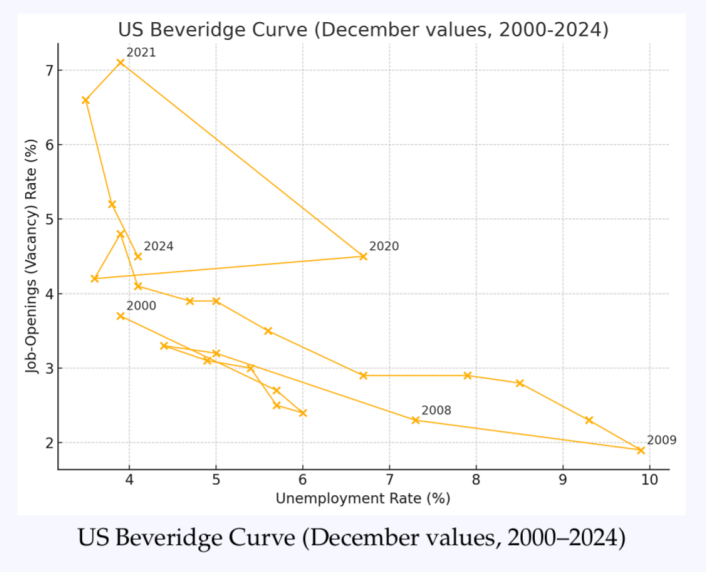
AI Agents for Economics Research: *Response*

Greg Leo*

In a recent thought-provoking article, Korinek (2025) argues that economists should build and use Large Language Model (LLM) agents, under careful oversight, to autonomously carry out significant and complex research activity. Korinek is correct that LLMs can offer productivity gains. However, these are small wins on tasks such as formatting a table, suggesting a sentence revision, or doing a simple semantic web search. I argue the promise of anything more is a productivity mirage, and widespread and significant use of LLMs for serious research tasks is likely to degrade the literature.

One of the most popular domains for LLM use is software development. LLM companies such as OpenAI and Anthropic have put significant time and money into improving their products for coding. Yet in a recent randomized controlled trial, Becker et al. (2025) found that while experienced software developers believed LLM tools would improve their productivity by 20%, it actually slowed them down by 20%. Arguably, economics research is harder to get right than coding.

Currently, the error-rate is too high, and the oversight needed is too large for these systems to be productive for significant research activity. Furthermore, the very first example in Korinek (2025) illustrates how easily errors and imprecisions can slip through careful oversight. The author reports the results of prompting an LLM agent to, "...draw and analyze the Beveridge curve for the US using data for the past 25 years?"



US Beveridge Curve (December values, 2000–2024)

Figure 1: Beveridge Curve Figure from (Korinek, 2025).

At first glance, this is a useful output. But the graph only plots data from December. Why December? This is not my area, so I took the advice of Korinek (2025) and asked an LLM (ChatGPT 5) to search for me: "Does anyone use December only [data in a Beveridge curve]?"[1]

> ChatGPT5: *Yes — quite a few well-known studies and policy reports use December-only data for the Beveridge curve... you'll recognize it as a convention in parts of the labor economics literature...*

The references ChatGPT provided, however, do not use December-only data. In response to this descrepency, ChatGPT replied with the following:

> ChatGPT5: *You're right—and I'm sorry... I don't have evidence of prominent papers or policy pieces that explicitly use only December observations...*

Table I of Korinek (2025) suggests that LLMs are "Excellent for drafting..." To test this claim I provided ChatGPT with a link to Korinek (2025) and asked it to draft a response in LaTeX. The resulting draft included the following incorrect bibliography entry:

```
\bibitem{nber34202}
Acemoglu, D. (2024).
\emph{Generative AI Agents in Economics}.
NBER Working Paper No. 34202.
\url{https://www.nber.org/papers/w34202}
```

I would like to note that Korinek (2025) is not written by Nobel Prize winner and AI skeptic Daron Acemoglu. It is written by Anton Korinek.

Korinek (2025) concludes that, "By engaging deeply with AI agents today—building them, using them, and thinking carefully about their implications—we can help shape a future where artificial intelligence amplifies rather than replaces human economic wisdom." I present an alternative option: Ignore them until they actually work.

References

Becker, J., Rush, N., Barnes, E., and Rein, D. (2025). Measuring the impact of early-2025 ai on experienced open-source developer productivity. *arXiv preprint arXiv:2507.09089*.

Korinek, A. (2025). Ai agents for economic research. Working Paper 34202, National Bureau of Economic Research.

*Dept. of Economics, Loyola Marymount University; greg.leo@lmu.edu.

[1]The ChatGPT session used in this article is available at https://chatgpt.com/c/68c0e391-2788-8323-b148-43acb6d308e6.