

Total Recall? Evaluating the Macroeconomic Knowledge of Large Language Models

ECONDAT Knowledge Session, 2025-06

Leland D. Crane, Akhil Karra, Paul E. Soto

Federal Reserve Board of Governors

Opinions expressed herein are those of the author alone and do not necessarily reflect the views of the Federal Reserve System.

Motivation: Two ways economists would like to use LLMs

Lots of interest in using large language models (LLMs) like ChatGPT, Claude, etc.

1. Ask LLM to help with historical analysis: suggest hypothesis, discuss mechanism, etc.
 - Example: To what extent was the early 1990s recession unanticipated, and what variables/models would have forecast it?
2. Use LLMs for forecasting
 - Example: Ask LLM to forecast next 4 quarters of GDP, perhaps using additional data
 - Often need to be validated in-sample (pre knowledge cutoff)

What we do

1. Test LLM recall of important macro data series: LLMs need to know history if they are going make good analyses
2. Test for lookahead bias:
 - Tell LLM to act as real time analyst living on date t . Shouldn't use post- t information.
 - Check if recall errors are correlated with post- t data values or revisions
 - If LLM can't act as a real time forecaster: problems for pre-knowledge cutoff results

Example query (truncated):

- Based on all knowledge available to you, tell me the fully-revised value of `{var}` for `{reference_quarter}`.
- Variables of interest:
 - GDP growth, CPI inflation, industrial production (IP) growth, unemployment rate
- Not giving LLM additional data/ context, see Hansen et al. (2025) for results with context
- LLM responds with a paragraph, extract the estimate with a summarization LLM call.

Example response:

Let's approach this step-by-step:

1. First, we need to understand what we're looking for: the initial release of real seasonally-adjusted GDP growth at an annual rate for the fourth quarter of 1945.
2. It's important to note that GDP as we know it today wasn't introduced until the late 1940s. Before that, GNP (Gross National Product) was the primary measure of economic output.
3. The concept of quarterly GDP growth rates wasn't widely used in the 1940s. Annual figures were more common.
4. 1945 was a pivotal year in economic history due to the end of World War II. The war ended in August 1945, which would have had a significant impact on the economy in Q4.
5. The post-war period typically saw a contraction as the economy transitioned from wartime to peacetime production. This transition often led to a decrease in government spending and a temporary increase in unemployment.
6. However, the release of pent-up consumer demand and the beginning of the post-war boom might have started to offset some of this contraction by Q4.
7. Given the time period, the initial estimates would have been less accurate and more subject to revision than modern figures.
8. Considering the economic conditions of the time, we might expect a negative growth rate, but not an extremely large one, as the immediate post-war contraction might have been starting to bottom out by Q4.
9. Modern data shows that 1946 saw a significant contraction, but this likely hadn't fully hit by Q4 1945.
10. Taking all of this into account, and considering that initial estimates tend to be more moderate than later revisions, we might estimate a figure in the range of -2% to -5% for the initial release.

Given all these considerations, my best estimate for the first print value of real seasonally-adjusted GDP growth at an annual rate for 1945Q4 would be:

-3.5%

This is a point estimate based on the available information and historical context, recognizing that actual data from this period is limited and that economic measurement techniques were less sophisticated than they are today.

Basic Setup

Model

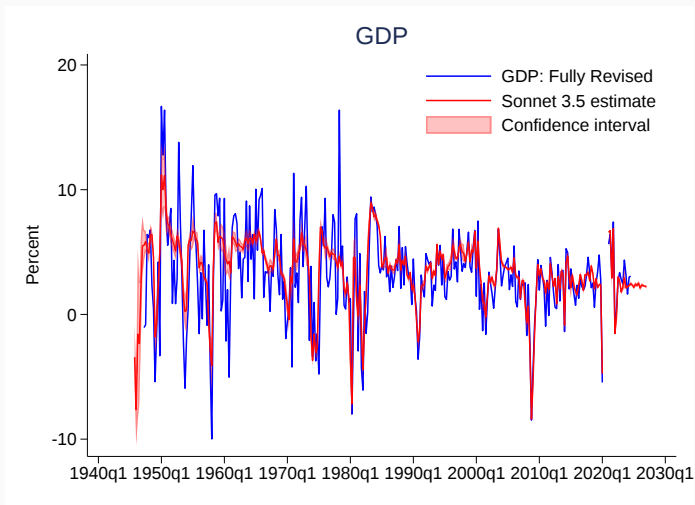
- Use Anthropic's Claude Sonnet 3.5 LLM
- Somewhat outdated now, but still very good

Randomness in responses

- LLM replies are generally stochastic: sample from the next word distribution.
- We repeat each query 10 times and average → more deterministic
- (Also find “temperature” parameter does almost nothing for Sonnet 3.5: Don't assume $\text{temp.}=0$ makes things deterministic!)

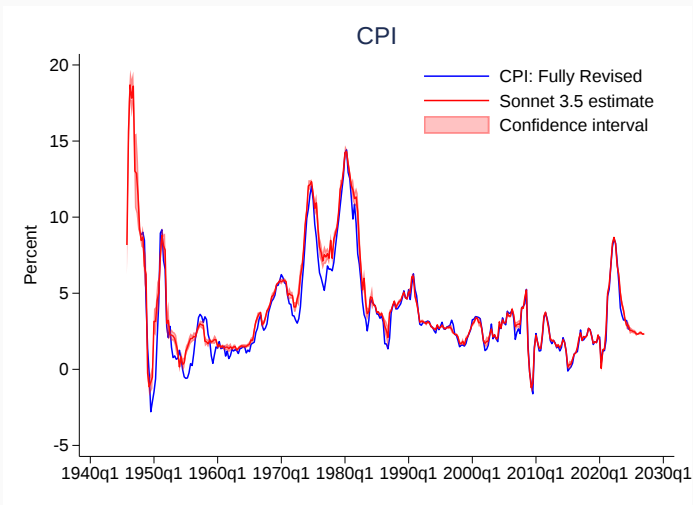
LLM Recall of Macroeconomic Indicators

- Sonnet tracks business cycle pretty well
- Misses much of the quarter-to-quarter volatility
- Worse farther back in time



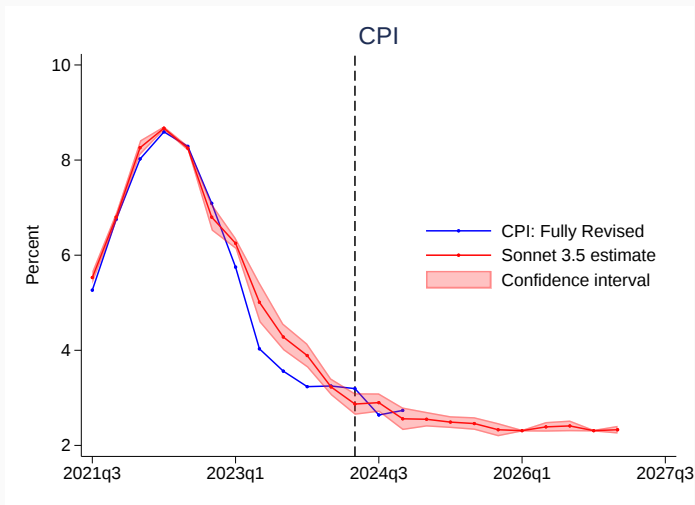
LLM Recall of Macroeconomic Indicators

- CPI looks much better: smoother series
- Pre-1990: starts missing troughs



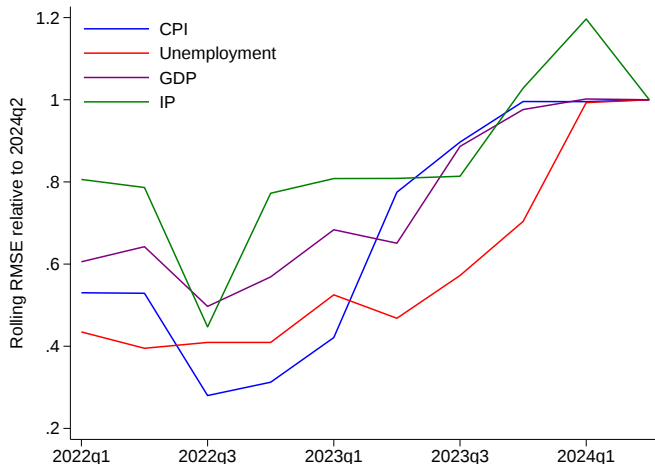
Recent history

- Vertical line: Knowledge cutoff/end of training data (April 2024)
- CPI estimates are somewhat worse as we get close to the knowledge cutoff
- Foresees us getting close to 2 percent inflation



Recent history

- Errors go up as we approach knowledge cutoff
- Small sample!
- But makes sense: training data is likely sparser right before the cutoff
- We haven't written everything about 2023 yet



- Sonnet 3.5 knows a lot about historical time series
- Misses much of the high-frequency stuff
- Worse for more volatile series
- Worse farther back in time and very close to the knowledge cutoff

Want to evaluate real time forecasts: LLM pretends to be a real time forecaster

- *"You are a forecaster living on March 15, 2018. Give me your forecast for 2025Q2 GDP using only information that was available as of that date"*
- Common approach in the literature

Does the LLM obey the real-time restriction? Hard to test

- If LLM forecast is very good: might be using real-time information other forecasters haven't processed

Conjecture: LLM is smoothing together past/future data values

- Focus on recall errors, not forecasts
- Ask LLM to recall period t GDP, using only data through t .
 - Recall error: ε_t
- SFP forecast error for next quarter GDP: ω_{t+1}
 - Unforecastable/hard to forecast using period t data
- Regress ε_t on ω_{t+1}

Correlation b/w ε_t and ω_{t+1} means LLM is using superhuman forecasting ability (ω_{t+1}) to make predictable recall errors for GDP_t

Correlation b/w ε_t and ω_{t+1} means LLM is using superhuman forecasting ability (ω_{t+1}) to make predictable recall errors for GDP_t

- LLM might have more insight into $t + 1$ GDP than SPF: legitimately forecasts some of ω_{t+1}
- But: implausible it uses superhuman insight to predictably misremember historical data
- Much more likely it is averaging future and past values: lookahead bias

What are the capabilities and limitations of LLMs?

- Good understanding of the “broad strokes” of macro data
- Can’t fully follow directions to ignore in-sample training data

What Explains Recall Errors?

$$\varepsilon_t = y_t - \hat{y}_t \quad (1)$$

- ε_t is the recall error: true quarter t GDP minus the LLM's estimate

$$\omega_{t+1} = y_{t+1} - SPF_{t+1}^t \quad (2)$$

- SPF_{t+1}^t : The quarter t Survey of Professional Forecasters forecast of $t + 1$ GDP: 1 quarter ahead
- ω_{t+1} : SFP forecast error: (roughly) the unforecastable part of y_{t+1}

What Explains Recall Errors?

- ε_t is the recall error: true quarter t GDP minus the LLM's estimate
- ω_{t+1} : SFP forecast error: (roughly) the unforecastable part of y_{t+1}

If ε_t is correlated with ω_{t+1} :

- Recall error is correlated with the surprise in future GDP.
- LLM is using future data to estimate current GDP (when told not to)

What Explains Recall Errors?

- ω_{t+1} helps forecast ε_t
- Controlling for information as of t :
GDP, forecast of $t + 1$
- ω_{t+1} noticeably reduces RMSE: economic significance

	1960-2024	1960-1989	1990-2024, ex. 2020
	(1)	(2)	(3)
ω_{t+1}	-0.015 (0.053)	-0.268*** (0.043)	-0.164*** (0.039)
SPF_{t+1}^t	0.092 (0.167)	-0.297*** (0.073)	-0.581*** (0.086)
GDP_t	0.213** (0.094)	0.768*** (0.052)	0.515*** (0.044)
GDP_{t-1}	0.043 (0.061)	-0.041 (0.054)	-0.146*** (0.038)
GDP_{t-2}	-0.018 (0.038)	-0.072* (0.037)	-0.009 (0.017)
Constant	-1.395*** (0.317)	-1.761*** (0.209)	0.204 (0.214)
Adjusted R ²	0.254	0.816	0.648
RMSE	1.692	1.209	0.635
alt RMSE w/o ω_{t+1}	1.682	1.471	0.699

Table 7: Tests for Smoothing. Dependent Variable: ε_t

- Some evidence that LLMs don't fully follow instructions to ignore future data
- Problems for evaluating LLM forecasts before the knowledge cutoff: may be peeking at future data
- May be attenuated if more data is provided: rely less on the fuzzy knowledge of the LLM

Other results

Data release dates:

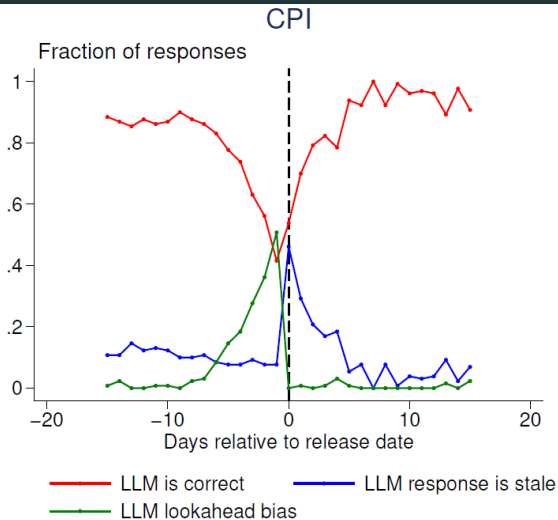
- Do LLMs recall *when* data became available?
- Real-time forecasters needs to know when data become public
- Tell LLM *“It is 5pm on March 15 2018, what is the reference month of the most recent available CPI release?”*

LLMs can recall data release dates to within a couple days

- But mistakes add up: on any given day the LLM will mistakenly think some series has been release
- “Lookbehind bias”: LLMs make opposite mistake too, believing stale data is current.

Release dates

- Accuracy falls around release date
- More lookahead bias just before release, more staleness just after
- Totally consistent with LLM having fuzzy recall



Release date

Inspecting the full text of responses:

- LLM often appears “conservative”: really wants to avoid saying future data was already released
- Despite us only asking for the most recent release.

Alternative prompt:

- Emphasis symmetric loss
- *“It is equally bad to make mistakes in either direction: if you think there is a 51 percent chance the more recent release has occurred, that should be your answer.”*

Release dates

- Lower average accuracy
- Symmetric penalty makes LLM more willing to risk lookahead bias
- More lookahead bias, fewer stale results
- Can trade off types of errors, haven't found a way to reduce total error

