

Discussion of
*LLM Survey Framework: Coverage, Reasoning, Dynamics,
Identification*
by Wu, Xi, and Xie

Daniela Puzzello
Indiana University

Inflation: Drivers and Dynamics Conference 2026

Large Language Models in Economics and Finance: Three Emerging Research Agendas

1. Simulating Economic Behavior

- ▶ LLMs are increasingly used as synthetic agents to explore whether they can reproduce choices observed in laboratory experiments and strategic interactions
 - Horton et al. (2023), Aher et al. (2023), Park et al. (2023), Mei et al. (2024)...

2. Forecasting with LLMs

- ▶ LLMs are used as forecasting tools for macroeconomic variables, asset prices, earnings, and financial markets
 - Faria-e-Castro & Leibovici (2024), Carriero et al. (2024), Lopez-Lira & Tang (2023), Kim et al. (2024), Alam et al. (2025)...

Large Language Models in Economics: Three Emerging Research Agendas, Cont'd

3. Simulating Expectations

- ▶ LLMs are used to generate households, professional forecasters, CFOs and firms' expectations and to study how they respond to information, communication, and narratives
- Bybee (2023), Zarifhonarvar (2026), Hansen, Horton, Kazinnik, Puzzello & Zarifhonarvar (2024), [Wu, Xi & Xie \(2025\)](#), Puzzello & Zarifhonarvar (2026), Graham, Harvey, and Jha (2026)...

While LLMs are increasingly used to generate synthetic agents, we still know relatively little about which aspects of human expectations and behavior they capture well, where they systematically differ from humans, and for which applications they are most useful

Can LLMs Simulate Human Respondents?

- ▶ **Ambitious research agenda:** Can LLM agents complement human surveys for studying expectations?
- ▶ **This paper:** Develops an LLM-based survey framework with
 - ▶ retrospective coverage,
 - ▶ economic reasoning,
 - ▶ dynamic treatment effects,
 - ▶ “clean” experimental identification.
- ▶ **Application:** Household inflation expectations and randomized information treatments.
- ▶ **Why important?** Surveys are informative, but expensive and difficult to run retrospectively or at very high frequency.

Main Contribution

Feature	Prior work	This paper
Synthetic survey respondents	✓	✓
Historical / retrospective simulation	✓	✓
Controlled information sets	✓	✓
Randomized information treatments	✓	✓
Multi-wave RCT	limited	✓
Dynamic treatment effects	limited	✓
Reasoning / mechanisms	limited	✓

Examples: Bybee (2023); Hansen, Horton, Kazinnik, Puzzello, and Zarifhonarvar (2025); Zarifhonarvar (2026); Graham, Harvey and Jha (2026), Puzzello and Zarifhonarvar (2026)

My reading: the novelty lies in the combination and extensions of an emerging methodology.

Key Strengths of the Paper

- 1. Starts from a demanding human benchmark**
 - ▶ Multi-wave RCT rather than a one-shot survey (Weber et al., 2025)
 - ▶ Same treatments across different inflation environments
- 2. Takes the historical-counterfactual problem seriously**
 - ▶ Date-restricted prompt
 - ▶ Explicit tests around major historical events
- 3. Pushes beyond replication**
 - ▶ Long retrospective panel
 - ▶ High-frequency treatment dynamics
 - ▶ Open-ended reasoning
- 4. Potentially big gains in cost and scale**

Validation: Beyond Average Responses

The validation focuses primarily on **central tendencies** and **treatment-effect correlations** (not to be taken for granted).

Can LLM agents reproduce other important features of human beliefs?

- ▶ **Cross-sectional disagreement**

- ▶ Compare the full distribution of expectations, not only medians.
- ▶ Do LLM agents generate enough heterogeneity across respondents?

- ▶ **Individual uncertainty**

- ▶ Compare dispersion of subjective probability distributions.
- ▶ Does uncertainty vary with the macroeconomic environment as in human data?

- ▶ **Heterogeneity in expectations and treatment effects**

- ▶ Do different types of LLM agents respond differently in the same way humans do?

Matching the average response may be easier than matching the distribution of beliefs and responses.

How Far Can We Extrapolate from the Validation?

Human–LLM treatment effects

Treatment	Corr.
Past inflation	0.62
Fed target	0.71
Fed forecast	-0.24

The framework is used for:

- ▶ Historical counterfactual waves
- ▶ Monthly treatment dynamics
- ▶ “Clean” counterfactual experiments
- ▶ Economic reasoning/mechanisms

How Far Can We Extrapolate from the Validation?

Human-LLM treatment effects

Treatment	Corr.
Past inflation	0.62
Fed target	0.71
Fed forecast	-0.24

The framework is used for:

- ▶ Historical counterfactual waves
- ▶ Monthly treatment dynamics
- ▶ “Clean” counterfactual experiments
- ▶ Economic reasoning/mechanisms

If validation does not automatically transfer across treatments, why should we assume it transfers from numerical answers to reasoning, or from benchmark settings to counterfactual historical settings?

How Should We Interpret LLM-Generated Reasoning?

- ▶ The paper uses LLM-generated explanations to identify economic mechanisms behind expectation formation.
- ▶ An explanation generated by an LLM is not necessarily the same as the reasoning a comparable human would use.

Key question

How can LLM-generated explanations be validated as informative about human economic reasoning?

Suggestion

Existing papers on narratives or new experiments may help

Additional Comments and Robustness Checks

- ▶ **Model dependence.** Results are robust across GPT-4.1 and GPT-5. Would the main findings survive with a different model family?
- ▶ **A stronger test of date restriction.** Consider historical episodes in which subsequent outcomes were highly surprising given information available at time t . Are date-restricted LLM expectations nevertheless closer to what actually happened? This would test for *latent hindsight bias*, beyond explicit knowledge of future events.
- ▶ **Repeated LLM draws.** How much variation is there across repeated runs with the same personas, prompts, and model? This could separate LLM sampling uncertainty from persona sampling uncertainty.
- ▶ **Sample size.** Why 200 personas given the very low marginal cost? Showing convergence as N increases could help establish how much synthetic sample size matters.
- ▶ **Historical composition.** The same persona distribution is projected backward to 1990. Is the intended object a *constant-composition counterfactual* or a reconstruction of the historical population?

Concluding Thoughts

- ▶ **Promising framework:** LLMs can substantially expand the scale, speed, and scope of survey experiments.
- ▶ **Contribution:** The paper integrates and extends existing approaches to LLM-based survey simulation.
- ▶ **Binding constraint: validation.** Performance in one treatment, outcome, or population need not generalize to others.
- ▶ **Way forward: integrate LLMs with human evidence.** Experiments, surveys, and field data can be used to
 - ▶ validate LLM simulations,
 - ▶ provide more relevant training data
 - ▶ inform the design and calibration of synthetic agents, and
 - ▶ identify the settings in which LLM-based inference is reliable.

LLM surveys can be a potentially powerful complement to human surveys, rather than a substitute for them.