Disclamer: The views expressed in this presentation are those of the authors and do not reflect those of the Deutsche Bundesbank or the Eurosystem.

LLM Survey Framework: Coverage, Reasoning, Dynamics, Identification

Jing Cynthia Wu, Jin Xi, Shihan Xie

Discussion by Olga Goldfayn-Frank

23 March 2026

13th ECB Conference on Forecasting Techniques

Can AI agents replace humans in surveys?

Key methodological innovation:

- **Date-restricted prompting**: Forces LLMs to answer as if in a specific period
- **Internal consistency (fixed personas)**: Same agents across treatments and over time reduces noise

Validation:

- Replicates main results from Weber et al (2025): Similar treatment effects
- High correlation with Michigan Survey of CE

Enables:

- Retrospective coverage
- Dynamic treatment effects and high frequency
- Reasoning extraction

At a fraction of costs of human survey...

Why does it matter?

- **Addresses key limitations of human surveys:**
  - Cannot be run retrospectively
  - Are costly and small(er)-scale
  - Difficult to capture reasoning well
  - Attrition and noise

- **Enables new research designs:**
  - Long historical panels (Du, Monninger, Qui, Wang, 2025)
  - High-frequency dynamics
  - Controlled counterfactuals
  - Direct study of expectation formation and information processing

LLM surveys could become a powerful complement to human surveys!

Comments: External Validity

- **Concern: Do LLMs replicate human reasoning or echoing training data?**
  - LLM "agents" are defined by few observable SoDe characteristics (age, gender, education, mariage, income, location).
  - In literature, these variables explain a very small share of variation in inflation expectations.
  - **where do the LLM agents' prediction actually come from?**

The LLM is mapping: (demographics) → (typical narrative learned from training corpus)

- **In real world:** Most variation is unobserved but economically meaningful (Cognition, attention, information sources, experience etc.)
- **LLM-Models:** Overly structured, may over-rely on "economist-like" narratives, not necessarily meaningful for a given point in time or *geography*

Comments: Time Restriction

- **Concern: Identification relies on prompt compliance ”ignore future information”**
  - LLM agents do not have true time-indexed memory
  - Store compressed knowledge of all periods

- **”Events test” does not guarantee absence of ”knowledge leaking”**
  - Does LLM know and uses findings of replicated or relevant papers?
  - Example: LLM states ”Mean-reversion” as reasoning...

    Armantier et al (2022) ”The curious case of the rise in deflation expectations *“... we find that those who expect deflation are more likely to expect price mean reversion and generally expect better, not worse, economic outcomes”*

- **Suggestions:**
  - Use models trained only on pre-period data
  - Placebo-test: future strongly differs from the past

Comments: Evaluation of the results

- **Trust-building exercise:** Beyond correlation of medians - full distributions?

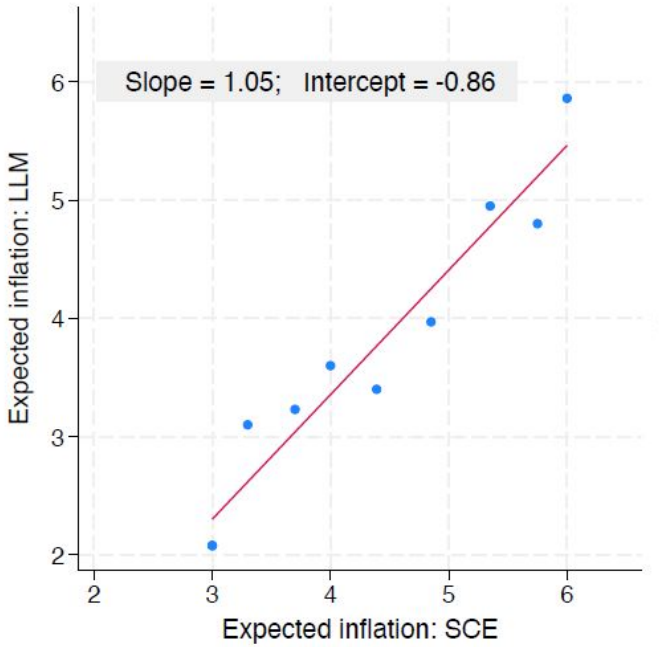
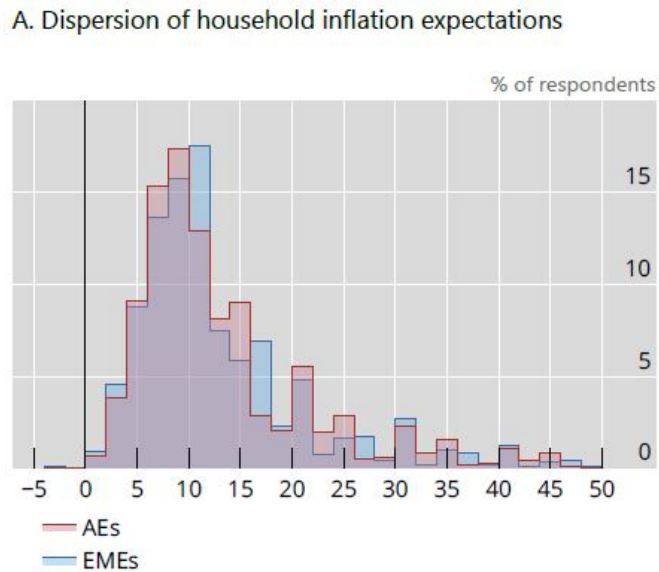


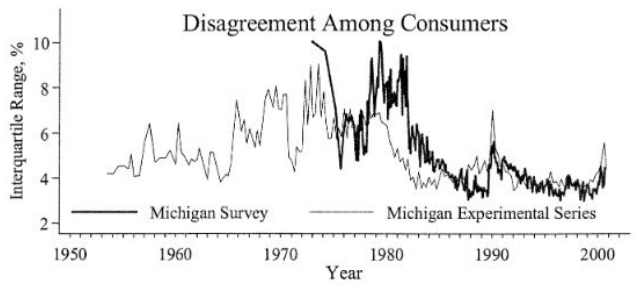
Figure 2: PRIOR INFLATION EXPECTATION: LLM VS. HUMAN



- **Individual uncertainty: LLM vs. Human respondents**?

Heterogeneity in expectations: Noise or valuable insights?

- Paper highlights **lower variance as an advantage**
- **Disagreement is informative** and consequential for expectations formation (Mankiw et al, 2004) transmission of monetary policy (Falk et al 2021), macro dynamics (Pedemonte et al., 2025)
- **Disagreement level *also* changes with inflation**:



- May contain counterintuitive predictions: Link et al 2025: Households and firms with higher attention to macro economy have higher inflaiton expectations

Comments: Story behind the numbers

1. "The weaker human survey-inflation correlations, ... likely reflect greater sampling noise". What about higher?

Table 2: CORRELATION OF SCALED SLOPES WITH INFLATION

Agent type	Correlation with inflation			
	T1: Past inflation	T2: Fed target	T3: Fed forecast	Pooled
LLM	0.92	0.73	0.85	0.79
Human	0.69	0.87	0.13	0.57

2. Humans consistently overpredict inflation vs. LLM. Interaction with high vs. low inflation periods?

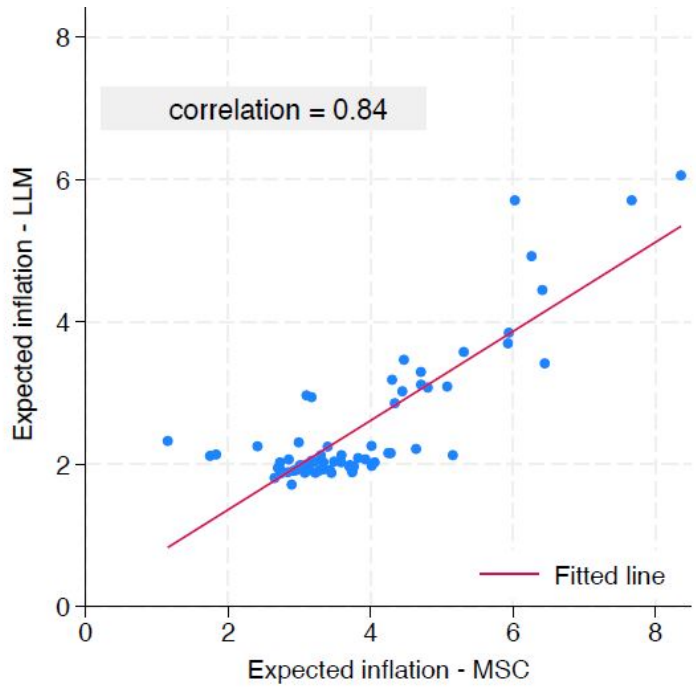


Figure 5: PRIOR INFLATION EXPECTATION: LLM VS. HUMAN (MSC)

Suggestions and open questions

1. Do LLM represent **professional forecasters** better?
2. What about **firms' expectations**? - even more difficult to survey them. But - we also know less about firms
3. Are LLMs-based survey methods useful for countries with less training material?

Very interesting and promising work:

Important, timely discussion and interesting angle!