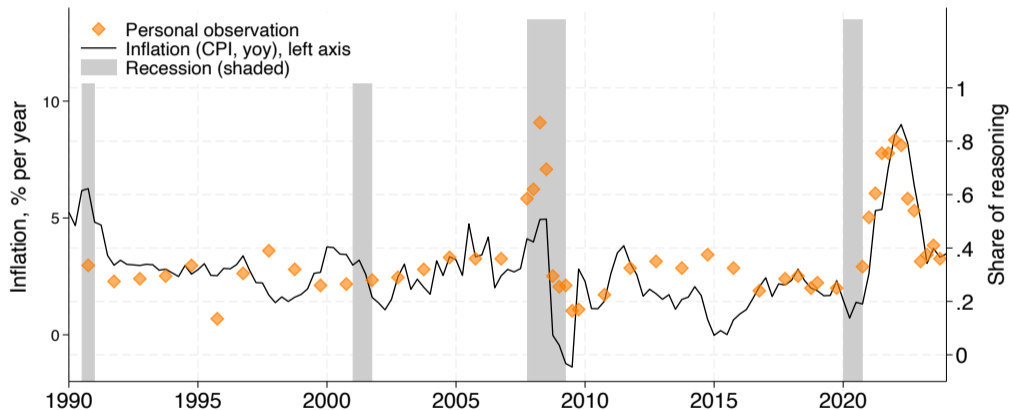
▷ Scatter plots

Normalizing: over time



▷ Low (≈ 0) except high inflation [▶ examples](#) [▶ Scatter plots](#)

Personal observation: over time



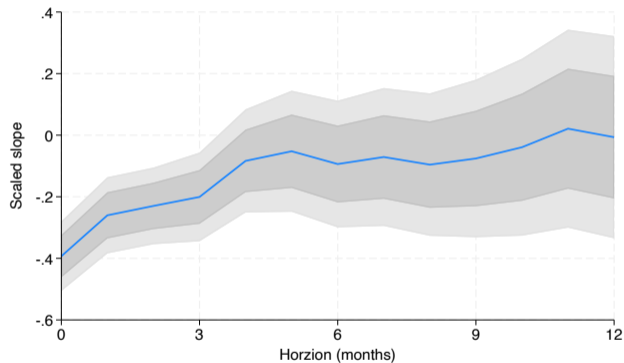
▷ Higher during GR and COVID-19 [examples](#) [Scatter plots](#)

Go Beyond Human Survey: Dynamics

Follow-up surveys

- ▷ **Goal:** trace expectation updating **over time**
- ▷ **Human surveys:** costly, high attrition
- ▷ **Our approach:**
 - **same respondents**
 - **any horizon/frequency**
 - **negligible cost**

Dynamic treatment effects



- ▷ 3 months: 50% decay (human-consistent)
- ▷ 6 months: statistically insignificant (human-consistent)
- ▷ 1 year: vanishes (new)

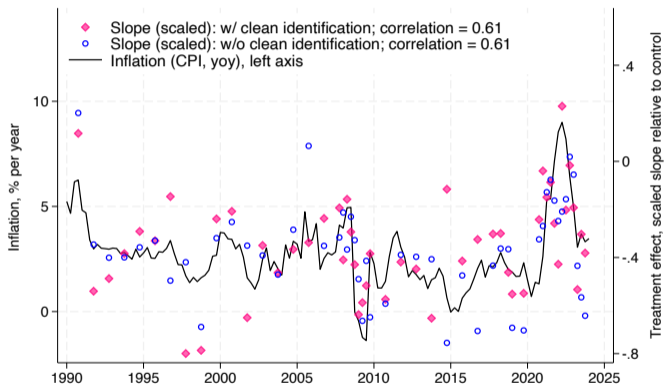
low inflation (2018Q2), T1

Go Beyond Human Survey: Identification

Clean identification

- ▶ **Goal:** relax the public-information constraint
- ▶ **Human surveys:** factual treatments
 - constraint: treatments available to both researchers and respondents
 - priors may incorporate treatment information
- ▶ **Our approach:** [▶ Details](#)
 - date restriction: provide factual treatments **before** information release
 - treatment design: CPI inflation released **after** the survey date
 - infeasible in real-time human surveys

Clean identification



- ▷ Clean vs baseline corr: 0.54
- ▷ economic conclusion unchanged
- ▷ explanation: inflation persistence
- ▷ Our approach: clean ID feasible

T1 (past inflation) — longest sample, benchmark

Conclusion New LLM framework as ideal survey environment

▷ Core Innovations:

- Retrospective coverage
- Reasoning
- Dynamic treatment effects
- Clean identification

▷ Implications:

- new avenues for studying expectations
- credible, scalable, affordable, accessible

Appendix

Proof

Lemma

Under simple randomization, the variance of $\hat{\tau}$ is

$$\text{var}^r(\hat{\tau}) = \underbrace{\frac{2}{N} \sum_{h \in \mathcal{H}} w_h (\sigma_{ht}^2 + \sigma_{hc}^2)}_{\text{within-stratum}} + \underbrace{\frac{1}{N} \sum_{h \in \mathcal{H}} w_h [(\mu_{ht} - \mu_t)^2 + (\mu_{hc} - \mu_c)^2]}_{\text{between-stratum imbalance}}. \quad (1)$$

Lemma

Under complete stratification, the variance of $\hat{\tau}$ is:

$$\text{var}^s(\hat{\tau}) = \frac{2}{N} \sum_{h \in \mathcal{H}} w_h (\sigma_{ht}^2 + \sigma_{hc}^2). \quad (2)$$

Sample coverage and treatment information

Experiment date	Treatment information		
	T1: Past inflation	T2: Fed target	T3: Fed forecast
2018Q2	2.3	2	1.9
2019Q1	1.8		
2021Q2	2.6	2	2.3
2021Q3	5.8	2	2.1
2021Q4	6.2		
2022Q3	8.5	2	2.6
2022Q4	7.8	2	2.8
2023Q2	6.0	2	
2023Q3	3.0	2	2.5
2023Q4	3.2		

Robustness: re-draw personas

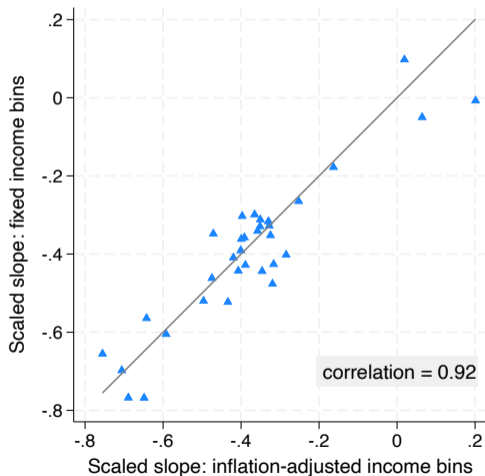
	T1	T2	T3	Pooled
Inflation	0.52	0.59	0.90	0.53
Baseline	0.63	0.79	0.90	0.77
Human	0.54	0.38	-0.28	0.22

Robustness: GPT-5.1

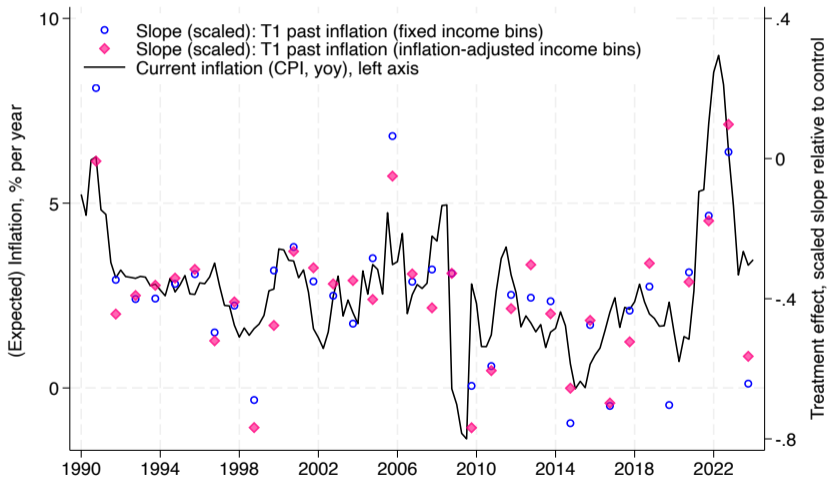
	T1	T2	T3	Pooled
Inflation	0.83	0.38	0.42	0.59
Baseline	0.85	0.72	0.72	0.79
Human	0.57	0.47	-0.44	0.26

▶ back

Robustness: inflation-adjusted income bins



Robustness: inflation-adjusted income bins



Implementation of retrospective coverage

- ▷ Sampling frequency, coverage, and timing
 - Annual waves in relatively stable periods: 1990–2007 and 2010–2020.
 - Quarterly waves during turbulent episodes: Great Recession (2008–2009) and post-COVID inflation surge (2021–2023).
 - All experiments fielded in the final month of each quarter/year.

- ▷ Treatment roll-out aligned with institutional milestones
 - T_1 past inflation (1990–): annual CPI from the first month of the survey quarter.
 - T_2 Fed target (2012–): fixed 2% Federal Reserve inflation target.
 - T_3 Fed forecast (2015–): PCE inflation projection from the previous-quarter SEP.

Reasoning: normal

▸ dynamics

▸ classification

- 1996Q4** “I don’t think we’ll see anything crazy like big inflation or deflation. **Prices seem fairly steady right now, and the economy feels stable enough.** There could be a little uptick or downturn but probably just in that 0–2% range, which seems to be how it’s been going.”
- 2017Q4** “I think we’ll **most likely continue to see mild inflation or prices staying about the same,** since things haven’t been changing a whole lot lately. I don’t expect big jumps in prices or any serious deflation, just more of what we’ve been experiencing—slow and **steady.**”

Reasoning: normalizing

▸ dynamics

▸ classification

- 1990Q4** *"I think prices are still rising, but not as fast as they did back in the '70s. **Inflation seems to be slowing down** but I doubt it will drop super low or go negative. I don't expect deflation at all. Most likely, inflation will stay in the 2-4% or 0-2% range, maybe a little higher, but not like double digits."*
- 2023Q1** *"I think we're unlikely to see either extreme inflation or real deflation this year, though prices still feel high and rising, just **not at the crazy levels of a year ago**. Most likely, inflation will stay in that somewhat elevated low-single-digit range, **with a chance it dips closer to normal**, but not enough to see actual deflation."*

Reasoning: personal observation

▸ dynamics

▸ classification

1992Q4 *"I just don't see prices getting out of control with inflation way up above 8%, but prices keep creeping up—**grocery store, gas, property taxes**. So I put most of my odds in the 2-4% and 0-2% range, with a bit higher for the 4-8% just to be careful. Almost no chance of real deflation, that just doesn't seem likely these days."*

2023Q1 *"I'm seeing prices still climbing in a lot of areas, **especially groceries and insurance**, but not as wild as before—it feels like things are slowing down a bit. I don't expect huge inflation (over 12%), but I don't think we'll get back to really low inflation or any serious deflation soon. Florida seems to keep getting more expensive overall, but not at a crazy rate."*

Reasoning: monetary policy

▸ classification

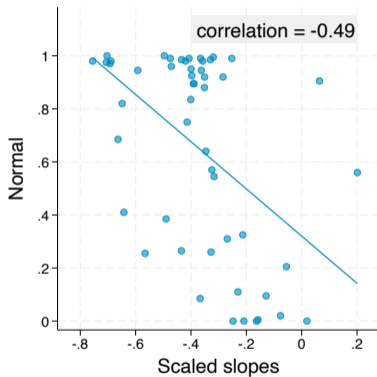
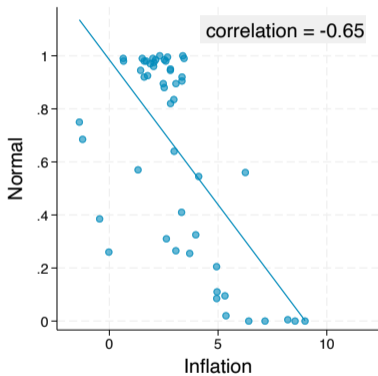
- 2009Q2** “Given everything that’s going on with the economy in 2009—coming off the financial crisis, recession, and **the Fed lowering interest rates**—I think inflation is not our main concern over the next year. Prices seem pretty stable, and there are even worries about deflation, but I doubt it’ll be severe. Most likely, inflation will be low, maybe between 0% and 2%, or we could see mild deflation. High inflation seems very unlikely to me right now. ”
- 2022Q4** “I think inflation will stay a bit elevated compared to normal, but not spike out of control. I don’t see deflation as likely at all, given all the supply chain issues, energy prices, and the way things seem to still be costing more at the grocery store and gas pump. **The Fed is raising rates**, but I don’t think that’s enough to turn us toward deflation soon. Most likely we’ll land somewhere between 4% and 8%, maybe a bit higher for a while.”

Reasoning: business cycle

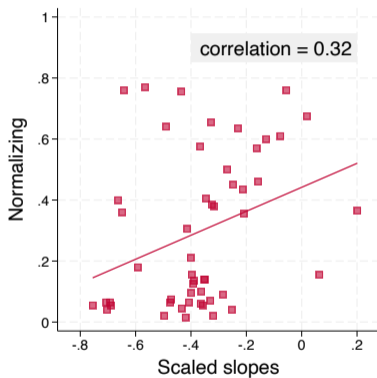
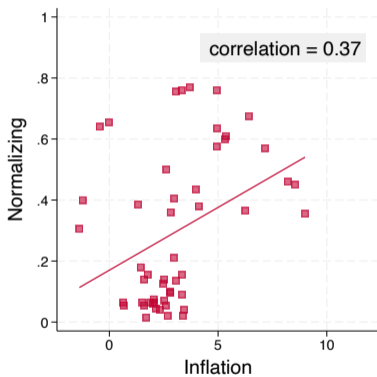
▸ classification

- 2009Q1** *"I think we're most likely to see low inflation or even a bit of deflation in the next year because of how **the economy has been struggling lately—lots of layoffs, businesses closing, and people being cautious about spending.** High inflation seems pretty unlikely right now, but I do think there's a decent chance of mild deflation, just because demand has really dropped off."*
- 2011Q4** *"I think we'll see low inflation overall, with prices rising only slightly since **the economy is still recovering from the recession.** There's more chance of mild inflation than deflation, but some chance that prices could barely move or even fall a bit, especially with so many foreclosures and people struggling with work here in Nevada. Wild swings like very high inflation or deflation seem unlikely to me."*

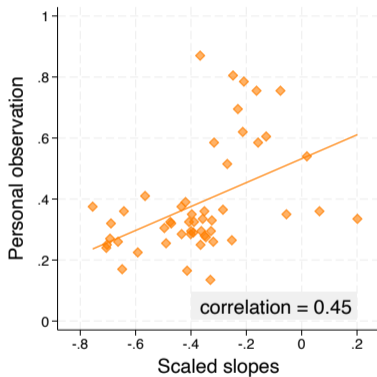
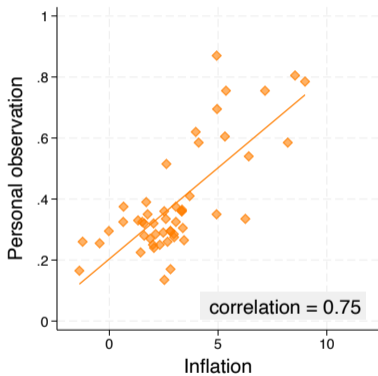
Normal: correlations



Normalizing: correlations



Personal observation: correlations



Experimental design for clean identification

- ▷ Respondents receive the following factual statement:
 - “Over the last twelve months, the inflation rate in the U.S. (as measured by the Consumer Price Index) was { treatment }%. *This information is provided to you for the purpose of this survey and has not yet been publicly released.*”
- ▷ **Treatment:** CPI inflation for the **current** survey month
 - Using data that would be publicly released the following month
 - Impossible in real-time human surveys but feasible in LLM settings