No Humans Were Harmed in the Making of This Paper

Andi Qu

Envisioning the Future of Computing Prize
Social and Ethical Responsibilities of Computing
Massachusetts Institute of Technology

Summary

Economic experiments are often constrained by ethical, logistical, and methodological barriers that limit their feasibility. Traditional studies involving human participants can be slow, costly, and difficult to interpret due to subjective variables. Large language models (LLMs) offer a novel approach to economic research by simulating human behavior in controlled environments. These models replicate key cognitive biases, social preferences, and decision-making processes, making them valuable proxies for human participants.

Recent studies demonstrate that LLMs can accurately model economic interactions, such as fairness considerations, status quo bias, and labor-consumption choices. Additionally, LLMs can be conditioned with demographic attributes to simulate behavior across specific population subgroups, enabling highly scalable and controlled experiments. Compared to traditional agent-based models, LLM-driven simulations provide greater autonomy and reduce researcher bias, as their behaviors emerge from learned patterns rather than direct programming. The computational efficiency of LLMs allows large-scale simulations to be conducted rapidly and cost-effectively, offering a powerful alternative to human-based studies.

Applications of LLM-based simulations include policy evaluations in areas such as immigration reform and school-choice mechanisms. By modeling potential outcomes without real-world interventions, these simulations help policymakers design evidence-based strategies. However, concerns about data validity, biases, and potential misuse must be addressed. Solutions such as multi-model validation, prompt engineering, and human oversight can enhance reliability and mitigate risks.

While LLMs should not replace human experiments entirely, they offer a transformative tool for economic research. Integrating LLM-driven simulations with traditional methods can improve data-driven policymaking, enabling more informed and ethical decisions in complex economic environments.

No Humans Were Harmed in the Making of This Paper

Introduction

In an ideal world, all economic policymakers would rely on sophisticated computer models and experimental data to guide their decisions. However, ethical, practical, and methodological barriers often render this data-driven approach infeasible. Consider, for example, an experiment to evaluate the impact of eliminating affirmative action in college admissions. While such data would be invaluable for shaping educational policies – highlighted in several landmark U.S. Supreme Court cases like *Fisher v. University of Texas* and *Students for Fair Admissions v. Harvard* [1] – experimenting ethically would be almost impossible. Intervening this way would likely create long-term, irreversible disadvantages for many individuals, a tradeoff no ethical researcher would accept outside a natural experiment.

Beyond ethical concerns, studies in human development may take years or decades to yield meaningful data, and researchers often need to compensate human participants financially. Moreover, complex social constructs like "diffusion of knowledge" (a core goal of higher education [2]) are difficult to quantify, so researchers often resort to qualitative measures instead. These measures are inherently subjective and prone to variation, even under controlled conditions, making the results challenging to interpret, reproduce, and trust.

Data-driven policymaking is much less common than it should be in today's high-tech age. Instead, world leaders often default to heuristics like common sense and anecdotal evidence. These heuristics may work well in everyday decision-making but are not nearly rigorous enough to address societal-scale issues affecting tens of millions of people. We clearly need new ways to conduct economic and social science experiments that overcome these barriers.

Recent advances in AI, particularly large language models (LLMs), offer a promising solution to this problem. By simulating large-scale economic experiments with LLM "agents" as proxies for human participants, researchers can explore policy outcomes without the ethical or practical constraints of experiments involving humans. These simulations offer economists a cheap, fast, and accurate alternative way to study economic phenomena, helping policymakers shape evidence-based policies to address society's most pressing issues.

The Promise of LLMs as Economic Agents

LLMs are uniquely well-suited for modeling complex economic environments because they combine natural human behaviors with modern computational power. As economic agents, they can accurately emulate how humans behave and interact with each other, including their reasoning biases and cultural beliefs. As computer programs, they give researchers fine-grained control over each simulation's speed, scale, and cost at the click of a button.

There is strong evidence that LLMs learn and make mistakes in ways that mirror humans. For one, LLMs infamously exhibit several reasoning flaws: they "hallucinate" (generate incorrect information that still sounds plausible), forget where they learned certain pieces of information, respond inconsistently after minor wording changes in prompts, and fall for misleading questions. However, Huijzer and Hill (2023) [3] argue that humans frequently make these same mistakes, as shown in numerous behavioral studies. These "undesirable" behaviors – typically seen as fundamental limitations of LLMs – are precisely what align the learning process of LLMs with that of humans. This evidence suggests that LLMs can accurately emulate how humans learn and adapt to new information in complex multi-step, multi-agent environments.

Moreover, LLMs can emulate how individuals from specific human subgroups (millennials, women, etc.) behave. Argyle et al. (2023) demonstrate that by conditioning (prompting) LLMs on sociodemographic backstories of real people, the LLMs would generate nuanced and multifaceted responses closely matching real responses sampled from many different human subgroups. Essentially, they create virtual humans representing those specific subgroups. This demonstrates that the "algorithmic bias" LLMs exhibit is not uniform across the model but demographically correlated, due to the same correlations in their training data. By combining elements from different backstories, the researchers could also accurately model the intersections of those subgroups [4]. This property of LLMs would make it easy for researchers to design a participant pool matching arbitrary target demographics or analyze interactions between specific subgroups – tasks that are much more challenging in experiments involving humans.

Several recent studies have also shown the efficacy of LLM agents as proxies for humans in economic experiments. Horton (2023) successfully replicates several behavioral economics experiments, thereby showing that LLM agents exhibit realistic social preferences like status quo bias (Samuelson and Zeckhauser (1988)) and fairness over profit-seeking (Kahneman et al. (1986), Charness and Rabin (2002)). These social preferences, he argues, emerge as a result of LLMs' latent social information (such as decision-making heuristics and economic laws) from being trained on data generated from real individuals describing their choices [5].

Similarly, Li et al. (2024) simulate experiments on macroeconomic activities and demonstrate that LLM agents can make realistic decisions regarding work and consumption — two key

macroeconomic components. In these simulations, each agent was conditioned with a unique real-world economic profile (such as their wage and whether they chose to work in the previous month), forming a heterogeneous set of interacting agents [6]. This approach overcomes a key challenge in macroeconomic modeling, where the heterogeneity of a population causes emergent macroeconomic phenomena but has traditionally been difficult to model at an individual level. Furthermore, these experiments show that multi-agent simulations over a long period are effective. Li et al. (2024) attribute this success to two additional benefits of LLMs: they behave autonomously and adaptively to their environment without explicit instructions and possess enough intelligence to plan and schedule their actions as a human would [6].

Simulations of LLM agents are not just comparable to experiments involving humans; they are also much more efficient to run. Simulation speeds are only limited by the computers they run on, as there is no need to wait for interactions to play out live. As such, experiments normally spanning months or years could be simulated in just a few hours. Likewise, these simulations have practically no size limit thanks to distributed computing, allowing researchers to add arbitrarily many agents on demand. The only major limiting factor is cost; even so, generating data using LLMs is significantly cheaper than surveying real humans. Argyle et al. (2023) report that one of their experiments, equivalent to surveying over 2000 individuals, cost only $29 using OpenAI's paid GPT-3 API [4]. Thanks to more efficient LLM inference methods and the increasing popularity of open-source LLMs, we can expect future experiments to run even faster and cost even less.

These simulations are also easy to rerun, which has two major advantages. First, it allows researchers to try many different combinations of variables (such as prices and income levels) to test the robustness of their results. Second, verifying results from other experiments that used LLMs would be easy, which would help prevent data fraud and cherry-picking. Given that these issues have been the basis of several recent controversies that have undermined public trust in the social sciences [7], the ability to replicate experiments easily is extremely valuable.

A New Generation of Agent-Based Models

The idea of using an ensemble of autonomous agents to model the economy is not entirely new. Early agent-based models (ABM) emerged in the 1990s as an alternative to empirical statistical approaches to macroeconomic modeling like the Kydland and Prescott model and the DSGE model. These empirical models were often criticized for assuming a perfect world in economic equilibrium, preventing the effective study of major changes or once-in-a-lifetime economic crises. Like empirical models, ABMs output quantitative data allowing researchers to analyze the results using statistical inference techniques. However, they avoid the assumption of a set economic equilibrium, potentially making them a better fit for analyzing these more complex situations.

Despite this potential, ABMs have historically seen limited use in economics research. A common criticism of these models is that they can be self-fulfilling, as the researcher acts as both creator and evaluator, programming the agents and observing their behavior. Horton (2023) likens this to asking, "What would [this model that does what I tell it to do] do?" [5], which is understandably unexciting.

Thankfully, this problem does not apply to LLM agents. Unlike earlier agent-based models, LLM agents are not controlled or programmed directly by the researchers. Although researchers can condition LLMs with beliefs, experiences, and similar factors, their behavior is ultimately a result of their underlying AI models.

This lack of direct control has two additional benefits. First, it minimizes the risk of experimenter bias influencing the results, as these simulations require no direct interaction between researchers and agents. Second, researchers no longer need to program every possible interaction between agents when designing them. Instead, the emergent behaviors observed during the experiment are also a result of the underlying AI model.

Some Motivating Use Cases

In addition to the affirmative action example outlined in the introduction, below are two more areas in which LLM-based simulations could help shape public policy.

Immigration and Visa Policies

Immigration reform has been a contentious topic in the last several U.S. elections. Policies like the number of H-1B visas available and the number of refugees to accept have far-reaching effects on the economy, such as employment rates, average worker wages, university enrollment, cultural diffusion, and birth rates.

Although there are many arguments for and against immigration in the economic literature, the actual effects of immigration policies (and changes to them) are still not well understood. Real-life experiments to determine their impacts are infeasible for several reasons. First, they would be unethical – given the far-reaching and long-lasting effects of immigration, making any major changes to the system (for example, halving the number of H-1B visas) would be incredibly dangerous without knowing the likely outcomes. Second, they would face many legal barriers without compelling evidence supporting them – immigration is a highly sensitive topic, so any major changes would be challenged in court by opposing lawmakers and corporations alike.

LLM-based simulations would help economists model the complex dynamics of immigration policy changes without subjecting real people to those changes. Thus, they would help

lawmakers create policies that best serve their country in a more objective and non-partisan manner.

School-Choice Mechanisms

School-choice mechanisms are algorithms that public school systems use to match students with schools. These algorithms typically consider each student's preference ranking of schools and some priority factors, such as where their siblings attend school.

Modern school-choice mechanisms work well and assign most students to one of their top-choice schools. However, they rely on the assumption that everyone reports their preferences truthfully. Many parents may try to rank schools strategically, often collaborating with other parents and sometimes in unexpected and irrational ways that still affect the outcomes. This generally happens because the mechanisms are hard to explain and poorly understood by parents.

Economists currently design and analyze school-choice mechanisms under a mathematical, game-theoretic lens. However, this approach cannot capture the social dynamics and decision-making heuristics of parents, nor can it model how well they would interpret a new mechanism. Deploying a new school-choice mechanism without understanding these factors could be extremely controversial and undermine trust in the public school system.

LLM-based simulations would help economists model how reported preferences compare against true preferences and find an effective way to communicate how the school-choice mechanism works. Thus, they would help ensure that any new mechanism economists deploy maximizes efficiency while minimizing confusion and controversy.

Potential Pitfalls and Mitigations

Given how little we currently understand about LLMs, quantitative-minded researchers would understandably be skeptical of this unconventional approach to experimentation or even reject it outright. Thus, a primary concern regarding this approach is data validity – how can we identify (and ideally prevent) rogue LLMs that generate garbage data in our experiments?

Indeed, LLMs are notoriously opaque. Although their responses are human-like, they are still fundamentally unpredictable because, like most deep-learning models and the human brain, they are inscrutable "black boxes". Moreover, the training data for these models are typically not made open-source, so any given LLM might be secretly trained on "poisoned" data, either intentionally (a malicious actor plants the data) or not (the data are carelessly scraped from the internet).

This unpredictability opens the doors for many undesirable behaviors during simulations. First, numerical data generated by LLM agents may come from flawed numerical reasoning, rendering any statistical analysis of the results meaningless. Second, data poisoning could change how LLM agents behave in certain situations (for example, uniformly preferring some political ideology) and sway experiment outcomes. As such, it would not be entirely unreasonable to expect someone seeking to influence major policy decisions (a foreign adversary, an unethical corporation, etc.) to attempt a data poisoning attack.

A simple yet effective way to mitigate these transparency-related issues is to rerun simulations with several different LLMs multiple times, just as one would collect many data points in a physics experiment. This way, simulations affected by rogue LLMs would emerge as outliers. Conversely, if all simulations yield similar results, then those results are very likely valid.

Clever prompt engineering is another effective way to make LLMs more reliable. For example, prompting agents to list their step-by-step reasoning (known as "chain-of-thought" prompting) empirically improves LLMs' reasoning abilities. Furthermore, "anchoring" agents to some reference helps them generate numerical responses on a common scale, making statistical analysis on those responses more reasonable [8].

Combined with chain-of-thought prompting, having a human in the loop monitoring each interaction between agents would help researchers identify rogue LLM behavior and intervene if necessary. Individual interactions should be simple and unsurprising, so this task should be relatively easy for the human monitor. Although this mitigation seems more labor-intensive than simply rerunning the simulations, it would not require extra effort in practice, as researchers are already expected to inspect their data manually.

Another valid concern regarding these simulations is that they might cause the social sciences to rely too heavily on LLM-based simulations, rendering experiments involving humans obsolete. This outcome is undesirable because the simulations' results may occasionally be inaccurate, and we may not notice. However, this outcome is also unrealistic, seeing how simulations are so prevalent in the physical sciences, yet experimental data remain as relevant as ever. For example, fluid dynamics simulations are used to design aerodynamic structures; even though these simulations are extremely accurate, scientists and engineers still opt to test those structures in physical wind tunnels whenever possible. Likewise, economic researchers will likely always be motivated to reproduce simulation results with real humans whenever possible.

A broader pitfall of LLMs is that they can be used to spread misinformation or commit fraud because they can convincingly imitate real humans. Although today's LLMs are already being used for these nefarious purposes, popularizing LLM-based simulations in research may worsen this problem. Researchers would likely be incentivized to develop more human-like LLMs,

which, despite being more capable of modeling human behavior and the economy, would also be more effective at causing harm.

Preventing these dangers would require a coordinated effort across the scientific community. First, we must establish strict ethical guidelines prioritizing transparency and accountability for researchers and software engineers working on LLMs, such as mandating that training data be made open-source. Second, we need more robust content verification systems, such as adding covert "watermarks" to LLMs' outputs that are not noticeable by humans but are still cryptographically detectable. Third, we could add technical safeguards to the LLMs themselves, such as built-in constraints preventing them from generating highly manipulative or deceptive content.

Finally, we cannot ignore LLMs' environmental impact. LLM-powered applications like ChatGPT consume over a million kilowatt-hours of energy daily (equivalent to 33,000 U.S. households) [9], which results in enormous carbon emissions and contributes heavily toward climate change. Although economic research happens at much smaller scales than ChatGPT, simulations with thousands of LLMs would still result in significant carbon emissions if done naively.

To reduce the environmental impact of LLM-based simulations, researchers could use smaller, less powerful models with a fraction of the parameters (and consequently power consumption) of state-of-the-art models like OpenAI o1. These smaller models may not have the best reasoning capabilities, but that is not a problem – we only need them to emulate regular humans, not superhumans.

There is also a lot of active research into making LLMs more computationally efficient. Recent advances, such as parameter quantization and FlashAttention [10, 11], have made LLMs efficient enough to run entirely on a consumer laptop without sacrificing quality. These techniques dramatically reduce LLMs' power consumption and, consequently, their climate impact.

Conclusion

LLMs are ideal proxies for human participants in economic research because they enable researchers to model humans – their biases, social preferences, and interactions – across a wide range of environments with unprecedented efficiency and accuracy. As such, they have the potential to shape data-driven policies that best address society's most challenging issues.

Of course, this does not mean that LLMs should entirely replace humans in all experiments. I only advocate for using LLMs in experiments that would be infeasible to conduct with humans or preliminary trials to probe for new research directions. After all, humans are ultimately our most reliable source of human behavior. Still, I encourage researchers to run LLM-based

simulations and compare their results to corresponding experiments involving humans, as such comparisons would help us better understand how LLMs align with humans.

Although there are some risks involved with this technology, several effective mitigations against them exist. Above all, we must focus on transparency and accountability – to understand the inner workings of LLMs, make them more reliable, and prevent their use for nefarious purposes.

Creating an accurate AI-based model of the entire economy would be no small feat. It would require significant investment in computing resources and the cross-disciplinary collaboration between economists and computational scientists with experience in large-scale simulations. Given the enormity of the potential benefits, though, it is a feat well worth undertaking.

References

[1] Supreme Court of the United States. *Students for Fair Admissions, Inc. V. President and Fellows of Harvard College*. Supreme Court of the United States, 2023.

[2] Coleman, and E James. *Diffusion of Knowledge: The Ignored Goal of Public Education*. Vol. 8, no. 1, 1 Jan. 2016, pp. 45–58.

[3] Huijzer, Rik, and Yannick Hill. *Large Language Models Show Human Behavior.* 31 Jan. 2023, https://doi.org/10.31234/osf.io/munc9.

[4] Argyle, Lisa P., et al. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis*, vol. 31, no. 3, 1 July 2023, pp. 337–351, https://doi.org/10.1017/pan.2023.2.

[5] Horton, John. *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* 1 Apr. 2023, https://doi.org/10.3386/w31122.

[6] Li, Nian, et al. "EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities." *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, vol. 1, 1 Jan. 2024, pp. 15523–15536, aclanthology.org/2024.acl-long.829/, https://doi.org/10.18653/v1/2024.acl-long.829.

[7] Leif, Uri, Joe, &. "[109] Data Falsificada (Part 1): "Clusterfake."" *Data Colada*, 17 June 2023, datacolada.org/109.

[8] O'Hagan, Sean, and Aaron Schein. "Measurement in the Age of LLMs: An Application to Ideological Scaling." *ArXiv.org*, 2023, arxiv.org/abs/2312.09203.

[9] McQuate, Sarah. "Q&A: UW Researcher Discusses Just How Much Energy ChatGPT Uses." *UW News*, 27 July 2023, www.washington.edu/news/2023/07/27/how-much-energy-does-chatgpt-use/.

[10] Ji, Lin, et al. "AWQ: Activation-Aware Weight Quantization for LLM Compression and Acceleration." ArXiv.org, 1 June 2023, https://doi.org/10.48550/arxiv.2306.00978.

[11] Dao, Tri, et al. "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness." ArXiv.org, 23 June 2022, arxiv.org/abs/2205.14135.