

On The Limits of Agency in Agent-based Models

Ayush Chopra
Massachusetts Institute of Technology
Cambridge, MA, USA
ayushc@mit.edu

Shashank Kumar
Massachusetts Institute of Technology
Cambridge, MA, USA
shashank3256@gmail.com

Nurullah Giray Kuru
Massachusetts Institute of Technology
Cambridge, MA, USA
ngkuru@mit.edu

Ramesh Raskar
Massachusetts Institute of Technology
Cambridge, MA, USA
raskar@mit.edu

Arnau Quera-bofarull
University of Oxford
Oxford, UK
arnauq@proton.me

ABSTRACT

Agent-based modeling (ABM) offers powerful insights into complex systems, but its practical utility has been limited by computational constraints and simplistic agent behaviors, especially when simulating large populations. Recent advancements in large language models (LLMs) could enhance ABMs with adaptive agents, but their integration into large-scale simulations remains challenging. This work introduces a novel methodology that bridges this gap by efficiently integrating LLMs into ABMs, enabling the simulation of millions of adaptive agents. We present LLM archetypes, a technique that balances behavioral complexity with computational efficiency, allowing for nuanced agent behavior in large-scale simulations. Our analysis explores the crucial trade-off between simulation scale and individual agent expressiveness, comparing different agent architectures ranging from simple heuristic-based agents to fully adaptive LLM-powered agents. We demonstrate the real-world applicability of our approach through a case study of the COVID-19 pandemic, simulating 8.4 million agents representing New York City and capturing the intricate interplay between health behaviors and economic outcomes. Our method significantly enhances ABM capabilities for predictive and counterfactual analyses, addressing limitations of historical data in policy design. By implementing these advances in an open-source framework, we facilitate the adoption of LLM archetypes across diverse ABM applications. Our results show that LLM archetypes can markedly improve the realism and utility of large-scale ABMs while maintaining computational feasibility, opening new avenues for modeling complex societal challenges and informing data-driven policy decisions.

KEYWORDS

Million-scale Agent-based Models; Differentiable ABMs; LLM as ABM agents; Generative agents

ACM Reference Format:

Ayush Chopra, Shashank Kumar, Nurullah Giray Kuru, Ramesh Raskar, and Arnau Quera-bofarull. 2025. On The Limits of Agency in Agent-based Models. In *Proc. of the 24th International Conference on Autonomous Agents*

and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025, IFAAMAS, 10 pages.

1 INTRODUCTION

Many of the today’s challenges — from epidemics to housing shortages to humanitarian crises — emerge from the complex interplay of countless individuals making decisions and interacting over time. Agent-based models (ABMs) aim to capture these dynamics by simulating collections of agents that act and interact within digital worlds. ABMs have proven useful across various domains, including epidemiology [7, 21, 23], economics [5, 6], and disaster response [8, 18]. For instance, they were used to evaluate vaccination protocols during the COVID-19 pandemic [7, 21, 43], predict the crash of housing markets [6, 17], and design evacuation programs for war refugees [18, 29]. Their utility in addressing policy questions stems from the ability to simulate interplay between individual behaviors and environmental dynamics.

However, the practical application of ABMs has been hindered by two major challenges. First, the high computational costs associated with simulating and calibrating large-scale models have limited their widespread adoption [9]. Recent advancements in deep learning have addressed some of these challenges, enabling the simulation of complex dynamics over millions of agents using vectorized operations [11, 42] and the calibration of models to heterogeneous data sources using differentiable programming [2, 11, 14, 30, 41, 46]. In such differentiable ABMs, deep neural networks also help specify complex environment dynamics [31], and autograd facilitates sensitivity analysis in zero-shot [39]. Hence, it is now feasible to simulate, calibrate, and analyze ABMs with millions of agents using commodity hardware. Second, and perhaps more critical, is the lack of expressiveness in ABM agents. Many ABMs rely on simplistic, rule-based agent behaviors that fail to capture the nuanced, adaptive decision-making of real-world individuals.

Large language models (LLMs) have shown remarkable performance in text-based applications [3, 35, 47], suggesting a possible solution to the challenge of creating more realistic agent behaviors. LLM-powered agents have demonstrated potential to enable more general and adaptive human-like behavior [22, 45]. However, integrating LLMs into large-scale ABMs remains problematic. While promising work on multi-agent simulations with LLM-agents has been performed [36, 48], it has primarily been limited to tabletop games and small population scenarios (few hundred agents). Querying an LLM for each individual agent’s decision at every time step quickly becomes computationally infeasible as the number of



This work is licensed under a Creative Commons Attribution International 4.0 License.

agents grows into the millions, which is often necessary for studying large-scale complex systems like epidemics or supply chain networks. The goal of our research is to bridge this gap.

Contribution: This work introduces LLM archetypes, a novel methodology that efficiently integrates LLMs into ABMs while maintaining the ability to simulate millions of agents. Our key insight is that by querying LLMs for representative agent types rather than individual agents, we can achieve a balance between behavioral complexity and computational efficiency. LLM archetypes identify representative agent types and use LLM queries to inform the behavior of entire groups of similar agents. Importantly, our approach does not lead to a degenerate solution where all agents within a group make identical decisions. Archetypes maintain action heterogeneity within each group through probabilistic sampling, while significantly reducing the computational burden.

Our approach demonstrates two crucial advantages of LLM archetypes: computational feasibility and enhanced performance. First, we show that LLM archetypes can reproduce population-wide individual behaviors with significantly fewer queries compared to fully LLM-powered agents. This dramatic reduction in computational overhead enables the simulation of millions of agents, which is often necessary for studying large-scale complex systems. Second, LLM archetypes achieve better calibration, and enable flexible counterfactual analysis by preserving simulation scale. We highlight the nuanced trade-off between individual agency and simulation scale, demonstrating that archetypes outperform both fully-adaptive LLM agents (limited to smaller populations due to computational constraints) and simple heuristic agents (which lack adaptivity).

To validate the concept of archetypes, we present a case study of the COVID-19 pandemic in New York City, simulating 8.4 million interacting agents. This case study showcases how LLM archetypes offer a balanced solution that preserves both adaptive behaviors and computational efficiency. We merge our contributions to the AgentTorch framework [12] to enable utility across diverse ABMs.

2 BACKGROUND

We define the simulation, calibration and analysis of an ABM. We formalize the notion of an ABM agent and highlight challenges in scaling LLM-based agent simulations to large populations.

2.1 Agent-based Modeling

Consider an ABM composed of N agents. We denote by $s_i(t)$ the state of agent i at simulation time t , which contains both static and time-evolving properties of agents. For instance, s may represent the age and sex of a person and their infected status. As the simulation proceeds, an agent i updates their state $s_i(t)$ by interacting with their neighbours $\mathcal{N}_i(t)$ and their environment $e(t)$, which can both also be time evolving. The neighbourhood of an agent can be specified using a graph, a proximity metric, or other methods. We denote by $m_{ij}(t)$ the message or information that agent i obtains from their interaction with agent j . In the case of an epidemiological ABM, this may represent the transmission of infection from agent j to agent i . We can then write the agent’s update rule as

$$s_i(t+1) = f\left(s_i(t), \cup_{j \in \mathcal{N}(i)} m_{ij}(t), e(t), \theta\right), \quad (1)$$

where θ are the structural parameters of the ABM. For instance, θ may correspond to the infectivity of a virus, or the vaccination efficacy. Similarly, the environment can also have its own dynamics that depend on the agent’s updates and actions,

$$e(t+1) = g(s(t), e(t), \theta). \quad (2)$$

The specific choices of f and g define the dynamics of the ABM system and they are typically stochastic functions which can be mechanistically specified or learned from data.

Simulating an ABM consists of picking an initial condition for the agents and environment states ($s(0), e(0)$) and recursively applying Equation 1 and Equation 2. Despite the very large size of the simulated state space, we are mainly interested in a collection of aggregate outcomes over agent states. For most ABMs, this corresponds to a multivariate time-series $x_t = h(s(t))$. For example, in epidemiological ABMs h corresponds to summing over the infected agents to obtain the daily number of infected agents. Once the functional form of an ABM has been set, the simulation of an ABM can be conceived as a stochastic simulator,

$$x = F(\theta, s(0), e(0)), \quad (3)$$

where $F = (f, g) \circ \dots \circ (f, g)$. The composition is repeated for T time-steps.

Calibrating an ABM refers to the process of finding a set of structural parameters $\hat{\theta}$, or a probability distribution over θ , such that $F(\hat{\theta}, s(0), e(0))$ produces an output x that is consistent with real-world data y . There are various techniques for calibrating ABMs, including approximate Bayesian computation [37] and neural likelihood and posterior estimation [14], among others.

Once calibrated, we can execute sensitivity **analyses on ABMs** to understand past events (retrospective), explore alternative scenarios (counterfactual), and design future policies (prospective) [40, 43]. This analytical capability makes ABMs powerful tools for policy design and positions them to address the Lucas critique [26].

2.2 Scaling Agent-based Models

Recent advancements, particularly in differentiable ABMs, have made it feasible to simulate, calibrate, and analyze ABMs with millions of agents using commodity hardware. A **differentiable ABM** [4, 11, 41] is an ABM for which the gradient

$$\eta = \nabla_{\theta} \mathbb{E}[F(\theta)] \quad (4)$$

exists and can be computed. This allows ABMs to improve calibration by using gradient-assisted techniques to integrate heterogeneous data [15, 41], accelerate simulations on CPUs and GPUs via tensorization [10], compose with neural networks in end-to-end differentiable pipelines [11, 32, 33] and accelerate sensitivity analyses with gradients [40].

AgentTorch [12] is an open-source framework that allows to generalize these capabilities across diverse ABMs. Its key feature is the ability to differentiate through the simulation (f, g) , enabling gradient-based optimization of model parameters θ . Through smoothing and reparameterization techniques, AgentTorch achieves differentiability in discrete stochastic programs and allows for the simulation of tens of millions of agents on consumer-grade GPUs. However, these large-scale simulations have only focused on heuristic. The focus on our work is to preserve this simulation scale while

incorporating LLMs to capture adaptive agent behavior. We build on top of AgentTorch for our experiments and analysis. Its flexibility in composing various agent rules and environments makes it a suitable framework for benchmarking agents defined by both heuristic and LLM-based behaviors.

2.3 Agency in Agent-based Models

Conventional ABMs typically use heuristic agents update rules f (Equation 1) derived from observational data or grounded in theory. However, these rules may not explicitly differentiate between components that depend on agent behaviour and environmental dynamics. An illustrative example of this is the dependence of disease risk on the basic reproduction number R_0 . R_0 corresponds to the expected number of cases directly generated by one infected individual. This parameter definition, however, does not allow to distinguish whether a high number of cases is driven by the agent’s behaviour (i.e., they interact more), or an increase in the infectivity of the virus (i.e., each contact is more infectious).

Recent advances in Large Language Models (LLMs) have opened new possibilities for creating more realistic and adaptive agent behaviors in ABMs. Integrating LLMs into ABMs can help decouple agent behavior dynamics from environmental dynamics. This modification to the agent update rule (Equation 1) is expressed as:

$$s_i(t + 1) = f \left(s_i(t), \cup_{j \in \mathcal{N}(i)} m_{ij}(t), e(t), \theta, \ell(\cdot | s_i(t), e(t), \theta) \right), \quad (5)$$

where $\ell(\cdot | s_i(t), e(t), \theta)$ is the LLM output. For example, when modeling the infection probability of an agent, an LLM could parameterize behaviour related mask-wearing compliance. To interpret the LLM output as an action within the ABM environment, we instruct the ABM to return yes / no answers to our prompts. In other words, given an action α (i.e., will the agent isolate at home?) with unknown probability p , we use the LLM as a proxy,

$$\alpha \sim \text{Bernoulli}(p) = \ell(\cdot | s_i(t), e(t), \theta). \quad (6)$$

Several recent works has explored integrating LLMs as ABM agents. Notably, "Smallville" [36] simulates 25 LLM-powered agents coordinating to plan a party together, [19] simulates disease spread over 100 LLM-powered agents, [25] build an expressive environment of macroeconomic dynamics but simulate only 300 agents, [49] simulates 1000 LLM-powered agents interacting in minecraft with the aim to capture self-organization in societies. [48] built an ABM framework where both agents and environment are modeled using LLMs. While promising, these works have been limited to small population scenarios (few hundred agents) and not designed to integrate real-world population data [20]. Integrating LLMs into large-scale ABMs remains challenging but is necessary for evaluating population-scale complex systems and guiding real-world policy decisions. The focus of our work is to preserve the simulation scale while incorporating LLMs to capture adaptive agent behavior, addressing the critical trade-off between individual agent expressiveness and computational feasibility in large-scale simulations.

3 LLM ARCHETYPES: SCALING ABM AGENTS

Understanding complex systems often requires the simulation of the entire population of agents to correctly capture emergent scale-sensitive effects. For instance, while the agency or intelligence of

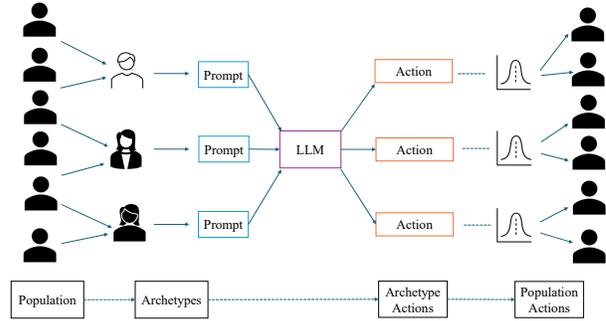


Figure 1: Schematic for sampling individual agent behavior using LLM archetypes. The process involves: (1) assigning individuals to representative archetypes (based on prompt template), (2) querying LLMs for archetype behaviors and estimating action distributions, and (3) sampling individual agent decisions from action distribution of representative archetype. This approach enables efficient scaling of adaptive behaviors to large agent populations.

an individual ant may be quite limited, the simulation of the entire colony captures coordination processes wherein ants use themselves as bridges for other ants to use. In these large systems, however, it is infeasible to query Equation 6 for each agent, time-step, and specific action. This problem can be overcome by recognizing that the number of different behaviors is typically much smaller than the number of agents. In other words, we only need to query the LLM to inform the behaviour over each unique set of agents’ characteristics. For instance, if we consider that the behavior is purely informed by age and gender, we only need to consider one LLM query per different combination of age and sex. We refer to each of these unique combinations as **archetypes**.

For each possible agent action α , we can estimate its probability p_α in Equation 6 using Monte-Carlo,

$$p_\alpha(s_i(t), e(t), \theta) = \mathbb{E}[\ell(\cdot | s_i(t), e(t), \theta)] \approx \frac{1}{M} \sum_{j=1}^M \xi_j \text{ with } \xi_j \sim \ell(\cdot | s_i(t), e(t), \theta). \quad (7)$$

By estimating $p_\alpha(k)$ for each archetype k , we can simulate the action of its agent by sampling the action from the archetype to which it belongs. Let K be the number of agent archetypes and A be the number of LLM-queryable actions; we can then simulate the behaviour of all agents with $K \times A$ queries, which will be typically much smaller than the number of agents N , allowing us to scale the simulation to millions of agents. This is shown in Figure 1.

4 EXPERIMENTAL SETUP

We consider a large-scale ABM of New York City simulating 8.4 million individuals during COVID-19. The pandemic exemplifies the intricate interplay between individual behavior, policy interventions, and environmental factors that our approach aims to capture. Disease spread initially triggered fluctuating mobility patterns and multiple infection waves [44]. Government lockdowns, while controlling spread, caused severe economic consequences, including unprecedented unemployment [13]. Stimulus programs, introduced to mitigate economic hardship and encourage compliance to health measure [24], had unintended effects on labor markets and resource allocation [16]. As the pandemic progressed, "pandemic fatigue" emerged, further complicating public health compliance and economic recovery [38]. This feedback loop between health outcomes, economic conditions, and human behavior is ideal to study the trade-off between simulation scale and individual agency.

Environment: The agent states have static (age, gender, income, occupation) and dynamic (disease, employment status) attributes. We use 2022 American community survey (ACS) for demographic and household characteristics, the Bureau of Labor Statistics for employment data, and the center for disease control (CDC) reports for data on disease dynamics, consistent with prior work [10, 21, 43]. Agent attributes are specified at census-resolution with demographic and income information discretized into bins. Interactions occur over household, workplace, and mobility networks, with recreational and workplace mobility parameterized using Google Mobility trends. Our simulations focus on the dynamics of disease spread and labor market. For disease spread, we consider a standard epidemiological model [7, 11, 21] wherein infection spreads through contact and the probability of agent i getting infected at step t is:

$$p_i(t) = 1 - \exp\left(-\frac{\beta S_i}{n_i} \sum_{j \in \mathcal{N}(i)} I_j(t)\right), \quad (8)$$

where $\mathcal{N}(i)$ is the set of neighbors of agent i , S_i the susceptibility of agent i , I_j the infection status of each neighbour, $n_i = \#\mathcal{N}(i)$ the total number of neighbors, and β a structural parameter of the ABM called the effective contact rate. The neighbourhood $\mathcal{N}(i)$ is given by contact networks constructed via household and mobility data in the US census.

For labor market, we consider a standard econometric model [25] which relates participation behavior of individual agents with aggregate unemployment rate at time t ($\mu_{w,t}$) as:

$$\mu_{w,t} = \sum_{j \in \mathcal{N}} \gamma_0 W_j(t) + \gamma_1 C_t \quad (9)$$

where $W_j(t)$ is the willingness to work for agent j at time t and C_t is the history of unemployment claim rates in the region, obtained from census data; and γ_0 and γ_1 are the structural parameters.

The epidemiological model forecast cases, while the economic model forecast unemployment rates. The models are coupled through a feedback loop: case numbers affect agents' willingness to work, which influences labor-force participation rates and workplace interaction networks, which in turn affect disease transmission. We implement this environment using the AgentTorch framework [12] which enables differentiate through these stochastic dynamics and

scales to large populations (8.4 million agents). The parameters ($\beta, \gamma_0, \gamma_1$) are calibrated to real-world data for cases and unemployment rates, using a standard protocol for differentiable ABMs [11, 12, 41] (visualized in figure 3).

Behavior: We use LLMs to model isolation and employment behavior of individual agents. Our prompt includes agent demographics, disease dynamics, information about extrinsic interventions (stimulus payment) and intrinsic behavior adaptation (duration of pandemic to capture effect of "pandemic fatigue"). The user prompt template, motivated by [25], is given below:

User Prompt

You are a {gender} of age {age}, living in the {location} region and receiving a monthly income of {income}.

The number of new cases in your neighborhood is {cases}, which is a {change}% change from the previous month. It has been {duration} months since the start of the pandemic.

This month, you have received a stimulus payment of {payment} to support your living expenses.

Given these factors, do you choose to isolate at home? (isolation behavior)

Given these factors, do you choose to work? (employment behavior)

"There isn't enough information" and "It is unclear" are not acceptable answers. Give a "Yes" or "No" answer, followed by a period. Give one sentence explaining your choice.

The input prompt receives case numbers and pandemic duration from the past simulation trajectory, instead of ground-truth data, and outputs agents' willingness to work $W_j(t)$ which is further used in simulation. We conduct such auto-regressive prompting for two reasons: i) when simulating for prospective interventions, ground-truth data is not available and hence prompt needs to be specified entirely using simulation, ii) when simulation is un-calibrated, the model peaks may not align well with real-world data. In such case, using ground-truth data is unsuitable for capturing the adaptability of behavior (especially when incorporating time-varying information like infections). Since prompt at step t depends upon simulated trajectory at step $t - 1$, LLMs need to be sampled online during the simulation. As behavior cannot be sampled offline, the trade-off between simulation scale and agent behavior becomes particularly critical. Finally, the stimulus eligibility, timing and amounts are based on the policy implemented in NYC at the time. Specifically, December 2020 - March 2021 overlaps with the second stimulus check which provided adults \$ 600 and additional \$ 600 for every child [34]. We use GPT-3.5 for our experiments.

5 VALIDATING ARCHETYPES

The section benchmarks the capacity of Archetypes to prompt behavior consistent with measurable population-wide observations. We use LLM archetypes to prompt individual-level willingness to work for 8.4 million synthetic agents (consider responses only for eligible working adults), focusing on two time-periods: December 2020 to March 2021 (coinciding with the second stimulus round) and

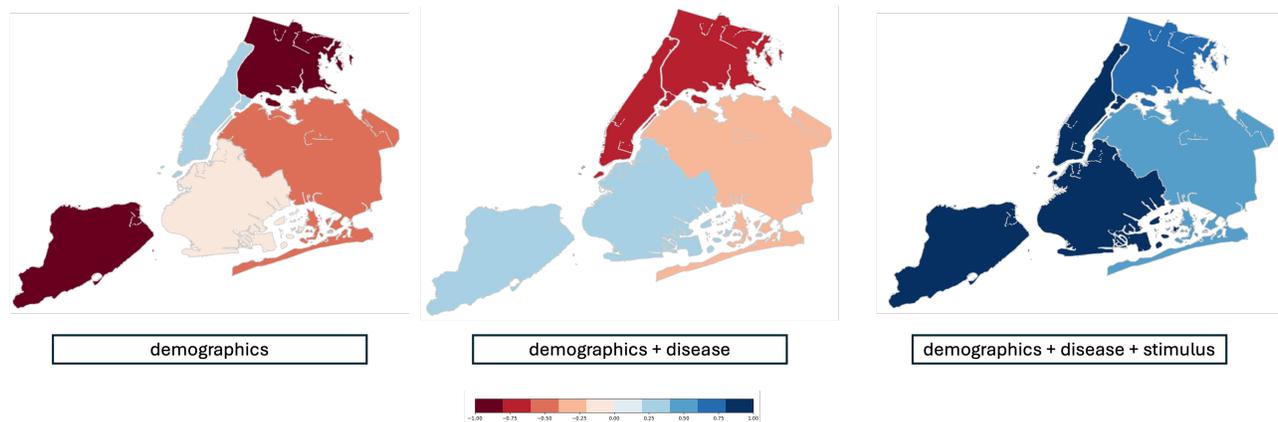


Figure 2: Prompting agents via LLM archetypes: Correlation between population-wide employment behavior predicted by LLM archetypes and observed data for 8.4 million NYC agents. Prompt 1 (left) corresponds to scenario where LLMs only see demographic attributes. Prompt 2 (middle) and Prompt 3 (right) add further contextual information regarding disease dynamics and stimulus payments. Increased correlation with additional contextual information highlights the ability of LLMs to capture behaviour trends across demographics and geography.

March to May 2022 (post-Omicron wave). We test three scenarios with increasing contextual information: (Prompt 1) we only provide demographic attributes of the agent, (Prompt 2) we add information about disease dynamics, (Prompt 3) we further include information about access to stimulus payments. For each scenarios, we initialize 3 (M in equation 7) queries per archetypes representing different combinations of unique prompts over the considered time frame. Given the census-resolution (demographic and income attributes binned for privacy) and prompt design (disease dynamics and stimulus information shared by all agents), this approach requires only 400 LLM queries to sample one decision (weekly) for each individual in the population. This is a significant reduction compared to millions of queries required in the conventional paradigm.

Following Equation 7, we obtain the probability of each archetype k performing action α ("will you work?") for each week, $p_{\alpha}(k, t)$. By sampling from the induced Bernoulli distributions, we generate a time-series of work attendance. We aggregate these individual-level actions to calculate the change in labor force participation rate across New York boroughs and compute correlation of this time series with observed data from the US Bureau of Labor Statistics¹. Results are averaged across 5 independent runs for robustness.

Figure 2 shows increasing correlation between behavior generated by LLM-archetypes and census data as we add more contextual information to the prompt. This demonstrates the method's ability to incorporate nuanced factors - like evolving environments and incentives - into agent decision making. Notably, archetypes capture positive time-varying correlations in census-level behaviors for 3 of 5 boroughs (roughly 5 million people across income and demographic), which is an encouraging result.

In the second experiment (March to May 2022), we test the ability of LLM-archetypes to simulate adaptive behavior over time, particularly "pandemic fatigue". This period represents the scenario post-Omicron wave, when stimulus and unemployment payments

had also declined considerably. We modify the prompt to highlight the time duration since the start of the pandemic and the lack of financial incentives, testing whether our representative (and individual) agents are sensitive to these changes. We then repeat the population sampling procedure described in the previous experiment to obtain the time-series and examine cross-correlation with real-world data on participation rates.

This analysis demonstrates the potential of LLM archetypes to capture complex, adaptive behaviors in large-scale populations while maintaining computational efficiency. Despite high-variance individual responses, archetypes successfully capture positive time-varying correlations with census-level behaviors, an encouraging outcome for large-scale ABMs. To mitigate biased LLM responses, we estimation archetype distributions via multiple LLM generations (M queries per archetype), as motivated by [28]. While these results are promising, we acknowledge ongoing challenges. First, real-world behaviors can be significantly more complex than what our prompt can capture and second, data contamination in LLMs remains an open challenge with no formal technique to design prompts for LLM queries. Despite these limitations, the computational efficiency of archetypes - requiring only 400 queries for 8.4 million agents - compared to millions in the conventional approach represents a significant advance. This efficiency, combined with the ability to capture adaptive behaviors, makes LLM archetypes a promising tool for large-scale ABMs. We demonstrate the viability of this approach for large-scale simulations in Section 6, presenting performance benchmarking results of online LLM sampling coupled with a highly complex environment.

6 BEHAVIOR AGENCY VS SIMULATION SCALE

This section investigates the impact of incorporating adaptive agent behavior within ABMs and analyzes the trade-off between agency and simulation scale when calibrating to real-world data. We simulate dynamics of disease spread and labor market in New York City from December 2020 to April 2021 and, compare three kinds

¹<https://www.bls.gov/charts/employment-situation/civilian-labor-force-participation-rate.htm>

of agent architectures: LLM-as-agent, Heuristics and Archetypes. LLM-as-agent instantiates a unique LLM query per individual agent, Archetype instantiates M LLM queries per representative agent and heuristic agents used hand-crafted behaviors shared across all agents. Specifically, heuristic agents use Equation 8 and Equation 9 as they are. For archetype and LLM-as-agent, these equations are modified to incorporate an action α determined by the LLM output and obtain agent decision to isolate (I_j) and work (W_j). In terms of I_j , this is defined as:

$$I_j^{\text{LLM}} := I_j(t)(1 - \ell_\alpha(s_j(t), e(t))), \quad (10)$$

where ℓ_α is the LLM output for the action, and

$$I_j^{\text{archetype}}(t) := I_j(t)(1 - A_j(t)), \quad (11)$$

where $A_j(t)$ is sampled from Bernoulli($p_j(\alpha)$) with $p_j(\alpha)$ is estimated using the LLM (see section 3). Similarly, $W_j(t)$ is also modified for LLM-as-agent and archetypes.

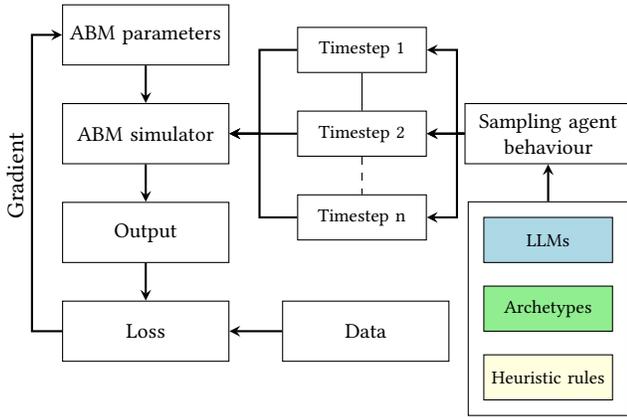


Figure 3: Calibration protocol for the three types of agent behaviours considered. This involves simulating ABM by sampling agent behavior at each step, comparing outputs to real-world data, and adjusting parameters through gradient-based optimization.

When simulating, we sample agent decisions at each step to execute dynamics and, aggregate infection and employment states after N steps. The aggregated outputs are used to calibrate structural parameters $\theta = (\beta, \gamma_0, \gamma_1)$ to historical time-series of infections and unemployment rates, using the protocol visualized in Figure 3. LLMs-as-agents query behavior at individual level and hence using it to simulate 8.4 million agents is computational infeasible. This enforces a trade-off between simulation scale and individual agency. For fair comparison, we fix the prompt budget to 300 LLM queries per step and compare the following configuration: a) Heuristic agents simulate 8.4 million agents using hand-crafted behaviors, b) Archetypes simulate 8.4 million agents with representative-level LLM queries and c) LLM-as-agents simulate a smaller population of 300 agents with individual-level LLM queries. Output of LLM-as-agents are scaled to the full population to compare with historical data and evaluate performance. To evaluate each calibrated model, we simulate a future time-series of 16 weeks for infection data

(measured weekly) and 80 weeks for employment rates (measured monthly) and report forecasting errors.

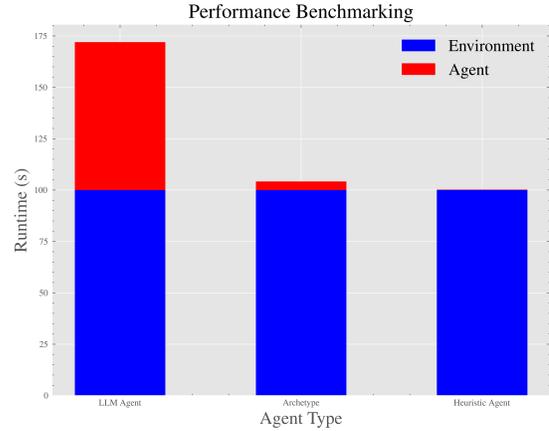


Figure 4: Runtime benchmarks for the environment and agent. Archetypes introduce much lower runtime overhead, enabling the simulation to scale to larger population size

Analysis: Results presented in Table 1 show that archetype-based model achieves the best performance, highlighting both the need for adaptive and expressive agents and the requirement of simulating the entire scale of the system. Further, heuristic agents simulated at population scale outperform LLM agents constrained to small population samples which shows that computational scaling is crucial. The benefit of LLM-archetypes over heuristic agents shows the benefit of capturing behavioral adaptations can be extremely useful at the right simulation scale. Figure 4 shows that Archetypes achieve this superior performance while consuming 95% less run-time compared to LLM-as-agents and marginally more than heuristic agents, which is encouraging for practical utility.

Agent Error Rate	Unemployment (\downarrow)	Infection (\downarrow)
Archetype	24.59 \pm 1.5	95.17 \pm 20.23
Heuristic	41.05 \pm 0.1	2914.73 \pm 300.25
LLM-as-agent	56.98 \pm 2.5	4311.70 \pm 674.14

Table 1: Benchmark results showing the mean-square errors for each of the considered agent architectures. Archetype achieves lower test error when forecasting both infections and unemployment rates as they capture adaptive agent behaviors without compromising simulation scale.

7 COUNTERFACTUAL SIMULATIONS

The Lucas Critique [27] posits that historical data alone can never predict the outcome of new policy, since behavior may adapt while outcomes are realized. By balancing individual agency and simulation scale, LLM archetypes allow us to analyze the relative impact of behavior adaptation and environment dynamics in shaping real world outcomes. We explore the interplay of pandemic fatigue and variant transmissibility on severity of disease outbreaks.

We consider transmission of two variants of COVID-19 - Delta ($\beta = [2.5 - 4.0]$, April 2021) and Omicron ($\beta = [5.5 - 8.0]$, November

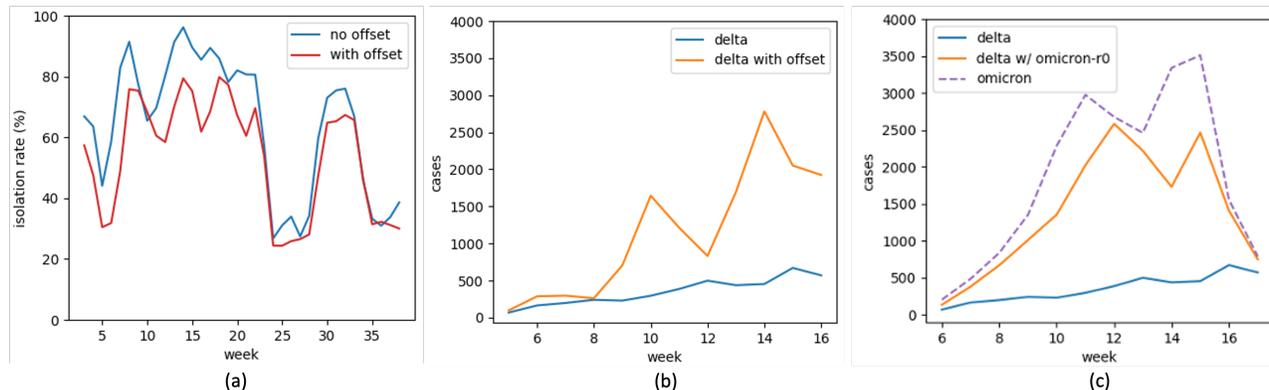


Figure 5: LLM archetypes help explore the interplay between behavior adaptation and environment dynamics in shaping epidemic outcomes. (left) Introducing pandemic fatigue ("the offset") to the prompt reduces relative rates of isolation behavior in the population. (middle) This decrease in isolation behavior translates to increased disease transmission in the population. (right) Comparing the original delta wave (in blue), delta wave with "omicron-like" transmissibility (in orange) and the omicron wave (shown in dashed purple to indicate this emerges at a later time). The omicron wave achieves a higher peak than both the original and "omicron-like" delta wave due to coupled impact of viral transmission and time-induced pandemic fatigue.

2021) - which emerged at different stages of the pandemic. While Omicron was roughly 2-3 times more transmissible than Delta, it produced 5-20 times the case intensity [1]. We hypothesize this is due to the coupled dynamics of time-induced fatigue and increased transmissibility. Using our model calibrated to the Delta wave, we conduct two counterfactual simulations:

Q1: What if we had the delta wave later?: To simulate time-induced behavior change ("pandemic fatigue"), we update the user prompt with an artificial offset: "it has been number of {weeks + OFFSET} since start of the epidemic", where OFFSET is 30 weeks.

Q2: What if we had the omicron wave earlier?: We update β in the simulation trace to mimic the Omicron variant while keeping the same behavioral dynamics.

Figure 5 illustrates the results of our counterfactual analysis, helping decouple the impact of behavior change and viral transmissibility on disease waves during the pandemic. First, we observe that time-induced fatigue can alter behavior of individuals with agents demonstrating lower willingness to isolate when prompted using the OFFSET (Figure 5(a)). This behavior change can result in a more severe outbreak even with viral dynamics do not change (Figure 5(b)). Second, an early onset of "omicron-like variant" would have been more destructive due to higher transmissibility, but not as severe as the actual omicron wave. The actual Omicron wave was exacerbated by the coupled influence of behavior change (additional "pandemic fatigue" due to extended duration). Such analyses can inform policy decisions during a pandemic helping appropriately allocate resources to both clinical and behavioral interventions.

We note that archetypes enable such analysis since they provide the ability to query adaptations in individual behavior, via expressive natural language, and also measure the cascading impact of individual decisions at a population scale. These analyses are challenging with other agent architectures due to trade-offs between simulation scale (compromised when using LLMs for each agent) and individual agency (lost when using heuristic agents). We note

that for small populations with high-personalized interventions, LLMs-as-agent can be viable architecture (as explored in works like [36, 48]). Archetypes are useful when analyzing outcomes in large populations with demographic-resolved interventions as often required for policy making.

8 DISCUSSION

This section addresses implementation considerations and limitations. We analyze sensitivity of Archetype-specific design choice when querying LLMs and introduces an API to generalize utility of our contributions to diverse scenarios. We also discuss some promising future capabilities and summarize limitations of the work.

LLM Consistency: LLM archetypes are sensitive to the quality and consistency of agent behaviors, which can vary with the choice of choice of model and number of queries per archetype (M in equation 7). We repeat the experiment in section 3 (using Prompt 3) and analyze sensitivity of individual decisions to model choice (GPT-3.5 vs GPT-4o) and number of instances per archetype ($M=1, 3, 6$). Results are shown in Figure 6. First, for lower M , using the superior GPT-4o model improves performance. Our algorithm is agnostic to the choice of LLMs and we anticipate the our results will become progressively better as LLMs mature, making ABMs more reliable. Second, for larger M - when archetype distributions are aggregated over multiple queries - performance significantly improves for GPT-3.5. We hypothesize that this mitigates biased LLM responses and helps capture realistic variability in individual behaviors. For future work, we plan to extensively benchmark different models (both open and closed-source) and design formal guidelines to specify archetype prompts. Finally, we also note that the choice of LLM is not the only factor affecting effectiveness. As demonstrated in Section 6, population scale is also critical for ABMs where heuristic agents can outperform LLM agents. This highlights the complex interplay between agent sophistication and simulation scale in determining overall ABM utility.

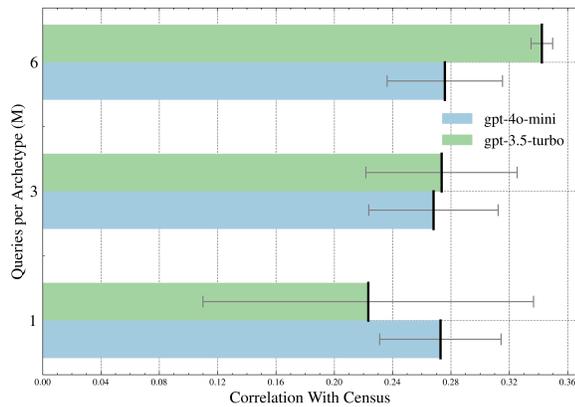


Figure 6: Sensitivity analysis to LLM model choice (GPT-3.5, GPT-4o) and number of queries per archetype (M=1, 3, 6)

Archetype API: We extend the AgentTorch framework [12] to integrate an ‘Archetype’ API, allowing the use of LLMs to prompt agent behavior in large-scale simulations. This API supports both of-line and online LLMs, facilitating wider adoption of LLM archetypes in ABMs. We present a code snippet in Figure 7.

```

from agent_torch.core import Archetype, Behavior
from agent_torch.populations import NYC

# Create an object of Archetype class
# n_arch estimates a predictive posterior over outcomes
archetype = Archetype(n_arch=7)

# Create an object of Behavior class
work_behavior = Behavior(archetype=archetype.llm(prompt),
                        region=NYC)

will_work = work_behavior.sample()

```

Figure 7: We extend the AgentTorch framework to generalize use of LLM Archetypes.

High-resolution Analysis: LLM archetypes enable measuring granular individual behavior and integrating it with population-scale simulations. Using our model calibrated to borough-level data, we measure the impact of stimulus payments on employment behavior at a granular zip-code level. In future, such analyses can help overcome limitations of historical data in policy design (Figure 8).

Limitations: While our work demonstrates the utility of LLM archetypes in large-scale ABMs, several areas for improvement remain. First, ensuring the robustness and fairness of LLM-driven agents remains an open challenge, as LLMs can produce inconsistent or biased outputs, potentially leading to unrealistic agent behaviors. Future work should focus on developing methods to detect and mitigate these biases. Second, while LLM archetypes aid scalability, they may not always capture the desired heterogeneity of individual agents, necessitating the development of more sophisticated archetype selection and interpolation methods. Third, the current implementation focuses on relatively simple agent actions, limiting the complexity of decision-making processes that

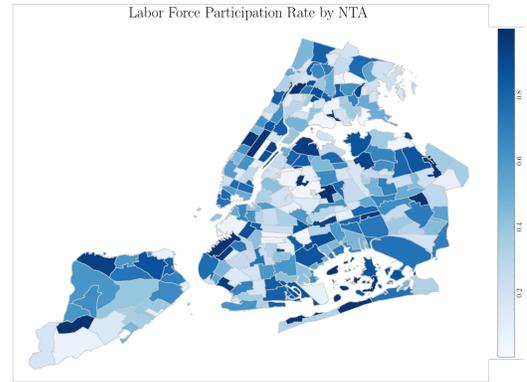


Figure 8: Zip-code level employment behavior for 8.4 million agents in NYC estimated using a model calibrated to coarse borough-level data. LLM Archetypes help overcome limitations of historical data for policy design.

can be modeled. Extending the action space and implementing multi-scale archetypes could address this limitation. Fourth, verifying the accuracy of individual agent behaviors generated by LLMs remains challenging, calling for the development of formal verification methods and benchmarks. We currently measure performance via comparisons with mesoscopic census data and generalization of macroscopic ABM predictions. Despite these limitations, we believe this work represents a significant step forward in agent-based modeling and opens new possibilities to understand and address societal challenges.

9 CONCLUSION

This work introduces LLM archetypes as a novel approach to scale adaptive agent behavior in large-scale agent-based models (ABMs). By efficiently integrating LLMs into ABMs, we enable the simulation of millions of agents with nuanced, context-aware behaviors while maintaining computational feasibility. Our case study on the COVID-19 pandemic in New York City demonstrates the power of this approach in capturing complex societal dynamics, balancing individual agency with population-scale outcomes. The framework’s ability to perform counterfactual analyses addresses key limitations in policy design, offering a powerful tool for tackling real-world challenges. While challenges remain in ensuring robustness and fairness of LLM-driven agents, this work represents a significant step forward in ABM capabilities. By bridging the gap between expressive individual agents and large-scale simulations, our approach opens new avenues for modeling complex systems and informing data-driven policy decisions across various domains.

ACKNOWLEDGMENTS

This research was funded by the MIT Media Lab consortium. Arnau Quera-bofarull acknowledges support by UKRI AI World Leading Researcher Fellowship awarded to Wooldridge (grant EP/W002949/1) and Trustworthy AI - Integrating Learning, Optimisation and Reasoning (TAILOR), a project funded by European Union Horizon2020 research and innovation program under Grant Agreement 952215.

REFERENCES

- [1] Mohammad Mamun Alam, Sumaiya Binte Hannan, Tanvir Ahmed Saikat, Md Baylayet Hasan Limon, Md Raihan Topu, Md Jowel Rana, Asma Salauddin, Sagar Bosu, and Mohammed Ziaur Rahman. 2023. Beta, Delta, and Omicron, Deadliest Among SARS-CoV-2 Variants: A Computational Repurposing Approach. *Evolutionary Bioinformatics* 19 (2023), 11769343231182258.
- [2] Philipp Andelfinger. 2023. Towards Differentiable Agent-Based Simulation. *ACM Transactions on Modeling and Computer Simulation* 32, 4 (Jan. 2023), 27:1–27:26. <https://doi.org/10.1145/3565810>
- [3] Rohan Anil, Andrew M Dai, and Yonghui et al Wu. 2023. PaLM 2 Technical Report. <https://doi.org/10.48550/arXiv.2305.10403> arXiv:2305.10403 [cs]
- [4] Gaurav Arya, Moritz Schauer, Frank Schäfer, and Christopher Rackauckas. 2022. Automatic Differentiation of Programs with Discrete Randomness. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 10435–10447.
- [5] R. Axtell and J. Farmer. 2022. Agent Based Modeling in Economics and Finance: Past, Present, and Future. *Journal of Economic Literature* (2022).
- [6] Robert L. Axtell. 2016. 120 Million Agents Self-Organize into 6 Million Firms: A Model of the U.S. Private Sector. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 806–816.
- [7] Joseph Aylett-Bullock, Carolina Cuesta-Lazaro, Arnau Quera-Bofarull, Miguel Icaza-Lizaola, Aidan Sedgewick, Henry Truong, Aoife Curran, Edward Elliott, Tristan Caulfield, Kevin Fong, Ian Vernon, Julian Williams, Richard Bower, and Frank Krauss. 2021. June: Open-Source Individual-Based Epidemiology Simulation. *Royal Society Open Science* 8, 7 (July 2021), 210506. <https://doi.org/10.1098/rsos.210506>
- [8] Joseph Aylett-Bullock, Robert Tucker Gilman, Ian Hall, David Kennedy, Emond Samir Evers, Anjali Katta, Hussien Ahmed, Kevin Fong, Keyrellous Adib, Lubna Al Ariqi, et al. 2022. Epidemiological modelling in refugee and internally displaced people settlements: challenges and ways forward. *BMJ Global Health* 7, 3 (2022), e007822.
- [9] Eric Bonabeau. 2002. Agent-Based Modeling: Methods and Techniques for Simulating Human Systems. *Proceedings of the National Academy of Sciences* 99, suppl_3 (May 2002), 7280–7287. <https://doi.org/10.1073/pnas.082080899>
- [10] Ayush Chopra, Esmat Gel, Jayakumar Subramanian, Balaji Krishnamurthy, Santiago Romero-Brufau, Kalyan S Pasupathy, Thomas C Kingsley, and Ramesh Raskar. 2021. DeepABM: scalable, efficient and differentiable agent-based simulations via graph neural networks. In *Winter Simulation Conference (WSC)*.
- [11] Ayush Chopra, Alexander Rodriguez, Jayakumar Subramanian, Arnau Quera-Bofarull, Balaji Krishnamurthy, B. Aditya Prakash, and Ramesh Raskar. 2023. Differentiable Agent-based Epidemiology. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1848–1857.
- [12] Ayush Chopra, Jayakumar Subramanian, Balaji Krishnamurthy, and Ramesh Raskar. 2024. A Framework for Learning in Agent-Based Models. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems.
- [13] Christian Dreger and Daniel Gros. 2021. Lockdowns and the US unemployment crisis. *Economics of Disasters and Climate Change* 5, 3 (2021), 449–463.
- [14] Joel Dyer, Patrick Cannon, J. Doyne Farmer, and Sebastian M. Schmon. 2024. Black-Box Bayesian Inference for Agent-Based Models. *Journal of Economic Dynamics and Control* 161 (April 2024), 104827. <https://doi.org/10.1016/j.jedc.2024.104827>
- [15] Joel Dyer, Arnau Quera-Bofarull, Ayush Chopra, J. Doyne Farmer, Anisoara Calinescu, and Michael Wooldridge. 2023. Gradient-Assisted Calibration for Financial Agent-Based Models. In *Proceedings of the Fourth ACM International Conference on AI in Finance (Icaif '23)*. Association for Computing Machinery, New York, NY, USA, 288–296. <https://doi.org/10.1145/3604237.3626857>
- [16] Elena Falcattoni and Vegard M Nygaard. 2020. Acts of Congress and COVID-19: A literature review on the impact of increased unemployment insurance benefits and stimulus checks. (2020).
- [17] Jiaqi Ge. 2017. Endogenous rise and collapse of housing price: an agent-based model of the housing market. *Computers, Environment and Urban Systems* 62 (2017), 182–198.
- [18] Saman Ghaffarian, Debraj Roy, Tatiana Filatova, and Norman Kerle. 2021. Agent-based modelling of post-disaster recovery with remote sensing data. *International Journal of Disaster Risk Reduction* 60 (2021), 102285.
- [19] Navid Ghaffarzadegan, Aritra Majumdar, Ross Williams, and Niyousha Hoseinichimeh. 2024. Generative agent-based modeling: an introduction and tutorial. *System Dynamics Review* 40, 1 (2024), e1761.
- [20] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).
- [21] Robert Hinch, William J. M. Probert, Anel Nurtay, Michelle Kendall, Chris Wymant, Matthew Hall, Katrina Lythgoe, Ana Bulas Cruz, Lele Zhao, Andrea Stewart, Luca Ferretti, Daniel Montero, James Warren, Nicole Mather, Matthew Abueg, Neo Wu, Olivier Legat, Katie Bentley, Thomas Mead, Kelvin Van-Vuuren, Dylan Feldner-Busztin, Tommaso Ristori, Anthony Finkelstein, David G. Bonsall, Lucie Abeler-Dörner, and Christophe Fraser. 2021. OpenABM-Covid19—An Agent-Based Model for Non-Pharmaceutical Interventions against COVID-19 Including Contact Tracing. *PLOS Computational Biology* 17, 7 (July 2021), e1009146. <https://doi.org/10.1371/journal.pcbi.1009146>
- [22] John J. Horton. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? <https://doi.org/10.48550/arXiv.2301.07543> [econ, q-fin]
- [23] Cliff C. Kerr, Robyn M. Stuart, Dina Mistry, Romesh G. Abeysuriya, Katherine Rosenfeld, Gregory R. Hart, Rafael C. Núñez, Jamie A. Cohen, Prashanth Selvaraj, Brittany Hagedorn, Lauren George, Michal Jastrzębski, Amanda S. Izzo, Greer Fowler, Anna Palmer, Dominic Delport, Nick Scott, Sherrie L. Kelly, Caroline S. Bennette, Bradley G. Wagner, Stewart T. Chang, Assaf P. Oron, Edward A. Wenger, Jasmina Panovska-Griffiths, Michael Famulare, and Daniel J. Klein. 2021. Covasim: An Agent-Based Model of COVID-19 Dynamics and Interventions. *PLOS Computational Biology* 17, 7 (July 2021), e1009149. <https://doi.org/10.1371/journal.pcbi.1009149>
- [24] Kangli Li, Natasha Zhang Foutz, Yuxin Cai, Yunlei Liang, and Song Gao. 2021. Impacts of COVID-19 lockdowns and stimulus payments on low-income population's spending in the United States. *PLoS one* 16, 9 (2021), e0256407.
- [25] Nian Li, Chen Gao, Yong Li, and Qingmin Liao. 2023. Large language model-empowered agents for simulating macroeconomic activities. *arXiv preprint arXiv:2310.10436* (2023).
- [26] Robert E. Lucas. 1976. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy* 1 (1976), 19–46. [https://doi.org/10.1016/S0167-2231\(76\)80003-6](https://doi.org/10.1016/S0167-2231(76)80003-6)
- [27] Robert E Lucas Jr. 1976. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, Vol. 1. North-Holland, 19–46.
- [28] Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896* (2023).
- [29] Zakaria Mehrab, Logan Stundal, Srinivasan Venkatramanan, Samarth Swarup, Bryan Lewis, Henning S Mortveit, Christopher L Barrett, Abhishek Pandey, Chad R Wells, Alison P Galvani, et al. 2024. An agent-based framework to study forced migration: A case study of Ukraine. *PNAS nexus* 3, 3 (2024), pgae080.
- [30] Corrado Monti, Marco Pangallo, Gianmarco De Francischi Morales, and Francesco Bonchi. 2023. On Learning Agent-Based Models from Data. *Scientific Reports* 13, 1 (June 2023), 9268. <https://doi.org/10.1038/s41598-023-35536-3>
- [31] Alexander Mordvintsev, Ettore Randazzo, Eyvind Niklasson, and Michael Levin. 2020. Growing neural cellular automata. *Distill* 5, 2 (2020), e23.
- [32] Alexander Mordvintsev, Ettore Randazzo, Eyvind Niklasson, and Michael Levin. 2020. Growing Neural Cellular Automata. *Distill* (2020). <https://distill.pub/2020/growing-ca/>
- [33] Elias Najjarro, Shyam Sudhakaran, Claire Glanois, and Sebastian Risi. 2022. HyperNCa: Growing developmental networks with neural cellular automata. *ICLR Workshop on Cells to Societies* (2022).
- [34] NYC 311. 2024. CARES Act Stimulus Payments in NYC. <https://portal.311.nyc.gov/article/?kanumber=KA-03373> Accessed October 16, 2024.
- [35] OpenAI. 2024. GPT-4 Technical Report. <https://doi.org/10.48550/arXiv.2303.08774> arXiv:2303.08774 [cs]
- [36] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. <https://doi.org/10.48550/arXiv.2304.03442> arXiv:2304.03442 [cs]
- [37] Donovan Platt. 2020. A Comparison of Economic Agent-Based Model Calibration Methods. *Journal of Economic Dynamics and Control* 113 (April 2020), 103859. <https://doi.org/10.1016/j.jedc.2020.103859>
- [38] Chenyuan Qin, Jie Deng, Min Du, Qiao Liu, Yaping Wang, Wenxin Yan, Min Liu, and Jue Liu. 2023. Pandemic fatigue and vaccine hesitancy among people who have recovered from COVID-19 infection in the post-pandemic era: cross-sectional study in China. *Vaccines* 11, 10 (2023), 1570.
- [39] Arnau Quera-Bofarull, Ayush Chopra, Joseph Aylett-Bullock, Carolina Cuesta-Lazaro, Anisoara Calinescu, Ramesh Raskar, and Michael Wooldridge. 2023. Don't Simulate Twice: One-shot Sensitivity Analyses via Automatic Differentiation. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (Aamas '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1867–1876.
- [40] Arnau Quera-Bofarull, Ayush Chopra, Joseph Aylett-Bullock, Carolina Cuesta-Lazaro, Anisoara Calinescu, Ramesh Raskar, and Michael Wooldridge. 2023. Don't simulate twice: one-shot sensitivity analyses via automatic differentiation. (2023).
- [41] Arnau Quera-Bofarull, Ayush Chopra, Anisoara Calinescu, Michael Wooldridge, and Joel Dyer. 2023. Bayesian Calibration of Differentiable Agent-Based Models. <https://doi.org/10.48550/arXiv.2305.15340> arXiv:2305.15340 [cs, stat]

- [42] Paul Richmond, Robert Chisholm, Peter Heywood, Mozghan Kabiri Chimeh, and Matthew Leach. 2023. FLAME GPU 2: A Framework for Flexible and Performant Agent Based Simulation on GPUs. *Software: Practice and Experience* 53, 8 (2023), 1659–1680. <https://doi.org/10.1002/spe.3207>
- [43] Santiago Romero-Brufau, Ayush Chopra, Alex J Ryu, Esmá Gel, Ramesh Raskar, Walter Kremers, Karen S Anderson, Jayakumar Subramanian, Balaji Krishnamurthy, Abhishek Singh, et al. 2021. Public health impact of delaying second dose of BNT162b2 or mRNA-1273 covid-19 vaccine: simulation agent based modeling study. *bmj* 373 (2021).
- [44] Clodomir Santana, Federico Botta, Hugo Barbosa, Filippo Privitera, Ronaldo Menezes, and Riccardo Di Clemente. 2023. COVID-19 is linked to changes in the time–space dimension of human mobility. *Nature Human Behaviour* 7, 10 (2023), 1729–1739.
- [45] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality Traits in Large Language Models. <https://doi.org/10.48550/arXiv.2307.00184> arXiv:2307.00184 [cs]
- [46] Alvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro J. Gonçalves, David S. Greenberg, and Jakob H. Macke. 2020. Sbi: A Toolkit for Simulation-Based Inference. *Journal of Open Source Software* 5, 52 (2020), 2505. <https://doi.org/10.21105/joss.02505>
- [47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. <https://doi.org/10.48550/arXiv.2302.13971> arXiv:2302.13971 [cs]
- [48] Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Duéñez-Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. 2023. Generative Agent-Based Modeling with Actions Grounded in Physical, Social, or Digital Space Using Concordia. <https://doi.org/10.48550/arXiv.2312.03664> arXiv:2312.03664 [cs]
- [49] Robert Yang. 2024. Project Sid: Building Digital Humans. <https://altera.al/> Accessed on October 16, 2024.