UNIVERSIDADE PRESBITERIANA MACKENZIE
GRADUATE PROGRAM IN
ELECTRICAL ENGINEERING AND COMPUTING

Fernando Tenório de Miranda Filho

SIMULATING OPINION DYNAMICS WITH LLM AGENTS: INSIGHTS FROM SILICON SAMPLING AND REAL-WORLD DATA

Qualification text presented to the Programa de Pós-Graduação em Engenharia Elétrica e Computação da Universidade Presbiteriana Mackenzie as a partial requirement for obtaining the Dr. degree in Electrical Engineering and Computing

Orientador: Prof. Dr. Pedro Paulo Balbi de Oliveira

São Paulo
2024

Agradecimentos

Agradeço à Universidade Presbiteriana Mackenzie pela bolsa de Doutorado concedida. Agradeço também ao meu orientador, PP, pelo enorme apoio durante essa jornada.

RESUMO

Simular opiniões humanas com precisão é essencial para compreender fenômenos sociais, como a polarização e a disseminação de desinformação. Modelos tradicionais baseados em agentes muitas vezes simplificam demais essa tarefa, reduzindo interações complexas em linguagem natural a regras estáticas, o que limita sua capacidade de capturar as sutilezas da comunicação humana. No entanto, avanços recentes em Inteligência Artificial, especialmente com Modelos de Linguagem de Grande Escala (LLMs), oferecem novas oportunidades para simular e entender a formação de opiniões humanas de maneira mais eficaz. Quando os LLMs participam de trocas de opinião em uma estrutura de rede social, tendem a alcançar um consenso sobre um determinado tema, mesmo quando inicialmente configurados com pontos de vista divergentes. Esse consenso é alcançado mais rapidamente quando o tema é um fato científico bem estabelecido. Embora esse comportamento destaque as capacidades dos LLMs, ele também apresenta uma limitação significativa: no mundo real, indivíduos de diferentes origens demográficas e culturais frequentemente mantêm perspectivas variadas sobre o mesmo assunto. Esta pesquisa visa avaliar quão precisamente os LLMs podem simular distribuições de opiniões humanas em diferentes tópicos, usando dados demográficos e comportamentais de pesquisas reais. Especificamente, utilizamos dados do American National Election Studies (ANES) para avaliar o desempenho de vários LLMs na replicação das distribuições de opinião do mundo real. Ao condicionar os LLMs a interpretar o papel de respondentes individuais—uma técnica conhecida como "silicon sampling"—investigamos como fatores como o tamanho do modelo e a censura influenciam os resultados, e quais variáveis desempenham o papel mais crítico em inferir os dados das pesquisas.

Palavras-chave: *llm, dinâmica de opiniões, consenso, modelos baseados em agentes, redes sociais, dados demográficos.*

ABSTRACT

Simulating human opinions accurately is essential for understanding societal phenomena such as polarization and the spread of misinformation. Traditional agent-based models often oversimplify this task by reducing complex natural language interactions to static rules, which limits their ability to capture the subtleties of human communication. However, recent advances in Artificial Intelligence, particularly with Large Language Models (LLMs), offer new opportunities for simulating and understanding human opinion formation more effectively. When LLMs engage in opinion exchanges within a social network framework, they tend to reach a consensus on a given topic, even when initially seeded with differing viewpoints. This consensus is achieved more quickly when the topic is a well-established scientific fact. While this behavior highlights the capabilities of LLMs, it also presents a significant limitation: in the real world, individuals from diverse demographic and cultural backgrounds often hold varied perspectives on the same issue. This research aims to assess how accurately LLMs can simulate distributions of human opinions across different topics, using demographic and behavioral data from real-world surveys. Specifically, we utilize data from the American National Election Studies (ANES) to evaluate the performance of various LLMs in replicating real-world opinion distributions. By conditioning LLMs to role-play individual respondents—a technique known as "silicon sampling"—we investigate how factors such as model size and censorship influence the results, and which variables play the most critical role in matching survey data.

Keywords: *llm, opinion dynamics, consensus, agent-based models, social networks, demographic data.*

Contents

Lista de Figuras

1 Introduction

Human beliefs and opinions play a central role in shaping societal interactions and behaviors. From our preferences for food and entertainment to our participation in political processes, these beliefs influence daily decisions and broader social dynamics.

The foundation of our beliefs and opinions is not genetically predetermined but acquired through a variety of learning experiences. Within family units, parents impart basic principles and beliefs to their children. Beyond the family, much of this learning occurs through "social learning" processes, where individuals gather information and update their beliefs based on personal experiences, observations of others' behaviors, communications within their social networks, and media consumption. Social learning is a multi-faceted process that involves [2].

1. **Social Networks**: An individual's learning is influenced by the behavior and information from a select subset of society, including friends, coworkers, family members, and trusted leaders. This network context is inseparable from the learning and opinion formation process.

2. **Interpretation in a Social Context**: Individuals interpret information through a social lens, which involves trusting certain sources over others and forming conjectures about the intentions behind the information provided by their network members.

3. **Dynamics of Information Spread**: Information obtained from one network member can be passed on to others, leading to a dynamic spread of beliefs and opinions across overlapping and non-overlapping social networks. This dynamic process can propagate both accurate and inaccurate information far beyond the initial source.

Given the critical role of beliefs and opinions in modern society, studying opinion dynamics is vital. Agent-based models (ABMs) are a classical approach in the study of opinion dynamics [27, 22, 20]. In ABMs, individuals are represented as agents that are distributed in the society according to some fixed network topology. When two agents interact, their opinions are updated according to some mathematical expression. As the simulation progresses, a stable state of opinion distribution typically emerges, which can represent consensus, polarization, or fragmentation of opinions within the population.

Despite their usefulness, traditional models of opinion dynamics often oversimplify this process by representing opinions as real numbers and failing to capture the intricate nature of human communication. However, recent advancements in artificial intelligence, particularly with large language models (LLMs), provide new opportunities to simulate and understand opinion dynamics more precisely.

1.1 Background and Motivation

Understanding and accurately simulating human beliefs and opinions is essential for analyzing societal phenomena such as political polarization, the spread of misinformation, and public sentiment on key issues. Public opinions shape social interactions and drive collective behaviors, influencing domains ranging from marketing and entertainment preferences to policy-making and electoral processes. Capturing these dynamics in computational models is a longstanding challenge in fields like sociology, political science, and artificial intelligence (AI), as human beliefs are shaped by a complex interplay of social, cultural, and demographic factors.

Agent-Based Models (ABMs) have been a popular method for studying opinion dynamics. ABMs represent individuals as autonomous agents in a simulated environment, where each agent interacts with others based on predefined rules and updates its state according to specific conditions. These models allow researchers to observe emergent behaviors, such as consensus formation, polarization, or clustering of opinions, from the bottom up (i.e., through simple local interactions among agents). ABMs have successfully illustrated how social structures and network topologies affect opinion formation and have contributed valuable insights into mechanisms like echo chambers and social contagion. However, traditional ABMs lack in their ability to simulate the linguistic and cognitive complexity of real human interactions. They often rely on simple numerical representations of opinions and rule-based interactions, which do not capture the rich, nuanced, and context-dependent nature of language-driven opinion formation [27, 20].

This gap has become more evideng role of digital media in shaping public opinion. In modern societies, much of opinion formation and information dissemination occurs through textual interactions—on social media platforms, news articles, and online forums—where people not only express opinions but also influence others through nuanced language. Traditional ABMs struggle to model these linguistic interactions, especially in

contexts where opinions are shaped by subtle differences in wording, framing, or tone. As a result, these models often fall short of capturing real-world opinion dynamics, especially for topics where cultural, demographic, or ideological backgrounds significantly influence responses.

Recent advancements in Artificial Intelligence, particularly in large language models (LLMs) like GPT-4,have opened new possibilities for simulating social phenomena with greater linguistic fidelity. LLMs are neural network models trained on massive datasets of human text, enabling them to generate language that is contextually coherent and, to some extent, reflective of human behavior and reasoning patterns. Unlike the agents in traditional ABMs, which operate based on static rules or simplified representations, LLMs can produce responses that incorporate a range of contextual, semantic, and demographic cues. This ability makes them suitable candidates for simulating more complex aspects of human social interaction and opinion formation, as they can "roleplay" individuals and provide responses influenced by simulated demographic or cultural backgrounds.

In particular, "silicon sampling" is a novel approach that roleplay survey respondents, conditioned on specific demographic or behavioral data. By guiding LLMs to respond as individuals from particular backgrounds—whether by age, gender, education, or ideology—researchers can simulate a distribution of opinions that mirrors the diversity of real-world survey respondents. This approach enables a new form of opinion simulation that not only includes more nuanced language-based interactions but also reflects demographic diversity, which is crucial for accurately modeling public opinion distributions. In this research, data from the American National Election Studies (ANES) is used for evaluating the effectiveness of LLMs in this role, with the aim of assessing how closely these models can replicate actual opinion distributions when conditioned on real demographic and behavioral variables [46, 10, 9].

Ultimately, this study addresses a key question in the social simulation domain: Can LLMs serve as effective human emulators in opinion dynamics models? By bridging the gap between traditional ABMs and advanced language models, this research aims to contribute to a more accurate and dynamic understanding of public opinion simulation, with potential implications for fields such as sociology, political science, and artificial intelligence. As LLMs continue to advance, they offer the possibility of more realistic simulations, which could lead to better-informed decisions in policy-making, marketing,

and other areas where understanding public sentiment is crucial.

1.2 Research Objectives and Questions

The primary objective of this research is to assess the capability of LLMs to simulate real-world distributions of human opinions across various topics, by conditioning the models on demographic and behavioral data. By examining the effectiveness of LLMs in replicating public opinion distributions from survey data, this research aims to bridge the gap between traditional agent-based modeling approaches and modern AI-driven simulations. This novel application of LLMs, known as "silicon sampling", involves configuring LLMs to roleplay individual respondents from specific demographic backgrounds, thus providing a closer approximation to real-world opinion diversity.

This objective is broken down into the following specific research goals:

1. **Evaluate the accuracy of LLMs in replicating opinion distributions**: Measure how closely the opinions generated by LLMs align with real survey data, using different metrics and statistics.

2. **Analyze the influence of demographic conditioning on LLM outputs**: Assess how effectively demographic information (e.g., age, gender, ideology) can be used to condition LLMs, and determine whether this conditioning leads to outputs that accurately reflect diverse perspectives in survey data.

3. **Investigate the impact of model characteristics on performance**: Explore how factors such as model size, architecture, and censorship mechanisms affect the LLMs' ability to simulate opinion distributions. This includes evaluating which model configurations yield the most realistic results and understanding the limitations of different LLMs in emulating human-like opinions.

4. **Identify key variables that affect the alignment of LLM-generated opinions with real data**: Determine which demographic and contextual variables (e.g., topic type, population subgroup) have the strongest influence on the alignment of LLM-generated opinions with real-world survey data.

To achieve these objectives, the research addresses the following research questions:

1. How accurately can LLMs simulate opinion distributions across different topics when conditioned on demographic data?

2. To what extent do demographic variables (e.g., age, gender, political affiliation) influence the outputs generated by LLMs?

3. What effect do LLM characteristics (such as model size, architecture, and censorship mechanisms) have on the accuracy of simulated opinion distributions?

4. Which variables and configurations are most critical for aligning LLM-generated opinions with real survey data?

The research is guided by several hypotheses:

1. **H1**: LLMs conditioned on demographic data can generate opinion distributions that closely match real-world survey distributions for **some topics**, **not all** topics.

2. **H2**: Larger LLMs, specifically those with hundres of billions of parameters, are more effective in producing opinion distributions that align with real data, as they have more capacity to represent complex demographic and linguistic nuances.

3. **H3**: Certain demographic variables, such as ideology and level of education, have a stronger impact on the alignment of simulated opinions with real-world data than others, due to their high influence on opinion formation.

4. **H4**: Censorship mechanisms in LLMs may reduce the model's ability to replicate certain opinions accurately, especially in polarized or sensitive topics.

By addressing these objectives and questions, this research aims to contribute to the emerging field of AI-driven social simulation, providing insights into how LLMs can be used as tools for simulating and analyzing public opinion. The findings are expected to offer new methodologies for simulating human beliefs and opinions, that could be used in policy analysis, market research, and other domains where understanding public sentiment is critical.

1.3 Thesis Structure

This thesis is organized into eight chapters, each building upon the foundation of using Large Language Models (LLMs) to simulate human opinion distributions. The chapters

explore theoretical backgrounds, methodological approaches, experimental findings, and conclusions related to the simulation of public opinion dynamics.

- **Chapter 2: Introduction to Agent-Based Models (ABMs)** This chapter introduces Agent-Based Models, a traditional method for studying social and opinion dynamics in computational simulations. It covers the fundamentals, including definitions, characteristics of agents, environment configurations, and different outcomes in ABM simulations. Additionally, it discusses the validation and calibration of ABMs, highlighting both their capabilities and limitations in modeling complex human behaviors. This chapter contextualizes the shift from traditional ABMs to LLM-based simulations by illustrating the challenges ABMs face in capturing nuanced, language-based interactions.

- **Chapter 3: Introduction to Opinion Dynamics** Chapter 3 provides an overview of opinion dynamics, exploring how individual beliefs and opinions are influenced by social networks and interpersonal interactions. It reviews classical models of opinion dynamics, such as the DeGroot model and bounded confidence models (e.g., DW and HK models), which describe how opinions evolve over time within a network. This chapter lays the theoretical groundwork for understanding the dynamics of opinion formation, which is essential for comparing the performance of traditional models with LLM-based approaches.

- **Chapter 4: Introduction to Large Language Models (LLMs)** This chapter presents the technological foundations of LLMs, tracing the development from early contextual embedding models like BERT to massive language models like GPT-4. It details key advancements in LLM training, scaling, and implementation that have made these models capable of generating human-like responses. The chapter also introduces the emergence of smaller, specialized LLMs, such as Phi-3 and LLaMA-3, which are particularly relevant for resource-efficient simulations. By discussing LLMs' capacity for social simulations, this chapter sets the stage for understanding their role as agents capable of simulating diverse opinion distributions.

- **Chapter 5: LLMs as Human Emulators: Literature Review** Chapter 5 reviews existing literature on the use of AI and LLMs for simulating human-like behaviors, focusing on opinion dynamics and multi-human simulation. It explores

the theoretical underpinnings and prior research on using LLMs to replicate human responses, specifically in social and opinion-based contexts. This chapter addresses both the potential and limitations of LLMs in emulating human behaviors, drawing attention to critical issues such as demographic conditioning and the impact of model censorship on output fidelity.

- **Chapter 6: Experimental Setup** This chapter outlines the experimental setup used to evaluate LLMs capacity to simulate human subjects accurately. It begins with an introduction to the ANES 2020 dataset and describes the sampling methodology used in the experiments. It then details the silicon sampling algorithm, which conditions LLMs to roleplay respondents from specific demographic groups, and provides an overview of the selected LLMs used in the study.

- **Chapter 7: Results and Insights** Chapter 7 presents a comprehensive discussion of the experimental findings from the previous chapter. It interprets the results in the context of the research objectives and questions, highlighting the extent to which LLMs can replicate real-world opinion distributions and identifying factors that impact simulation accuracy. This chapter addresses the efficacy of demographic conditioning in silicon sampling and evaluates the influence of model characteristics, such as size and architecture, on simulation fidelity.

- **Chapter 8: Concluding Remarks** The final chapter summarizes the research contributions, emphasizing the potential of LLMs as tools for simulating opinion dynamics. It discusses the broader implications of using LLMs for social simulation, the limitations encountered in this study, and directions for future research. Recommendations are offered for refining LLM-based simulation approaches to improve accuracy and applicability across various fields, including sociology, political science, and artificial intelligence.

2 Introduction to Agent-Based Models (ABMs)

2.1 Definition and Fundamentals of Agent-Based Models

Agent-Based Models (ABMs) are computational frameworks that simulate the actions and interactions of autonomous agents to assess their effects on the system as a whole. These models are rooted in the fields of complex systems, computational sociology, and artificial intelligence, providing a bottom-up approach to understanding macro-level phenomena from micro-level interactions. In essence, ABMs are comprised of individual agents, each possessing unique attributes and rules of behavior, situated within a defined environment. In an ABM, agents receive inputs from their environment and take actions in response to those inputs. The interactions of these agents generate emergent properties that can be observed and analyzed to gain insights into the dynamics of the system.

The fundamental components of ABMs include agents, rules, and the environment. Agents are the individual entities within the model that follow specified rules or strategies to make decisions and perform actions. These agents can represent various entities, such as individuals, groups, or organizations, depending on the context of the model. The rules governing agent behavior are designed to capture realistic decision-making processes, which can be deterministic or stochastic in nature. The environment provides the context within which agents interact, including spatial and social structures that influence agent behavior.

ABMs are particularly useful in studying complex adaptive systems where interactions between components are nonlinear and can lead to unexpected outcomes. This complexity is often characterized by feedback loops, adaptation, and evolution, making traditional analytical methods insufficient for capturing the dynamics of such systems. By allowing agents to interact based on simple rules, ABMs can simulate the emergence of complex patterns and structures that are difficult to predict through top-down approaches.

One of the strengths of ABMs is their flexibility and adaptability. Researchers can tailor models to fit specific research questions by adjusting agent characteristics, rules of interaction, and environmental parameters. This adaptability makes ABMs applicable across a wide range of disciplines, including economics, sociology, political science, ecology, and public health. By experimenting with different scenarios and parameters, researchers can explore how changes at the micro level influence macro-level outcomes, providing

valuable insights into the mechanisms driving system behavior [13].

2.2 Historical Development and Evolution of ABMs

The origins of ABMs can be traced back to the mid-20th century with the advent of cellular automata and early computer simulations. One of the pioneering works in this domain was John Conway's "Game of Life," developed in 1970, which demonstrated how simple rules applied to a grid of cells could generate complex patterns over time. This early example highlighted the potential of computational models to simulate emergent phenomena, laying the groundwork for the development of ABMs.

The formalization of ABMs as a distinct modeling approach began in the 1980s and 1990s, driven by advances in computer technology and the increasing recognition of the limitations of traditional mathematical models in capturing complex social phenomena. Researchers such as Robert Axelrod, Thomas Schelling, and Joshua Epstein made significant contributions to the field. Axelrod's work on the evolution of cooperation and Schelling's segregation model were particularly influential, demonstrating how ABMs could provide insights into social dynamics through the interaction of simple rules and behaviors.

The publication of Epstein and Axtell's "Growing Artificial Societies: Social Science from the Bottom Up" in 1996 marked a significant milestone in the development of ABMs. This work showcased the potential of ABMs to model a wide range of social phenomena, from population dynamics to economic markets, and highlighted the importance of agent heterogeneity and interaction in generating emergent properties. The book's emphasis on bottom-up modeling and the use of computational experiments inspired a new generation of researchers to adopt ABMs in their work [6, 21].

In the following decades, the field of ABMs expanded rapidly, benefiting from advancements in computational power, algorithm development, and data availability. The integration of ABMs with empirical data and the development of sophisticated modeling platforms, such as NetLogo, Repast, and AnyLogic, facilitated the creation and analysis of more complex and realistic models. The interdisciplinary nature of ABMs also led to their application in diverse fields, from studying the spread of infectious diseases to modeling financial markets and urban development [51].

2.3 Agent Characteristics and Behaviors

Agents in ABMs are autonomous entities that possess distinct characteristics and follow specific rules or strategies to make decisions and perform actions. These characteristics can include attributes such as age, gender, preferences, resources, and capabilities, which define the agent's state and influence its behavior. The heterogeneity of agents is a critical aspect of ABMs, as it allows for the representation of diversity within the modeled population and the exploration of how individual differences impact system dynamics.

The behavior of agents is governed by a set of rules or algorithms that dictate how they respond to various stimuli and interact with other agents and the environment. These rules can be simple, such as following a fixed strategy or probabilistic behavior, or more complex, involving learning and adaptation. For instance, agents may use heuristic rules to make decisions based on limited information, or they may employ sophisticated algorithms, such as reinforcement learning, to adapt their behavior based on past experiences and anticipated outcomes [13].

One of the key challenges in designing agent behaviors is ensuring that they are both realistic and computationally tractable. Researchers must strike a balance between capturing the complexity of real-world decision-making processes and maintaining the simplicity necessary for computational efficiency. This often involves abstracting certain aspects of behavior while retaining the core elements that drive system dynamics. Sensitivity analysis can help identify which aspects of agent behavior are most influential on model outcomes, guiding the refinement of behavioral rules.

The flexibility of ABMs in specifying agent behaviors allows researchers to explore a wide range of scenarios and hypotheses. By manipulating agent characteristics and behavioral rules, researchers can investigate how different assumptions and interventions influence the outcomes of the model. This ability to experiment with "what-if" scenarios makes ABMs a powerful tool for understanding the mechanisms driving complex systems and for designing policies and interventions that can lead to desired outcomes [43].

2.4 Environment and Interactions

The environment in an ABM provides the context within which agents interact and includes the spatial, social, and institutional structures that influence agent behavior. The environment can be modeled in various ways, from simple grids or networks to more

complex and realistic representations of geographic and social spaces. The choice of environmental structure depends on the research question and the specific dynamics being studied.

Spatial environments are commonly used in ABMs to represent physical spaces where agents move and interact. These environments can range from abstract grids to detailed geographic maps, depending on the level of realism required. Spatial interactions can include movement, proximity-based interactions, and resource consumption, among others. For example, in models of disease spread, the spatial environment can represent a city where agents move between different locations, influencing the transmission dynamics of the disease [47].

Social environments capture the relationships and interactions between agents, such as social networks, organizational structures, and institutional rules. Social networks are particularly important in ABMs that study information diffusion, social influence, and collective behavior. The structure of the social network, including the number and strength of connections between agents, can significantly impact the dynamics of the modeled system. Researchers can experiment with different network structures to explore how variations in social connectivity affect outcomes [28].

Interactions between agents and the environment are central to the dynamics of ABMs. These interactions can be direct, such as communication or competition for resources, or indirect, such as the influence of environmental changes on agent behavior. The rules governing these interactions are crucial for capturing the feedback loops and emergent properties characteristic of complex systems. For instance, in models of ecological systems, interactions between agents (e.g., predators and prey) and the environment (e.g., availability of resources) drive population dynamics and ecosystem stability.

The design of the environment and interaction rules requires careful consideration to ensure that the model accurately captures the relevant dynamics of the real-world system. Researchers must consider the scale and scope of the environment, the types of interactions, and the potential for feedback loops and emergent phenomena. Sensitivity analysis and validation against empirical data are essential for refining these aspects of the model and ensuring its robustness and relevance.

2.5 Types of Outcomes in ABMs

ABMs are capable of producing a variety of outcomes that can be broadly categorized into equilibria, cycles, randomness, or complex patterns. These outcomes are not predetermined but emerge from the dynamic interactions among agents [13]:

- Equilibrium Outcomes: Some ABMs converge to a stable state where agent behaviors and system properties do not change over time. For example, Schelling's model of racial segregation and Axelrod's culture model both demonstrate how local interactions can lead to stable, segregated patterns.

- Dynamic and Stochastic Outcomes: Many ABMs do not settle into a fixed point but instead exhibit dynamic behavior such as cycles or fluctuating patterns. These models can produce time series data, such as vote shares in electoral competition models, which vary over time and can exhibit complex, unpredictable dynamics.

- Complex and Emergent Patterns: ABMs are particularly well-suited for exploring complex systems where traditional analytical methods fall short. They can generate rich, emergent patterns that provide insights into phenomena like social norms, collective behavior, and systemic risks. This capability makes them invaluable for studying complex adaptive systems in various domains, including political science, economics, and sociology

2.6 Model Validation and Calibration

Validation and calibration are critical steps in the development of ABMs, ensuring that the model accurately represents the real-world system and produces reliable and meaningful results. Validation involves assessing the model's accuracy and credibility by comparing its outcomes with empirical data or established theories. Calibration, on the other hand, involves adjusting the model's parameters and rules to align its behavior with observed data or known benchmarks.

The validation process begins with defining appropriate metrics for evaluating the model's performance. These metrics can include statistical measures, such as goodness-of-fit tests, as well as qualitative assessments of the model's ability to reproduce key patterns and behaviors observed in the real world. For example, in an ABM of urban traffic,

validation metrics might include the average travel time, traffic congestion patterns, and the distribution of vehicle flows across different routes.

Empirical validation involves comparing the model's outputs with observed data from the real-world system being studied. This comparison can help identify discrepancies between the model and reality, guiding the refinement of model assumptions and parameters. Researchers often use historical data to validate ABMs, ensuring that the model can replicate past behaviors and trends. For instance, in an ABM of financial markets, historical price and trading volume data can be used to validate the model's accuracy in capturing market dynamics.

Calibration is an iterative process that involves adjusting the model's parameters to improve its fit with empirical data. This process can be automated using techniques such as parameter optimization and machine learning, or it can be performed manually through trial and error. Calibration helps to ensure that the model's behavior aligns with known benchmarks and that it produces realistic outcomes under different scenarios. For example, in an ABM of disease spread, parameters such as transmission rates and recovery times can be calibrated using epidemiological data [30].

Robustness and sensitivity analysis are also important components of model validation. Sensitivity analysis involves systematically varying model parameters to assess their impact on the outcomes, helping to identify which parameters are most influential and which are less critical. Robustness analysis examines the model's behavior under different assumptions and scenarios, ensuring that the results are consistent and reliable. These analyses help to build confidence in the model's validity and to identify areas where further refinement may be needed.

2.7 ABMs in Social Simulations

Agent-Based Modeling (ABM) is crucial for exploring and understanding social systems through computational simulations. Social systems involve complex networks of interactions among individuals, institutions, and their environments, encompassing individual behaviors, group dynamics, social norms, and institutional structures. Individuals in these systems often play multiple roles simultaneously, each with its own set of expectations and norms [41].

ABM simulates the actions and interactions of agents-representing individuals or enti-

ties within social systems-to predict and understand complex phenomena. This bottom-up modeling approach allows macro-level phenomena to emerge from micro-level interactions, providing valuable insights into social order, the evolution of norms and institutions, and the dynamics of social change. Social simulations enable the examination of hypothetical scenarios, testing of social behavior theories, and exploration of policy effects without real-world ethical and practical constraints.

However, analyzing ABM results to understand social systems is challenging due to the complexity and dynamism of the models and the systems they represent. ABMs generate vast amounts of data from numerous agent interactions over time, making pattern discernment and conclusion drawing difficult. Emergent phenomena, while valuable, can complicate analysis as outcomes are not always predictable or linear. Interpreting ABM results meaningfully for policy-making or theoretical advancement requires bridging the gap between technical model outputs and social science conceptual frameworks, often necessitating interdisciplinary collaboration.

In addition to ABM, social scientists use various methods to study social systems. Surveys and questionnaires gather large-scale data on individual attitudes and behaviors. Interviews and ethnography provide in-depth qualitative insights into social phenomena. Case studies offer detailed analysis of specific instances, experimental designs establish causal relationships, and statistical analysis and computational techniques, including network analysis and data mining, model large datasets. Content and discourse analysis examine communication patterns and meaning construction. Some of these methods are already employed in ABM studies [33].

3 Introduction to Opinion Dynamics

3.1 Social Networks

Social networks play a crucial role in disseminating information, shaping opinions, and influencing behaviors. They are responsible for spreading news about products, job opportunities, and various social programs. Additionally, they impact decisions related to education and criminal activities, and they affect political opinions and attitudes toward different groups. Given their significant role, it's essential to understand how beliefs and behaviors change over time, how these changes relate to the network's structure, and whether the outcomes are effective. Given the complex forms that social networks often take, it can be difficult for the agents involved to update beliefs properly [29].

In a simplified view, social networks can be broadly understood as the framework of social connections. It encompasses not just the relationships between individuals, but also the connections between organizations, cities, and even countries. This network perspective is invaluable for understanding the structure of complex systems, the spread of trust, and the mechanisms of influence. In the context of opinion dynamics, a social network specifically refers to the connections among agents, represented by a graph where nodes symbolize the agents and edges denote their relationships. Here we introduce three representative complex networks.

The Erdős–Rényi (ER) random network model is characterized by its random distribution of edges among nodes. In this model, the randomness pertains to how connections are made: each pair of nodes is connected with a probability p (where $0 \leq p \leq 1$). For a network with m nodes, this means that every pair of nodes has an independent chance p of being connected. In ER networks, both the clustering coefficient and average path length tend to be low. The clustering coefficient measures the likelihood that two neighbors of a node are also connected to each other, while the average path length is the mean distance between any two nodes in the network.

In a small-world (SW) network, most nodes are not directly connected to each other, but any node can be reached from any other node with just a few connections. This network model captures the phenomenon where people who seem distant are actually relatively close through a short chain of connections. While a SW network features a short average path length, similar to a random network, it also exhibits a significantly

higher clustering coefficient. This means that, unlike in random networks, nodes in a small-world network tend to form tightly-knit groups or clusters.

A scale-free (SF) network is characterized by significant heterogeneity in node connectivity. In such networks, the distribution of connections, or degrees, among nodes is highly uneven: a small number of nodes have a very high number of connections, while the majority have only a few. The nodes with many connections, often referred to as "hubs," play a crucial role in the network's structure and function. This uneven distribution of connections is a fundamental property of scale-free networks, reflecting the inherent imbalance found in many complex systems [57].

3.2 Classical Opinion Dynamics Models

Opinion dynamics is the process of studying the evolution of opinions through the social interaction between a group of agents [57]. In recent decades, various opinion dynamics models with different opinion evolution rules have been proposed. These models can be classified into either discrete or continuous, based on the chosen representation for the opinions. In physics literature, the discrete (especially binary) opinion dynamics model dominates the research due to its apparent analogy with the classic spin model [54]. Other examples are the voter [11, 18] model, the majority rule [25, 26] model and the Sznajd [48] model. Although researches with the binary and discrete models were fruitful, these models have serious limitations. In some situations, it is important to know not only whether two opinions are the same or not, but also how similar or dissimilar they are. Continuous opinion models were then established to deal with these limitations, and have since attracted increasing attention from researchers. These models include the DeGroot [15] model , the FJ [24] model, the DW [14] model and the HK [35] model.

Most opinion dynamics models consist of three fundamental components: the format of opinion expression, the rules governing opinion evolution, and the environment in which opinion dynamics occur. In discrete models, opinions are expressed as binary or categorical values, while environments are typically regular lattices or some other network structure. The opinion evolution mechanism can vary according to the model [54]. Given this setup, agents interact with each other and update their opinions on the same topic.

After each iteration, a stability check is made, comparing all agent's updated opinions with the opinions from the previous iteration. If these values are equal (according to some

threshold), stability is reached and the algorithm stops. If all agents converged to the same final opinion, we call it a *consensus*, otherwise, we call it a *fragmentation* with *n distinct opinion clusters*. In the special case of fragmentation where two clusters are formed and these clusters reflect contrasting viewpoints (according to some measure), most authors call it a *polarization*. When working with continuous models, it is important (and often crucial, see [34]), to consider the numerical stability of the algorithm. As noted above, the notions of consensus and fragmentation will depend on the floating-point arithmetic used to represent real numbers. For most studies, the simple adoption of thresholds can solve this problem.

The next sections provide a brief introduction to some classical continuous models, including the DeGroot [15], and the *Bounded Confidence models*, DW [14] and HW [35]. The goal here is to better contrast and illustrate the contributions of our work, regarding the limitations of these classical opinion dynamics models.

3.3 The DeGroot Model

Consider a finite set A = {1, 2, 3, . . . , n} of agents or nodes that interact according to a social network. The interaction patterns are represented through an $n \times n$ non-negative matrix $\mathbf{T}$, where $T_{ij} > 0$ indicates that i pays attention to j. The matrix $\mathbf{T}$ may be asymmetric, and the interactions can be one-sided, so that $T_{ij} > 0$ while $T_{ji} = 0$. We refer to $\mathbf{T}$ as the interaction matrix. This matrix is stochastic, so that each row sums to one.

At each iteration, agents update their opinions by taking the weighted average of their neighbors opinions, with T_{ij} being the weight or trust that agent i places on the current opinion of agent j in forming his or her opinion for the next iteration. In particular, each agent has a opinion $p_i(t) \in \mathbb{R}$ at time $t \in \{0, 1, 2, \ldots\}$. For simplicity, we take $p_i(t)$ to be a scalar, but it could as well be in $\mathbb{R}^n$ without loss of generality. The vector of all agent's opinions at time t is written $\mathbf{p}_{(t)}$. The updating rule is

$$\mathbf{p}_{(t)} = \mathbf{T}\mathbf{p}_{(t-1)} \tag{1}$$

then

$$\mathbf{p}_{(t)} = \mathbf{T}^t\mathbf{p}_{(0)} \tag{2}$$

Since the interaction matrix $\mathbf{T}$ is stochastic, standard results in Markov chain theory can

be used to state conditions under which the limit

$$\mathbf{p}(\infty) = \lim_{t\to\infty} \mathbf{p}_{(t)} = \lim_{t\to\infty} \mathbf{T}^t \mathbf{p}_{(0)} \tag{3}$$

exists for any initial opinions $\mathbf{p}_{(0)}$. A complete characterization of the convergence properties of Equation (3) can be found in [29]. For completeness, we state as a proposition the special case where consensus is always reached, so that all agents converge to the same opinion.

Proposition. *If the matrix* $\mathbf{T}$ *is strongly-connected and aperiodic, then* $\lim_{t\to\infty} \mathbf{T}^t$ *exists and a consensus is reached.*

3.4 Bounded Confidence models

The bounded confidence (BC) models are an extended version of the DeGroot model, where the weight T_{ij} changes with time. The BC models are based on the following idea: when the difference of opinions between two agents is lower than a given threshold, they will interact, otherwise they will not even bother to discuss.

Let $p_i(t) \in [0, 1]$ be the opinion of agent a_i at time t, and ϵ be the bounded confidence level (which is held constant for the entire simulation). The BC model include two main variants, the DW and HK models, both described in the next subsections.

3.4.1 DW model

Two agents are randomly chosen from the set of agents, and they will determine whether to interact or not according to the bounded confidence level. If $|p_i(t) - p_j(t)| > \epsilon$, agents a_i and a_j will not interact, since their opinions are too far apart according to ϵ. Otherwise, the update rule will be:

$$p_i(t+1) = p_i(t) + \mu(p_j(t) - p_i(t)) \tag{4}$$

$$p_j(t+1) = p_j(t) + \mu(p_i(t) - p_j(t)) \tag{5}$$

where $\mu \in [0, 0.5]$ is the convergence parameter. Depending on the parameters ϵ and μ, the system may reach a consensus, polarization or fragmentation [57].

3.4.2 HK model

Let $T_{ij}(t)$ be the weight that agent a_i gives to agent a_j at time t. This weight can be described as

$$T_{ij}(t) = \begin{cases} \frac{1}{|S_i^t|} & a_j \in S_i^t \\ 0 & a_j \notin S_i^t \end{cases} \tag{6}$$

where $S_i^t = \{a_j || o_i^t - o_j^t| \leq \epsilon\}$ is the confidence set of agent a_i, and $|\cdot|$ denotes the absolute value of a real number and the number of elements for a finite set. Then, the opinion evolution rule is as follows:

$$p_i(t+1) = \sum_{a_j \in S_i^t} T_{ij}(t) p_j(t) \tag{7}$$

In other words, at each round, agent a_i that has current opinion $p_i(t)$ will update its opinion with the simple average of all other agents opinion $a_j(t)$ that are within ϵ of $p_i(t)$. If there exists an ordering $p_{i_1} \leq p_{i_2} \leq \cdots \leq p_{i_m}$ such that two adjacent opinions are within the bounded confidence level ϵ, then the opinion profile $P = p_1, p_2, \ldots, p_m$ is called an ϵ-profile. Hegselmann and Krause [35] argue that the opinion profile will be an ϵ-profile for all times if a consensus is reached for an initial profile. Moreover, two agents will remain separated forever if they split at some point [57].

4 Introduction to Large Language Models (LLMs)

The landscape of natural language processing (NLP) has undergone a profound transformation with the advent of Large Language Models (LLMs). This evolution is rooted in several key advancements that have progressively enhanced the capabilities of language models, enabling them to perform a broad range of tasks with unprecedented accuracy and versatility. This introduction delves into the theoretical foundations and milestones of LLMs, tracing their development from early architectures to contemporary models.

4.1 The Advent of the Transformer Architecture

The concept of the Transformer model, introduced by Vaswani et al. in their groundbreaking paper "Attention is All You Need" (2017), represents a pivotal moment in the evolution of language models [5]. Prior to this, sequential data processing in NLP heavily relied on Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks. While RNNs were effective in capturing temporal dependencies, they struggled with long-range dependencies and parallelization issues. The Transformer model addressed these challenges by utilizing self-attention mechanisms, which allowed for the direct modeling of relationships between all tokens in a sequence, irrespective of their distance from one another.

The self-attention mechanism in the Transformer architecture processes sequences in parallel rather than sequentially, leading to significant improvements in computational efficiency and scalability. This parallelization capability not only reduced training times but also enabled the handling of much larger datasets. Furthermore, the Transformer's ability to capture contextual relationships through attention weights provided a more nuanced understanding of language compared to previous architectures. The introduction of positional encodings in Transformers also addressed the limitation of sequence order, ensuring that the model could retain the position information necessary for understanding the sequence.

Vaswani et al.'s work laid the groundwork for subsequent advancements in language models by demonstrating that an architecture purely based on attention mechanisms could outperform traditional RNN-based models in various NLP tasks. Their work has since become a cornerstone in the field, influencing the design of numerous subsequent

models and serving as a key reference for researchers exploring novel architectures and improvements in NLP.

4.2 Bidirectional Contextual Embeddings: The BERT Architecture

The introduction of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. in 2018 marked another significant leap in the development of language models [16]. Unlike previous models that processed text in a unidirectional manner, BERT utilized a bidirectional approach to capture context from both directions in a sentence. This bidirectional training allowed BERT to develop a more comprehensive understanding of context and meaning, addressing limitations associated with single-direction models.

BERT's training process involves two main stages: pre-training and fine-tuning. During pre-training, BERT is exposed to a large corpus of text to learn general language representations through masked language modeling and next sentence prediction tasks. In the fine-tuning stage, BERT is adapted to specific tasks by leveraging the pre-trained representations. This approach demonstrated significant improvements across a range of benchmarks, including question answering, named entity recognition, and sentiment analysis. The model's ability to understand context more effectively led to a substantial enhancement in performance on these tasks compared to previous models.

The impact of BERT extended beyond its immediate performance gains. It inspired a series of subsequent models and variations, including RoBERTa, ALBERT, and DistilBERT, which built on the core principles of BERT while optimizing various aspects such as efficiency and scalability. The introduction of BERT established a new paradigm in NLP, emphasizing the importance of bidirectional context and transfer learning in achieving state-of-the-art performance.

4.3 The Era of Massive Language Models: GPT-3

The release of GPT-3 (Generative Pre-trained Transformer 3) by Brown et al. in 2020 represented a landmark achievement in the field of LLMs [8]. The GPT-3 model, with its 175 billion parameters, demonstrated the power of scaling up language models to achieve remarkable performance across a diverse range of tasks. The model's capacity to generate

coherent and contextually relevant text in various domains showcased its versatility and the effectiveness of large-scale pre-training.

GPT-3's architecture builds upon its predecessors, including GPT-2, by increasing the model size and training data. One of the key innovations of GPT-3 is its few-shot learning capability, which allows the model to perform tasks with minimal examples. This feature contrasts with traditional models that often require extensive task-specific training data. GPT-3's ability to generalize from a few examples and generate high-quality text has opened up new possibilities for applications in content creation, dialogue systems, and more.

Despite its impressive capabilities, GPT-3 also raised important questions about the ethical implications and practical challenges associated with deploying large-scale models. Issues such as model bias, computational resource requirements, and the potential for misuse have become central topics of discussion in the research community. The release of GPT-3 has catalyzed ongoing debates and research into addressing these challenges while harnessing the potential of LLMs.

4.4 Insights into Scaling and Performance

The study of scaling laws has provided valuable insights into the relationship between model size, training data, and performance. Kaplan et al. (2020) explored these scaling laws in their paper "Scaling Laws for Neural Language Models," which analyzes how increasing model parameters and training data influences performance [40]. Their research demonstrates that larger models generally achieve better performance, though with diminishing returns beyond a certain point. This understanding has guided the development of increasingly larger models and informed decisions about resource allocation and training strategies.

Scaling laws also underscore the importance of data quality and diversity in training large models. Kaplan et al.'s findings highlight that while increasing model size can improve performance, the quality and breadth of the training data play a crucial role in determining the model's effectiveness. As models grow larger, ensuring the inclusion of diverse and representative data becomes increasingly critical to avoid issues such as bias and overfitting.

The insights from scaling laws have influenced the design and training of subsequent

LLMs, including GPT-3 and beyond. Researchers continue to explore ways to optimize scaling strategies, improve data curation practices, and develop methods for efficient training and deployment of large models. Understanding these scaling dynamics remains a key area of research for advancing the field of LLMs.

4.5 Efficient Training and Implementation

As LLMs become larger and more complex, efficient training and implementation have become critical challenges. Narayanan et al. (2020) addressed these challenges in their work on efficient large-scale language model training [44]. They explored techniques for optimizing training processes on distributed GPU clusters, which are essential for managing the computational demands of training massive models. Strategies such as model parallelism, data parallelism, and mixed-precision training are crucial for achieving scalability and efficiency.

By leveraging distributed computing resources and implementing efficient algorithms, researchers can reduce training times and resource consumption while maintaining model performance. These advancements are essential for supporting the ongoing development and deployment of LLMs.

Efficient training methods also contribute to the accessibility and democratization of large language models. By improving the scalability and affordability of training processes, researchers and organizations can more readily experiment with and deploy advanced models. The continued evolution of training techniques will play a significant role in shaping the future of LLMs and their applications.

4.6 The Emergence of Smaller LLMs: Phi-3, LLaMA-3 and Mixture of Experts

As the field of large language models (LLMs) continues to evolve, there has been a notable trend towards the development and deployment of smaller models that demonstrate performance levels comparable to their larger counterparts. This trend is driven by the need for more efficient models that balance performance with computational and resource constraints. Recent advancements, such as Microsoft's Phi-3 and Meta's LLaMA-3, exemplify this shift, offering high performance while maintaining relatively smaller model

sizes.

The emergence of these smaller, high-performing LLMs is reshaping the landscape of artificial intelligence. It is becoming increasingly evident that model size does not necessarily equate to superior performance. This trend opens up new possibilities for innovation and democratization of AI, as smaller models become more accessible to a wider range of developers and organizations, since these models do not need specialized highly-cost GPUs to run.

4.6.1 Phi-3 model

Microsoft's Phi-3 model represents a significant step in optimizing LLMs for both performance and efficiency. With number of parameters ranging from 7B (Phi-3 small) to 14B (Phi-3 medium), Phi-3 builds on the Transformer architecture with several enhancements aimed at reducing model size without compromising effectiveness. The model employs techniques such as parameter sharing, reduced precision training, and efficient attention mechanisms to achieve a balance between performance and resource consumption. Despite its smaller size compared to models like GPT-3.5, Phi-3 has been designed to perform competitively on various NLP benchmarks.

The key advantage of Phi-3 lies in its ability to maintain high levels of performance on tasks such as text generation, comprehension, and question-answering while requiring fewer computational resources. This efficiency makes Phi-3 suitable for deployment in environments with limited hardware capabilities or for applications where response times and cost-effectiveness are critical. Microsoft's approach with Phi-3 underscores the growing trend towards creating smaller, more efficient models that can deliver comparable results to larger, more resource-intensive models [1].

4.6.2 LLama3 model

Meta's LLaMA-3 model represents another noteworthy advancement in the realm of smaller LLMs. With number of parameters in the smaller variants ranging from 8B to 70B, LLaMA-3 is designed to provide high performance across a range of NLP tasks while operating with a more compact architecture. By leveraging innovations in model design and training techniques, LLaMA-3 achieves results that are competitive with larger models like GPT-3.5, despite having fewer parameters.

LLaMA-3's design focuses on optimizing the efficiency of both the training and inference processes. Techniques such as sparse attention mechanisms and advanced optimization algorithms contribute to the model's ability to perform well while keeping resource requirements manageable. The success of LLaMA-3 highlights the potential of smaller models to match or exceed the performance of larger counterparts, especially when tailored for specific applications and optimized for efficiency [36].

4.6.3 Mixture of Experts

One of the most promising architectural innovations in the LLM space is the "Mixture of Experts" (MoE) approach. This technique involves dividing the model into multiple specialized sub-models, or experts, each responsible for handling specific parts of the input data. By routing different parts of the input to the most suitable expert, MoE models can achieve significant improvements in performance and efficiency compared to traditional monolithic architectures [39, 42].

The core idea behind MoE is to leverage the principle of specialization. Each expert can be trained on a specific subset of the data, allowing it to develop a deeper understanding of that particular domain. This specialization leads to better performance and faster inference times, as the model only needs to activate the relevant experts for a given input.

A prime example of MoE in action is the Mixtral model, a variant of the Mistral LLM [38]. Mixtral employs an MoE architecture to enhance its capabilities. By combining multiple expert models, Mixtral demonstrates improved performance on a variety of tasks while maintaining efficiency. This showcases the potential of MoE as a powerful tool for building larger and more complex language models.

The performance of smaller and mixture of expert models is a indicative of a broader trend towards optimizing LLMs for practical deployment. Research and benchmarks have shown that these models can achieve results comparable to larger models like GPT-3.5, particularly in contexts where efficiency and cost are important considerations. Smaller models are often more accessible and feasible for integration into a wider range of applications, from mobile devices to edge computing environments.

Also, the competitive performance of smaller LLMs addresses several challenges associated with larger models, such as high computational costs and environmental impact. By reducing the size of the models and the associated computational requirements, de-

velopers can mitigate some of these concerns while still leveraging advanced language understanding capabilities. This trend towards smaller models reflects a growing emphasis on sustainability and practical utility in AI research and development.

4.6.4 LLMs in Social Simulations

Large Language Models (LLMs) have rapidly gained adoption across various research domains and industries due to their superior language understanding, generation, and translation capabilities. They are utilized in healthcare for patient care and medical research, in finance for market analysis and customer service automation, and in academia for analyzing large datasets, thus accelerating research in social sciences, linguistics, and computer science. LLMs' versatility and sophistication are reshaping industries and research methodologies [12, 7, 52, 49].

In social sciences, LLMs can be effectively applied to social simulations, which are used to model and analyze complex interactions within social systems, such as individual behaviors, group dynamics, social norms, and institutional structures. These simulations aim to understand, predict, or examine hypothetical scenarios within social systems [31].

Agent-Based Modeling is a technique commonly used for social simulations. In the context of ABMs, for most applications, each LLM instance can be viewed as an individual agent that will interact with other agents, usually via raw text messages. Exploring the usage of LLMs as agents in Social Simulations is a very recent research field, with many articles and ideas being published every day, and as such, its full potential has not yet been realized. Even though these studies already provide invaluable experimental results, the conceptual framework is still evolving.

4.7 Comprehensive Reviews and Recent Advances

To stay abreast of ongoing advancements in LLMs, recent comprehensive reviews and surveys offer valuable overviews of the state-of-the-art technologies and emerging trends. These reviews synthesize the latest research, highlight key developments, and provide insights into current and future directions in LLM research. They often cover a wide range of topics, including model architectures, training techniques, applications, and ethical considerations [55, 32].

Recent reviews also address the broader impact of LLMs on various fields, such as

healthcare, education, and industry. By summarizing the latest research and providing contextual analysis, these reviews help researchers and practitioners understand the implications of LLM advancements and guide future research efforts.

For the most up-to-date information, researchers should consult recent publications and reviews in top-tier journals and conferences. These sources provide a comprehensive understanding of the current state of LLMs and offer a roadmap for future research and development. In the following chapter we will review in more detail some of these works, since they directly relate to the central theme of this this thesis: *the use of LLMs to simulate Humans and replicate Human Subject Studies*.

4.8 LLM programming APIs

In this section we review the programming standards and conventions that most LLM models and programming APIs follow.

4.8.1 The messages chat format

The standard format that LLM APIs follow is a JSON structure containing a list of messages. Each message has a `role` and `content` field:

```
messages = [
    {
        "role": "system",
        "content": "You are a helpful AI assistant."
    },
    {
        "role": "user",
        "content": "Why is the sky blue?"
    },
    {
    "role": "assistant",
    "content": "The sky is blue because..."
    },
]
```

Where:

- `role` can be "system" (initial instructions), "user" (user messages), or "assistant" (AI responses).
- `content` contains the actual message text.

Thus, when prompting an LLM, this *messages* parameter is normally sent as part of the request, along with other parameters such as *temperature*, that will affect the model's response.

5 LLMs as Human Emulators: Literature Review

Large Language Models (LLMs), such as GPT-3 and its successors, have demonstrated remarkable capabilities in generating coherent and contextually appropriate text. These models, trained on vast amounts of data, can simulate human-like responses, making them valuable tools in various applications, including social network simulations and synthetic surveying. This chapter reviews some works that explore the potential of LLMs as human emulators, examining their behavior, limitations and possible biases in their generated opinions.

5.1 Multi-Human Simulation and Behavioral Replication

In [3] the authors introduce the concept of *Turing Experiments (TEs)*, a novel method to evaluate the zero-shot simulation capabilities of AI models, specifically large language models (LLMs). TEs aim to replicate human behavior in various scenarios by using text prompts to generate responses from LLMs, thereby providing insight into which human behaviors are captured by these models.

The primary contributions of the work include the proposal of TEs, a new approach to understanding AI capabilities by simulating human subject studies, and the development of a methodology for running TEs. This methodology involves designing prompts and generating responses from LLMs to create a text-based transcript of simulated experiments. Additionally, the researchers conducted four distinct TEs using various GPT models to assess their fidelity in replicating human behaviors.

The study conducted four TEs across different domains. The Ultimatum Game TE investigated fairness and rationality in behavioral economics, with findings indicating that simulation outcomes varied consistently by gender and name, replicating human gender differences. The Garden-Path Sentences TE examined parsing in psycholinguistics, where larger models provided more accurate simulations of human parsing difficulties compared to smaller models. The Milgram Shock Experiment TE studied obedience to authority in social psychology, replicating the diminishing obedience observed in the original Milgram experiment, with a notable spike at a critical shock level. The Wisdom of Crowds TE focused on collective intelligence, and contrary to expectations, larger models did not outperform smaller ones. Instead, they exhibited a hyper-accuracy distortion, providing

inhumanly precise answers.

The study acknowledges several limitations. There is a concern that LLMs may reproduce specific sentences and descriptions from their training data rather than generating novel responses. Ensuring the prompts accurately simulate experimental conditions without leading to "p-hacking" requires careful design and validation. While TEs can predate and inform costly human subject studies, they may not fully capture the complexity and variability of real human behavior. Additionally, the use of TEs must consider potential biases and ethical implications of simulating human subjects, especially regarding sensitive topics.

The study demonstrates that TEs can be a valuable tool for evaluating AI models, offering insights into their strengths and weaknesses in simulating human behavior. However, the approach requires careful consideration of training data, prompt design, and ethical implications.

5.2 Opinion Dynamics and LLMs: Investigating Principles and Mechanisms

The study on opinion dynamics utilizing large language models (LLMs) explores how these models can simulate and analyze the spread and evolution of opinions within a population. In [10] the primary focus is on understanding the impact of initial opinion distributions and the nature of the discussion subjects on the final opinion distributions. This study is rooted in the broader context of opinion dynamics and learning within social networks.

The authores investigate how different biases and mechanisms inherent in LLMs influence the formation and propagation of opinions. By simulating various scenarios, the study examines how LLM agents form and change their opinions based on initial conditions and interaction rules. This investigation is pivotal for applications where understanding collective human behavior is essential, such as in social media analysis, marketing, and political forecasting.

Two primary experimental setups are employed: the *FreeForm* case and the *ClosedForm* case. In the FreeForm case, LLM agents freely express their opinions, which allows for a wide range of opinion dynamics to be observed. In contrast, the ClosedForm case restricts agents to choose their opinions from a predefined list of options, providing a more

controlled environment to study opinion shifts.

The study finds that the final opinion distribution significantly depends on the initial distribution. For instance, if the initial distribution is polarized, the final distribution tends to remain polarized unless there is a strong bias towards consensus. Also, agents with memory of their past opinions show a different opinion dynamics compared to memoryless agents. Memory enables agents to maintain consistency with their previous opinions, reducing the impact of spontaneous biases like safety bias. This consistency leads to more stable opinion distributions, particularly when agents are surrounded by like-minded peers.

A bias towards equity-consensus is identified, where agents tend to align their opinions with the majority. This bias is less pronounced when agents have memory, as they exhibit more resistance to changing their opinions to maintain past consistency. In both FreeForm and ClosedForm cases, the study observes polarization under certain conditions. However, the presence of memory and the nature of interaction (e.g., reasons for funding) can shift opinions significantly, leading to either consensus or continued polarization.

The study acknowledges several limitations. Firstly, the simulations are based on idealized models of opinion dynamics and may not fully capture the complexities of real-world social interactions. The reliance on LLMs also introduces potential biases inherent in the training data and algorithms of these models. Additionally, the experimental setups may oversimplify the nuances of opinion formation and change, as real human interactions are influenced by a multitude of factors not accounted for in the study. The research provides valuable insights into the mechanisms of opinion dynamics within populations of LLM agents. It highlights the significant role of initial opinion distributions, memory effects, and biases towards consensus in shaping the final opinion landscape.

5.3 Using Language Models to Simulate Human Samples

In [4] the authors investigate the use of GPT-3 to simulate responses from diverse human sub-populations in social science research. The motivation behind the study is the recognition that while AI tools often replicate biases from their training data, these biases are not monolithic and can be fine-tuned to accurately emulate various human demographic groups. The researchers introduce the concept of *algorithmic fidelity*, which is the ability of a language model to replicate the nuanced relationships between ideas,

attitudes, and socio-cultural contexts of specific human sub-populations. They propose four criteria for assessing this fidelity:

1. Social Science Turing Test: Generated responses must be indistinguishable from human texts.

2. Backward Continuity: Responses must be consistent with the socio-demographic context provided.

3. Forward Continuity: Responses must proceed naturally from the provided context.

4. Pattern Correspondence: Responses must reflect the underlying patterns of relationships observed in human data.

To demonstrate their approach, the authors conducted three experiments using GPT-3 conditioned on thousands of socio-demographic backstories from actual survey participants.

In the first experiment, GPT-3 was tasked with generating free-form text describing outgroup partisans. The generated descriptions closely mirrored those produced by real humans, demonstrating high algorithmic fidelity. The second experiment focused on predicting voting behavior based on demographic backstories, where the model's output exhibited patterns consistent with actual human data. The third experiment is of particular interest for this thesis, and will be described in more detail.

The third study involved GPT-3 responding to closed-ended survey questions. For this, the researchers introduced a methodology called silicon sampling to correct skewed marginal statistics in the language model's training data. This experiment was designed to assess the model's ability to generate responses that mimic the complex correlations between demographics, attitudes, and behaviors typically observed in human survey data. The researchers used data from the American National Election Studies (ANES), which is a comprehensive, long-term data collection project that provides a rich source of information on American public opinion, electoral behavior, and demographic variables. The ANES dataset includes detailed socio-demographic profiles and responses to various survey questions related to political attitudes, voting behavior, and social issues.

A specific subset of the ANES dataset was selected, focusing on responses to specific closed-ended survey questions. This subset included information on respondents demographics (such as age, gender, race, education, and income) and their answers to questions

about political attitudes and behaviors. The selected demographic profiles were used to condition GPT-3. This involved creating prompts that included both the demographic information and the survey questions. For example, a prompt might include a detailed socio-demographic backstory followed by a survey question like "Do you favor an increase, decrease, or no change in government spending to help people pay for health insurance?", GPT-3 was then tasked with generating responses to these survey questions based on the provided demographic profiles, in other words, the LLM was asked to roleplay that specific person. The model's outputs were collected and analyzed to see how well they matched the patterns observed in the actual ANES responses.

The researchers evaluated the generated responses using the criteria for algorithmic fidelity. They looked at the consistency of the model's answers with the demographic context (Backward Continuity), the natural progression of responses (Forward Continuity), and the overall patterns of relationships between demographics and survey responses (Pattern Correspondence). The results showed that GPT-3's generated responses exhibited the same complex correlations between demographics, attitudes, and behaviors that are observed in real human survey data. For instance, the model could reflect how different demographic groups tend to align with certain political attitudes or voting behaviors, demonstrating its ability to accurately simulate the nuanced interplay of factors influencing human opinions and decisions.

The study shows that GPT-3 contains a striking degree of algorithmic fidelity within the realm of public opinion in the United States, being capable of replicating the viewpoints of demographically varied sub-populations within the U.S. Additionally, the ethical implications of using AI to simulate human behavior, such as potential misuse or over-reliance on simulated data, need careful consideration. Models with such fidelity, when coupled with future computational and methodological advances, pose a significant risk of being used to engineer social manipulation, disinformation campaigns, financial fraud, and so forth.

5.4 Related Works

Due to the remarkable performance of LLMs in natural language tasks, there has been a surge of interest in using these models to simulate and understand human opinion dynamics. The limitations of classical ABMs have paved the way for LLMs to revolutionize

this field, offering unprecedented opportunities for research and exploration.

In [46], they address the critical question of whose opinions LLMs reflect when used in open-ended applications. They propose a quantitative framework to evaluate the alignment of LLM opinions with those of various U.S. demographic groups, using a new dataset named OpinionQA. Their analysis reveals substantial misalignment between LLM-generated opinions and those of U.S. demographic groups. Notably, this misalignment persists even when LLMs are explicitly steered towards specific demographic profiles. In other words, LLM opinions tend to align more closely with the viewpoints of specific demographic groups. This results in a bias towards dominant cultural and demographic perspectives. Also, human feedback (HF) fine-tuning [45] that is intended to make models more human-aligned, seems to only amplify this misalignment.

In [9], the authors simulate a social-network where all LLM-based agents are fully connected. Each agent role-plays a different persona that is seeded with an initial opinion on a given topic (all topics explored have a well-established scientific ground truth). At each iteration, two agents are randomly paired and they can read each other's opinions before updating their own. Their findings reveal a strong inherent bias in LLM agents towards producing accurate information, leading simulated agents to reach a consensus with scientific reality. Again, this bias limits their utility for understanding resistance to consensus views on issues like climate change.

6 Experimental Setup

The **American National Election Studies** (ANES) is a renowned research organization in the United States, known for its long-standing contribution to the study of political behavior and public opinion. Founded in 1948, ANES conducts comprehensive surveys during U.S. national elections, capturing a wide array of data related to voter behavior, political attitudes, and public opinion on key issues. These surveys are conducted both before and after elections, enabling the study of political dynamics and changes over time. The richness of ANES datasets makes them invaluable for research into the factors that shape voter decisions and broader political trends in the U.S.

ANES is a collaborative effort, primarily supported by the National Science Foundation (NSF) and jointly managed by the University of Michigan's Institute for Social Research and Stanford University. The organization's commitment to methodological rigor ensures the quality and reliability of its data, making it one of the most trusted sources in the field of political science. Researchers and scholars across various disciplines rely on ANES data to analyze shifts in political behavior, track voter attitudes over decades, and understand the effects of socioeconomic and cultural factors on electoral outcomes.

The experimental part of this work will make use of the ANES 2020 dataset, consisting of interviews with respondents between the dates of August 18, 2020, and November 3, 2020. Our study extends the experimental framework established by [4]. While their research focused on exploring the correlation between real-world human answers and their simulated LLM answers, this work builds on their methodology by exploring different aspects of the problem.

First, we primarily make use of *small* LLMs, with parameter sizes ranging from 7B to 14B. For comparison, state-of-the art LLMs are also used. For example, as of October 2024, the official number of parameters for OpenAI's GPT-4o model is not disclosed, but it is estimated to be above 1 trillion, based on different sources [53]. Secondly, we use a different set of variables in our simulation. Our target variables were chosen with two ideas in mind: first was to avoid using highly correlated variables together (like *political party* and *ideology*), and second, we wanted to extend the variables domain beyond demographic data, to also include social, behavioral, or cultural variables. Third, we pay special attention to the results of *uncensored* models, i.e., models that either were not human-aligned [45], or that underwent another trainning process to try and revert

this alignment. Finally, we develop a custom *Feature Importance* algorithm, to determine which variables are the most important when emulating human respondents using LLMs.

6.1 ANES 2020 dataset

The ANES 2020 dataset contemplates interviews with respondents between August 18, 2020, and November 3, 2020, being the primary data source for comprehending the American public opinion. The dataset includes demographic, social and behavioral information, including responses to presidential votes and political surveys. The dataset also features re-interviews with 2016 ANES respondents, and has samples from the 2020 pre-election and post-election periods.

In this study, we focus exclusively on respondents from the 2020 pre-election sample. To begin our data selection, we exclude the ANES 2016 respondents as well as post-election samples. Next, we define two distinct groups of variables. The first group we call *backstory variables*, consisting of: *Age, Gender, Children, Race, Interested in politics, Ideology, Education, Occupation, Has health insurance, Income, Church, City or rural, Science, Trust media.* The second group we call *topic variables*, which comprises: *Race diversity, Gender role, Current Economy, Drug addiction, Climate change, Gay marriage, Refugee allowing, Health insurance, Gun regulation, Income inequality.* The backstory variables will compose the "personal profile" of our "silicon" or "virtual" person. The topic variables are the questions that this virtual person or silicon sample will answer when roleplayed by the LLM. It is important to notice that in the ANES survey all variables are formatted as question-answers, we adopt these two categories to make our methodology clearer. Tables 1 and 2 explains each of these variables in detail. In the backstory variables table we omit the questions in order to save space, as they are all neutral and direct.

So far our compiled dataset consists of 5441 samples and 24 variables of interest, all respondents from the 2020 pre-election period. In addition to the answer choices listed in Tables 1 and 2, respondents could also choose to refrain from answering a question. Since those unanswered questions are not relevant to our study, we simply exclude them from the dataset, making it now 3990 samples across 24 features. From now on we will denote this dataset simply as "full dataset" - all experiments in our work will actually use a *stratified sample* of this dataset. The sampling methodology is explained in the next

section. The ANES datasets and their corresponding manuals are publicly available and can be accessed at `https://electionstudies.org/` after a simple signup process.

Table 1: Backstory variables

Variable	**Choices**
Age	*(free form value)*
Gender	1. Male 2. Female
Children	0. No children 1. One child 2. Two children 3. Three children 4. Four or more children
Race	1. White 2. Black 3. Hispanic 4. Asian 5. Native American 6. Mixed
Interested in politics	1. Very interested 2. Somewhat interested 3. Not very interested 4. Not at all interested
Ideology	1. Extremely liberal 2. Liberal 3. Slightly liberal 4. Moderate 5. Slightly conservative 6. Conservative 7. Extremely conservative
Education	1. Less than high school credential 2. High school credential 3. Some post-high school, no bachelor's degree 4. Bachelor's degree 5. Graduate degree
Occupation	1. For-profit company or organization 2. Non-profit organization 3. Local government 4. State government 5. Military 6. Federal government, as a civilian employee 7. Owner of non-incorporated business 8. Owner of incorporated business 9. for-profit family business
Has health insurance	1. Yes 2. No
Income (all family)	1 Under $9,999 2. $10,000-14,999 3. $15,000-19,999 ... 15. $80,000-89,999 16. $90,000-99,999 17. $100,000-109,999 ... 21. $175,000-249,999 22. $250,000 or more
Church (attend?)	1. Yes 2. No
City or rural	1. City person 2. Suburb person 3. Small-town person 4. Country person 5. Neither a city nor rural person
Science (need experts?)	1. No 2. A little 3. A moderate amount 4. A lot 5. A great deal
Trust Media	1. No 2. A little 3. A moderate amount 4. A lot 5. A great deal

Table 2: Topic variables

Variable	**Question**	**Choices**
Race diversity	Does the increasing number of people of many different races and ethnic groups in the United States make this country a better place to live, a worse place to live, or does it make no difference?	1. Better 2. Worse 3. Makes no difference
Gender role	Do you think it is better, worse, or makes no difference for the family as a whole if the man works outside the home and the woman takes care of the home and family?	1. Better 2. Worse 3. Makes no difference
Current Economy*	What do you think about the state of the economy these days in the United States?	1. Good 2. Neither good nor bad 3. Bad
Drug addiction	Do you think the federal government should be doing more about the opioid drug addiction issue, should be doing less, or is it currently doing the right amount?	1. Should be doing more 2. Should be doing less 3. Is doing the right amount
Climate change*	How much, if at all, do you think climate change is currently affecting severe weather events or temperature patterns in the United States?	1. Not at all 2. A little 3. A lot
Gay marriage	Which comes closest to your view regarding gay and lesbian couples?	1. They should be allowed to legally marry 2. They should be allowed to form civil unions but not legally marry 3. There should be no legal recognition of gay or lesbian couples relationship
Refugee allowing	Do you favor, oppose, or neither favor nor oppose allowing refugees who are fleeing war, persecution, or natural disasters in other countries to come to live in the U.S.?	1. Favor 2. Oppose 3. Neither favor nor oppose
Health insurance	Do you favor an increase, decrease, or no change in government spending to help people pay for health insurance when people cannot pay for it all themselves?	1. Increase 2. Decrease 3. No change
Gun regulation	Do you think the federal government should make it more difficult for people to buy a gun than it is now, make it easier for people to buy a gun, or keep these rules about the same as they are now?	1. More difficult 2. Easier 3. Keep these rules about the same
Income inequality	Do you favor, oppose, or neither favor nor oppose the government trying to reduce the difference in incomes between the richest and poorest households?	1. Favor 2. Oppose 3. Neither favor nor oppose

6.2 Sampling Methodology

The full dataset, as outlined in the previous section, consists of 3990 samples and 24 features, grouped into 14 *backstory* variables and 10 *topic* variables.

Running LLMs is both time-consuming and costly. To optimize resources and streamline the analysis process, we opted to take a stratified sample of the dataset, based on a subset of the backstory variables, since these are the "profile generator" or *conditioning* variables. To ensure our sample accurately reflects the demographic and political diversity of our target population, we chose the following *strata variables*: *Age, Gender, Race, Education, Income* and *Ideology*. To show that this is indeed an effective choice of stratification variables, we use the Cramér's V statistic.

Cramér's V is a statistical measure used to assess the strength of association between two nominal categorical variables, with values ranging from 0 (indicating no association) to 1 (indicating a perfect association). To evaluate stratification effectiveness, for each strata varaible, we can calculate Cramér's V using a contingency table where columns represent dataset membership (original vs. sample) and rows represent the categories of our variable. A low Cramér's V indicates that knowing whether a data point comes from the original or sample dataset tells us very little about which category it belongs to - meaning the proportions are similar between datasets. This suggests successful stratification because the sample maintained the original distribution of the variable. Conversely, a high Cramér's V would indicate that the distributions differ significantly between the original and sample datasets, suggesting the stratification might not have preserved the original proportions well.

We apply the sampling idea discussed above using a sample fraction of 0.3. Thus, from the full dataset consisting of 3990 samples, our stratified sample dataset will consist of approximately 1200 samples, depending on the random seed. This is the final dataset used in all experiments. The association results are presented in Table 3, with the *Similarity ratio* column calculated by

$$\text{similarity\_ratio} = 1 - \frac{1}{2}\sum_{i=1}^{k} |p_i^{\text{orig}} - p_i^{\text{sample}}| \tag{8}$$

where p_i^{orig} represents the proportions of a given variable in the original dataset, while p_i^{sample} represents the proportions of the same variable in the stratified sample. The

absolute difference between these proportions is summed across all k categories, and this sum is divided by 2 to ensure the value ranges between 0 and 1. The similarity ratio closer to 1 indicates that the stratified sample closely matches the original dataset. Cramér's V association values can be interpreted using the following thresholds, commonly referenced in literature[23]. The results show that our final sample successfully maintains the original distribution of the variables. Figures 1 to 6 show the distribution of categories for all strata variables, for both the original dataset and the stratified sample. The values are coded for spacing reasons (except for age), but the meaning of each value for a specific variable can be found in Table 1.

Cramér's V Value	**Strength of Association**
$V < 0.1$	Negligible
$0.1 \leq V < 0.2$	Weak
$0.2 \leq V < 0.3$	Moderate
$V \geq 0.3$	Strong

Table 3: Stratification association results

Variable	**Cramér's V**	**Similarity ratio**	**Association**
Age	0.090	92.2%	Negligible
Gender	0.005	99.3%	Negligible
Race	0.150	87.3%	Weak
Education	0.045	96.9%	Negligible
Income	0.065	94.4%	Negligible
Ideology	0.038	97.2%	Negligible

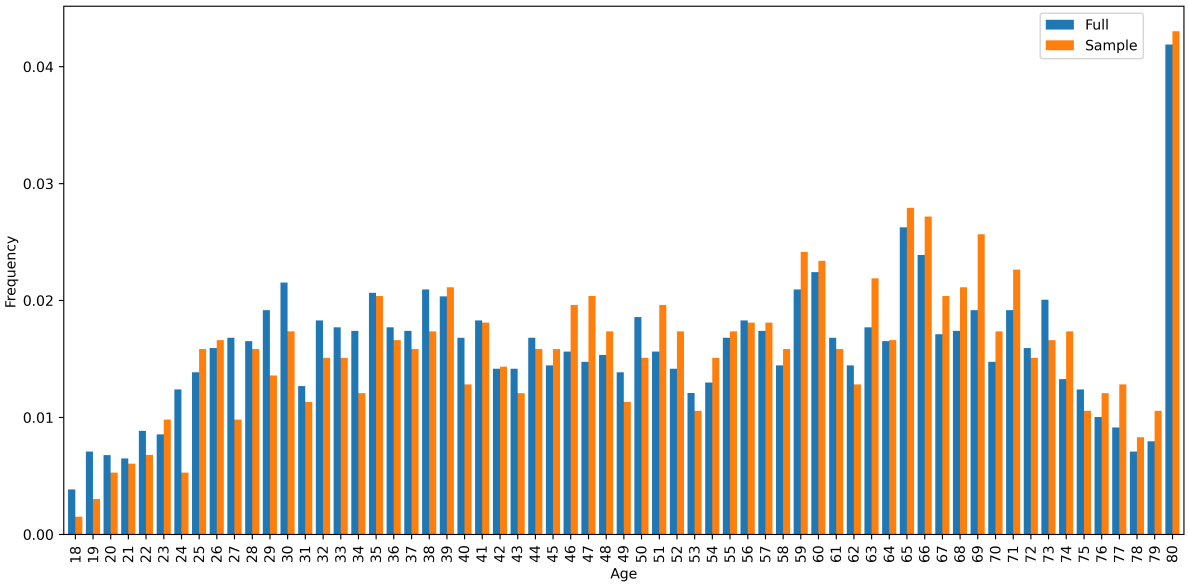

Figure 1: Age stratified sample comparison.

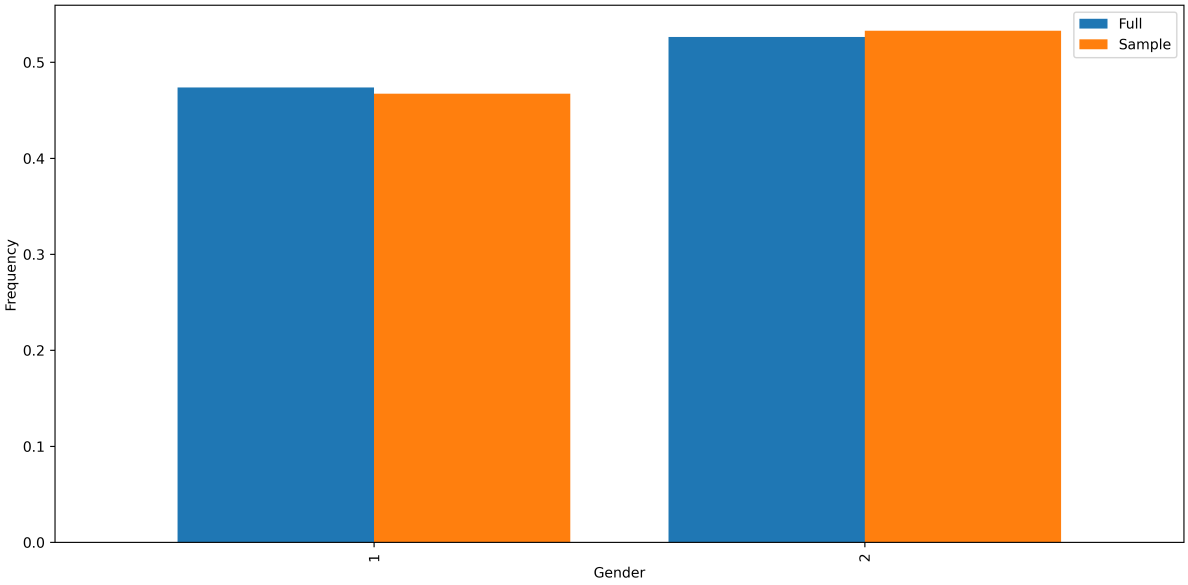

Figure 2: Gender stratified sample comparison.

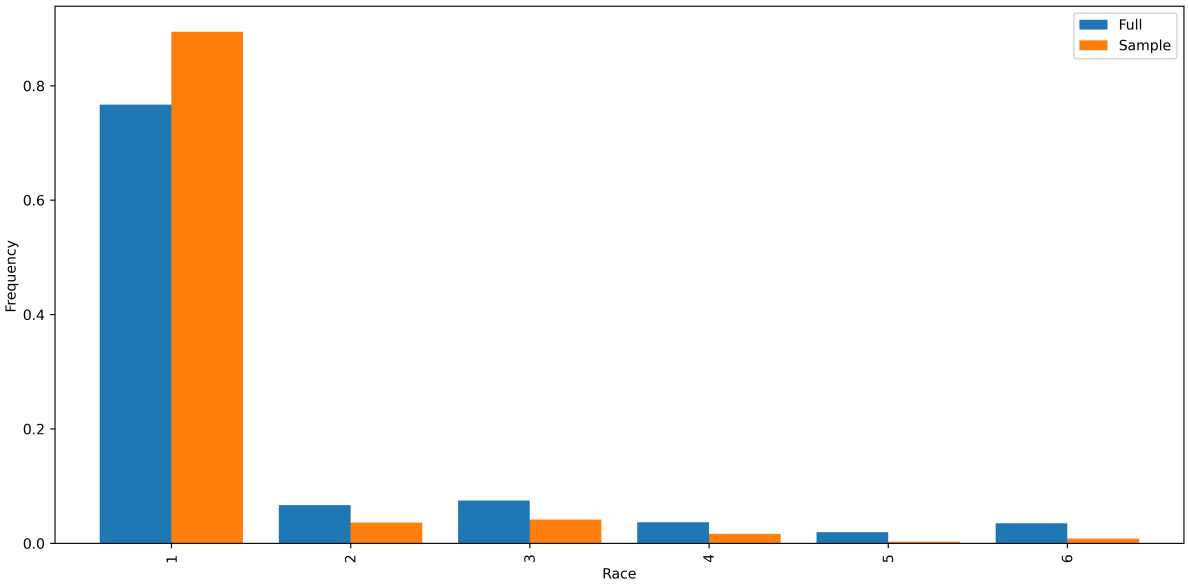

Figure 3: Race stratified sample comparison.

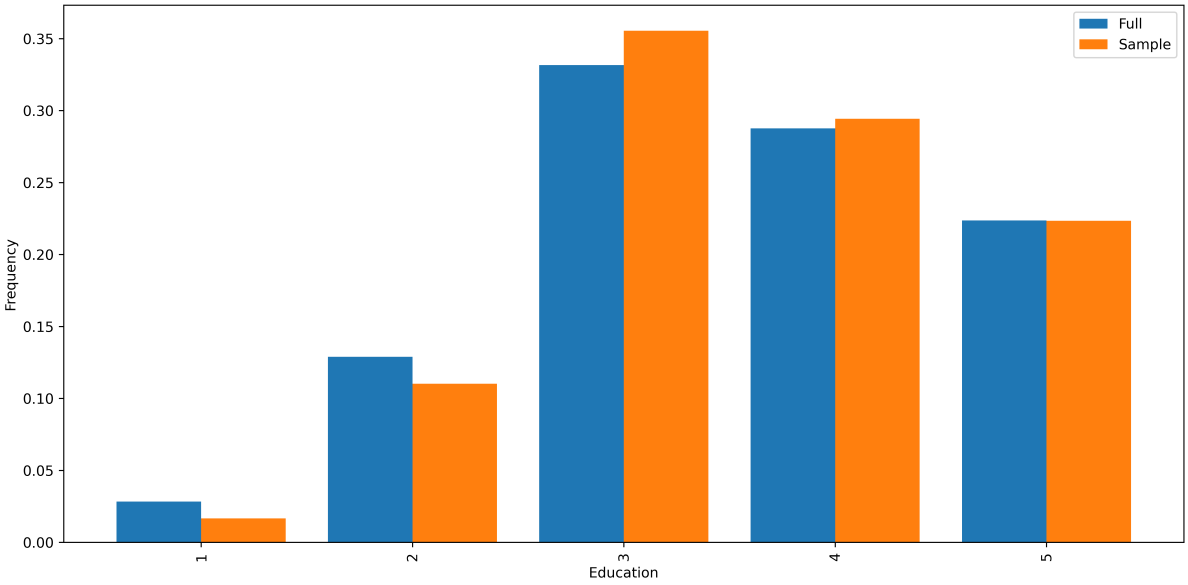

Figure 4: Education stratified sample comparison.

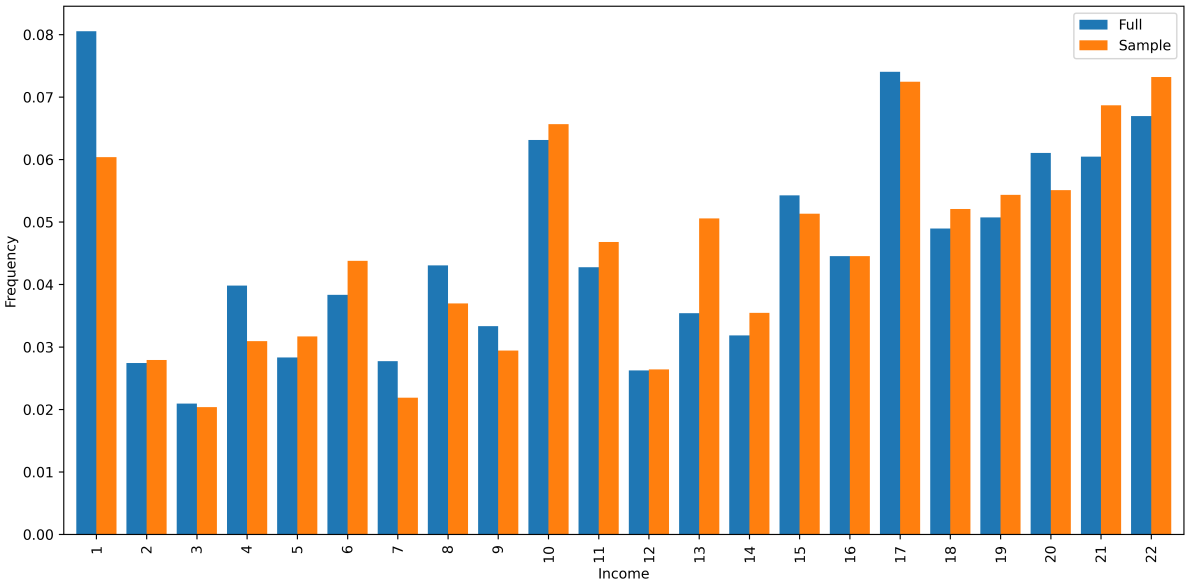

Figure 5: Income stratified sample comparison.

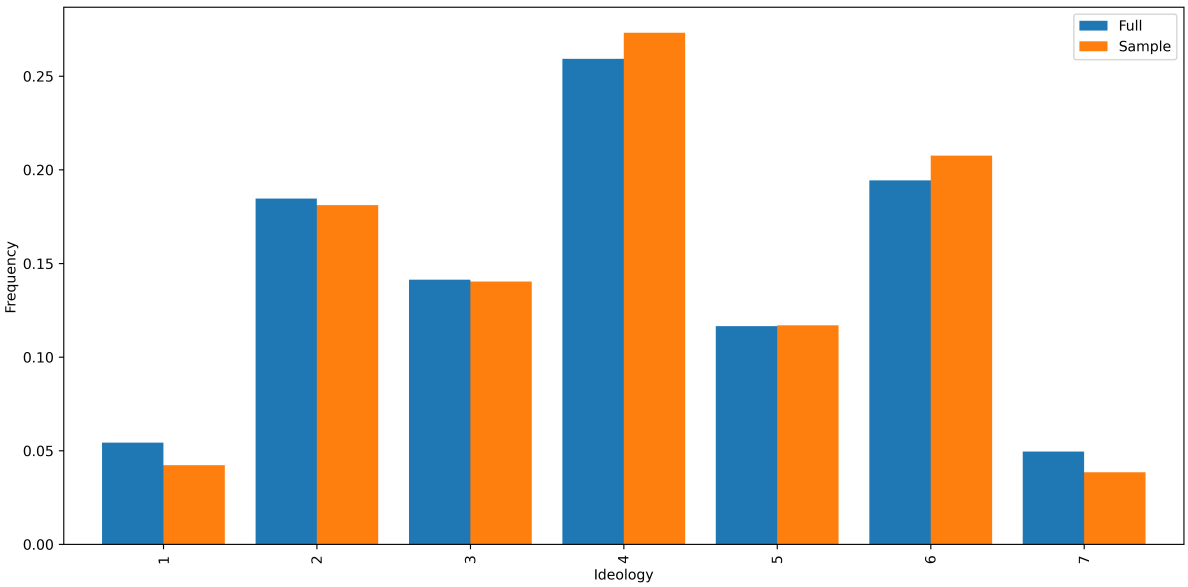

Figure 6: Ideology stratified sample comparison.

6.3 Silicon Sampling Algorithm

The central idea of our experiments can be described in the following pseudocode:

Algorithm 1 Silicon Sampling

```
function SILICON_SAMPLING_TOPIC(dataset, topic, model, temperature)
    llm_answers := []
    respondent_answers := []
    today := "The date is November 3, 2020."
    for each respondent in dataset.rows do
        backstory := create_backstory(respondent.backstory_variables)
        sys_prompt :=
          "Roleplay the person below. {today}
          When questioned, answer just with the option number and nothing more.
          {backstory}"
        llm := instantiate LLM(model)
        llm.setSystemPrompt(sys_prompt)
        llm.setTemperature(temperature)
        user_prompt := get_question_choices(topic)
        llm_text_answer := llm.ask(user_prompt)
        llm_answer := extract_choice(llm_text_answer)
        respondent_answer := get_respondent_answer(respondent, topic)
        llm_answers.append(llm_answer)
        respondent_answers.append(respondent_answer)
    end for
    return llm_answers, respondent_answers
end function
```

The algorithm aboves defines a single experiment for a given dataset, topic, model and temperature. Since each row of our sample dataset corresponds to a real respondent of the ANES 2020 survey, we simply iterate over all respondents and create its backstory. Then we condition the LLM by setting its *system prompt* to this backstory. Finally, we fetch the question and answer choices for the given topic (as given in Table 2) and ask the LLM to answer that question. After getting the LLM textual answer, we perform an additional step to extract the option number from this textual answer. This step is necessary, because even though we clearly instructed the LLM to answer directly with the option number, it may not follow this instruction and answer something like "My answer is option 3.", instead of just "3". Finally, we get the real world answer given by the respondent and store both answers in their corresponding list, for later accuracy measurements. The algorithm returns the list of answers given by each respondent and its LLM conditioned silicon version.

6.4 Silicon Sampling examples

In this section we provide some examples to better illustrate what exactly is feed into the LLM. Each example below is from a different randomly chosen respondent. The topic in question can be seen in the table caption.

Example 1 - Climate change

System prompt: *Roleplay the person below. The date is November 3, 2020. When questioned, answer just with the option number and nothing more.*

You are 23 years old. You are a woman. You have no children. Racially, you are White. You are somewhat interested in politics. Ideologically, you are extremely liberal.You have some post-high school, no bachelor's degree. Your employment is best described as working for a for-profit company or organization. In the moment you do have health insurance. Your total family income is $70,000-74,999. You do not attend church. You consider yourself a city person. You believe people need a great deal of help from experts to understand science. You have little trust in the media when it comes to reporting the news accurately and fairly.

User prompt: *Question: How much, if at all, do you think climate change is currently affecting severe weather events or temperature patterns in the United States?*

Answer choices:

1. Not at all
2. A little
3. A lot

My answer is

Example 2 - Health insurance

System prompt: *Roleplay the person below. The date is November 3, 2020. When questioned, answer just with the option number and nothing more.*
You are 27 years old. You are a man. You have no children. Racially, you are Black. You are somewhat interested in politics. Ideologically, you are moderate. You have some post-high school, no bachelor's degree. Your employment is best described as working for a for-profit company or organization. In the moment you do have health insurance. Your total family income is $45,000-49,999. You attend church. You consider yourself a country person. You believe people need a lot of help from experts to understand science. You have moderate trust in the media when it comes to reporting the news accurately and fairly.

User prompt: *Do you favor an increase, decrease, or no change in government spending to help people pay for health insurance when people cannot pay for it all themselves?*

Answer choices:

1. Increase
2. Decrease
3. No change

My answer is

Example 3 - Gun regulation

System prompt: *Roleplay the person below. The date is November 3, 2020. When questioned, answer just with the option number and nothing more.*

You are 32 years old. You are a woman. You have no children. Racially, you are White. You are somewhat interested in politics. Ideologically, you are extremely liberal. You have a Graduate degree. Your employment is best described as working for a non-profit organization. In the moment you do have health insurance. Your total family income is $80,000-89,999. You do not attend church. You consider yourself a city person. You believe people need a little of help from experts to understand science. You have moderate trust in the media when it comes to reporting the news accurately and fairly.

User prompt: *Do you think the federal government should make it more difficult for people to buy a gun than it is now, make it easier for people to buy a gun, or keep these rules about the same as they are now?*

Answer choices:

1. More difficult
2. Easier
3. Keep these rules about the same

My answer is

6.5 Overview of Selected LLMs

As discussed previously, this work focuses on smaller LLMs that can be run on consumer-grade GPUs. For this purpose, we rely on *Ollama*, an open-source project that enables users to operate LLMs on local machines. Ollama provides a REST API endpoint for developers to interact with these models, as well as a client tool to download the latest versions of supported models. The project can be found at `https://ollama.com/`.

The following models were selected for our experiments:

- **Qwen2-7B**: Qwen2 is a new series of large language models from the Alibaba group. It was trained on data in 29 languages, including English and Chinese. It is available in 4 parameter sizes: 0.5B, 1.5B, 7B, 72B. [56].

- **Llama3-8B**: Meta Llama 3, a family of models developed by Meta, available in both 8B and 70B parameter sizes (pre-trained or instruction-tuned). Llama 3 instruction-tuned models are fine-tuned and optimized for chat use cases and outperform many of the available open-source chat models on common benchmarks [17].

- **Gemma2-9B**: Google Gemma 2 is a high-performing and efficient model available in three sizes: 2B, 9B, and 27B. Featuring a brand new architecture designed for class leading performance and efficiency, both the 9B and 27B versions delivers performance surpassing models more than twice its size in benchmarks [50].

- **Solar-uncensored-10B**: SOLAR-10.7B is an LLM with 10.7 billion parameters, demonstrating incredible performance in various natural language processing tasks. In this work we use an uncensored version, trained using the Toxic-DPO v0.1 dataset to uncensor the model [58].

- **Mistral-nemo-12B**: Mistral NeMo is a 12B model built in collaboration with NVIDIA. Mistral NeMo offers a large context window of up to 128k tokens. Its reasoning, world knowledge, and coding accuracy are state-of-the-art in its size category. As it relies on standard architecture, Mistral NeMo is easy to use and a drop-in replacement in any system using Mistral 7B [37].

- **Phi3-14B**: Phi-3 is a family of open AI models developed by Microsoft. It is available in 2 parameter sizes: 3.8B and 14B. The Phi-3 family of models training

data includes a wide variety of sources, and is a combination of rigorously filtered public documents, selected educational materials and newly generated "textbook-like" synthetic data [1].

There are several reasons to justify this particular model selection. First, these are state-of-the-art models, trained and curated by leading technology companies. Second, each model differs slightly in architecture and training data (see referenced papers), with each reflecting substantial investment in research and innovation. Third, while the models are small, they have varying parameter sizes, ranging from 7B to 14B, which naturally leads to the question: "will larger models perform better?" - a question we aim to answer. Finally, we also include an uncensored model, which raises interesting questions regarding its relative performance and the factors that influence it, which we investigate as well.

7 Results and Insights

All experiments involving these local LLMs were conducted on an NVIDIA GeForce RTX 3060 GPU with 12GB of VRAM. Additionally, all downloaded models were chosen with the same quantization level, specifically the Q6_K quantization.

7.1 Data Generation and Algorithm Parameters

We denote by **experiment** a single run of Algorithm 1. As explained in Section 6.2, we use the same stratified sample dataset for all silicon sampling experiments. The *temperature* parameter is also kept constant for all experiments, with a value of $temperature = 0.7$. The temperature parameter is critical in the sense that it influences the balance between predictability and creativity in the model's generated response. We have two other important parameters that are held constant across all experiments: the *permutation order* of the backstory variables composing the system prompt, and the use of a *third-person* description. We plan to further explore the impact of these parameters in future experiments. Thus, we generate all experimental data by simply iterating over all topics and selected models and running Algorithm 1. Since we have selected 6 models and 10 topics, we generate a total of 60 output files. Each file corresponds to a particular topic and LLM combination and holds the two lists returned by the algorithm: one represents the real respondents answers for that topic, the other is the LLM answers for the same topic. In the next sections we explore the main results that can be derived from this data.

7.2 Result 1: Response Accuracy Analysis

Here we present the accuracy of the LLM responses for all topics and models, by calculating the proportion of matching answers between the LLM and the real respondents. We also display the average accuracy across models and topics.

Table 4: Silicon Sample accuracy

Topic	Qwen2	Llama3	Gemma2	SolarU	MistralN	Phi3	Avg.
Race diversity	0.543	0.395	0.408	0.468	0.594	0.615	0.504
Gender role	0.682	0.485	0.542	0.677	0.651	0.575	0.602
Current Economy	0.382	0.456	0.356	0.349	0.353	0.297	0.366
Drug addiction	0.456	0.476	0.666	0.646	0.72	0.72	0.614
Climate change	0.586	0.418	0.401	0.625	0.54	0.567	0.523
Gay marriage	0.686	0.617	0.681	0.668	0.717	0.717	0.681
Refugee allowing	0.459	0.498	0.44	0.498	0.536	0.579	0.502
Health insurance	0.63	0.53	0.524	0.566	0.597	0.58	0.571
Gun regulation	0.621	0.46	0.459	0.517	0.658	0.545	0.543
Income inequality	0.508	0.523	0.599	0.644	0.636	0.655	0.594
Avg.	0.555	0.486	0.508	0.566	0.6	0.585	

For visual clarity, Figures 7 and 8 showcase the same results from different viewpoints.

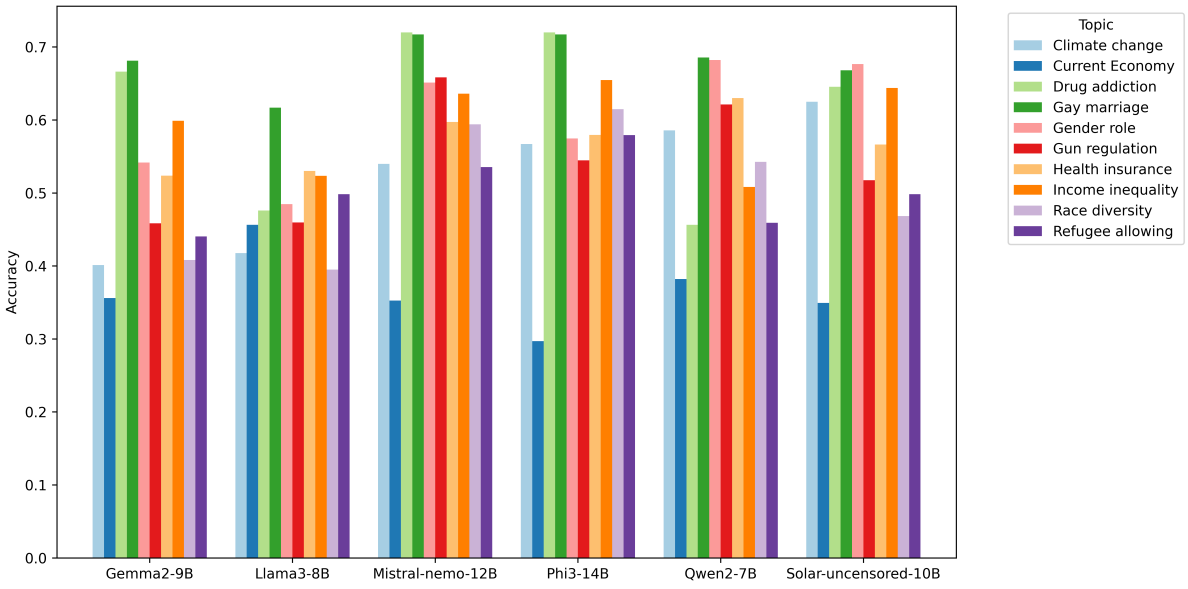

Figure 7: Accuracy grouped by model.

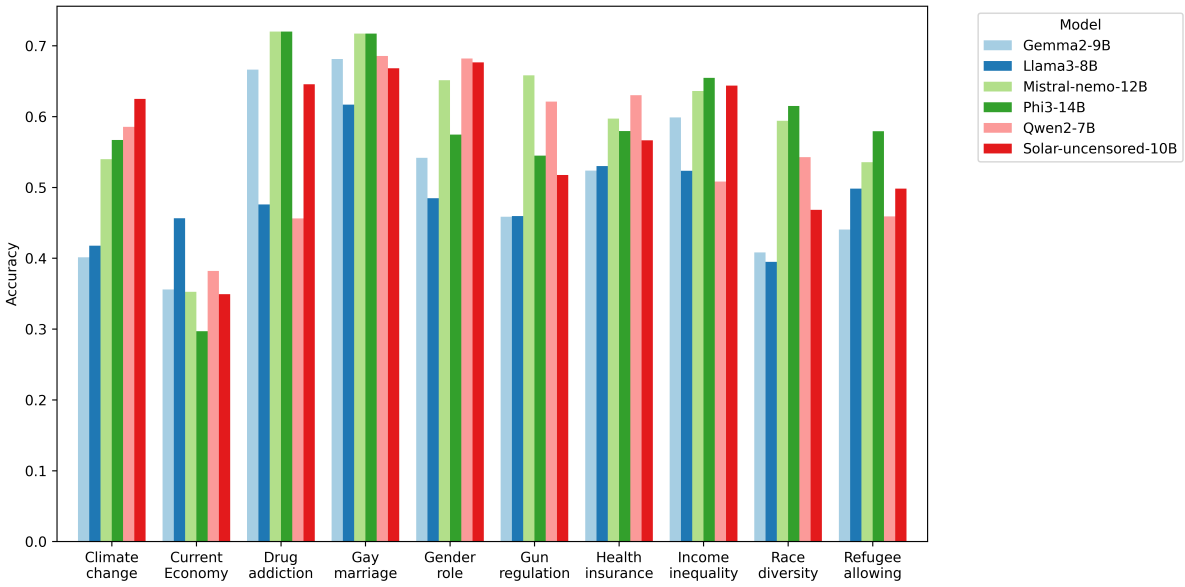

Figure 8: Accuracy grouped by topic.

We see from the data in Table 4 that **MistralN** has the highest overall accuracy, followed by **Phi3**. These two models consistently perform well across most topics, especially in *Drug addiction* and *Gay marriage*, where they show the highest alignment with human responses. On the other hand, **Llama3** shows the lowest average accuracy, particularly struggling with *Race diversity* and *Climate change*. Looking across topics, we see that *Gay marriage* and *Drug addiction* show high accuracy across all models, with averages of 0.681 and 0.614 respectively, indicating these topics might be easier for LLMs to simulate human responses accurately. *Current Economy* has the lowest average accuracy (0.366), suggesting it may be challenging for the models to replicate nuanced opinions on economic topics.

7.3 Result 2: Response Distribution Analysis

In this analysis we are interested in measuring the **distribution alignment** of the answers distributions generated by the LLM and the real respondents, providing insight into whether their response patterns across answers are similar, **irrespective of specific answer matches**. In other words, it is totally possible for two distributions to be maximally aligned, while their accuracy is zero. The **Jensen–Shannon divergence** is a method of measuring the similarity between two probability distributions. It is based on the Kullback–Leibler divergence, with some notable differences, including that it is sym-

metric and it always has a finite value. The square root of the Jensen–Shannon divergence is a metric (in the mathematical sense) and is called the **Jensen-Shannon Distance (JSD)**. We use JSD to assess the similarity between the distributions of the real respondents answers and the LLM-generated answers for all models and topics. This approach enables us to evaluate whether the LLMs capture the overall response distribution patterns, ensuring that even choices with lower frequencies are accounted for. Being bounded and symmetric, JSD is a good choice for comparing multiple models across various topics and facilitating interpretation by indicating whether generated responses closely approximate the observed distributions (lower JSD values) or diverge significantly (higher JSD values) [19]. Table 5 shows the results.

Table 5: Silicon Sample JSD

Topic	**Qwen2**	**Llama3**	**Gemma2**	**SolarU**	**MistralN**	**Phi3**	**Avg.**
Race diversity	0.287	0.319	0.297	0.266	0.049	0.274	0.249
Gender role	0.35	0.298	0.131	0.211	0.235	0.036	0.21
Current Economy	0.212	0.446	0.336	0.346	0.32	0.405	0.344
Drug addiction	0.316	0.224	0.265	0.09	0.329	0.329	0.259
Climate change	0.332	0.194	0.308	0.163	0.084	0.257	0.223
Gay marriage	0.097	0.144	0.082	0.15	0.083	0.062	0.103
Refugee allowing	0.267	0.221	0.243	0.189	0.325	0.204	0.242
Health insurance	0.096	0.108	0.194	0.168	0.268	0.335	0.195
Gun regulation	0.23	0.395	0.246	0.202	0.156	0.438	0.278
Income inequality	0.25	0.168	0.115	0.079	0.293	0.196	0.183
Avg.	0.244	0.252	0.222	0.186	0.214	0.254	

For quick visual comparison Figures 9 and 10 present the same results from different perspectives.

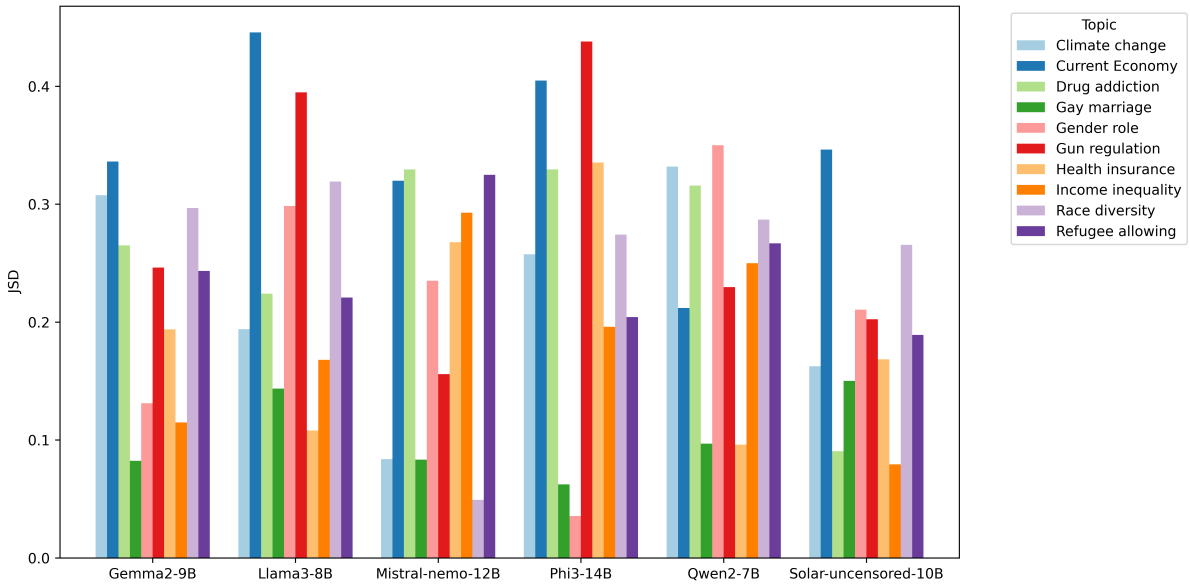

Figure 9: JSD grouped by model.

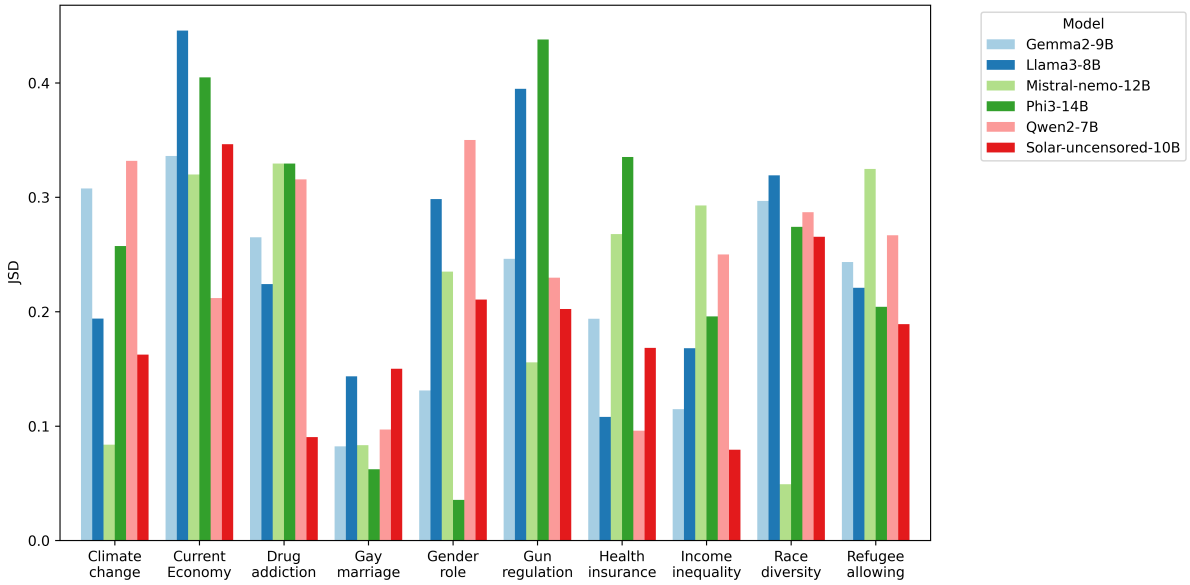

Figure 10: JSD grouped by topic.

The results in Table 5 shows that **SolarU** and **MistralN** have the best average JSD values, suggesting these models probability distributions are closest to human response distributions across topics. Across topics, *Gay marriage* and *Health insurance* exhibit the lowest JSD, indicating better alignment for these specific topics. **Phi3** has the highest JSD average, with particularly high values in *Current Economy* and *Gun regulation*, showing that this model may diverge from human response distributions in politically and economically sensitive topics. Also, *Race diversity* and *Gun regulation* have relatively

high average JSDs, indicating that these topics may be challenging for most models to accurately replicate human response distributions.

7.4 Result 3: Response Agreement Analysis

While accuracy and Jensen-Shannon Divergence offer valuable insights into the alignment between the LLM-generated responses and the observed data, they do so from distinct perspectives. **Accuracy** provides a measure of direct matching between individual responses, and **Jensen-Shannon Divergence (JSD)** captures the distributional similarity between response patterns across choices. However, both of these metrics have limitations when applied in cases where responses are unbalanced or where certain choices dominate the distribution, as they can overestimate agreement due to common, high-frequency responses. To address this, we introduce **Cohen's Kappa** as an additional metric, which is specifically designed to assess pairwise agreement while adjusting for chance. Unlike accuracy, which does not account for the likelihood of agreement due to chance alone, Cohen's Kappa calculates the expected agreement based on each response category's probability. This adjustment makes it especially valuable for imbalanced response distributions and for cases where certain response choices are more likely than others, which can otherwise inflate accuracy scores and obscure true alignment. High Kappa values indicate that the LLM is reliably predicting specific responses in line with the observed data, while low values suggest that any observed agreement might be largely due to common choice patterns rather than actual alignment. It is also possible for the statistic to be negative, which can occur by chance if there is no relationship between the ratings of the two raters, or it may reflect a real tendency of the raters to give differing ratings.

Table 6 shows the results for the Cohen's Kappa.

Table 6: Silicon Sample Kappa

Topic	Qwen2	Llama3	Gemma2	SolarU	MistralN	Phi3	Avg.
Race diversity	0.221	0.12	0.137	0.16	0.236	0.068	0.157
Gender role	0.005	0.07	0.153	0.101	0.037	0.112	0.079
Current Economy	0.038	-0.0	0.04	0.013	-0.09	0.028	0.005
Drug addiction	0.089	0.078	-0.006	0.112	0.0	0.0	0.045
Climate change	0.101	-0.029	0.134	0.248	0.127	0.081	0.111
Gay marriage	0.294	0.294	0.323	0.303	0.28	0.276	0.295
Refugee allowing	0.172	0.228	0.184	0.236	0.189	0.165	0.196
Health insurance	0.297	0.241	0.253	0.287	0.136	0.189	0.234
Gun regulation	0.317	0.098	0.175	0.234	0.37	0.0	0.199
Income inequality	0.302	0.288	0.393	0.427	0.334	0.405	0.358
Avg.	0.183	0.139	0.179	0.212	0.162	0.132	

As before, we provide a visual representation of these results in Figures 11 and 12.

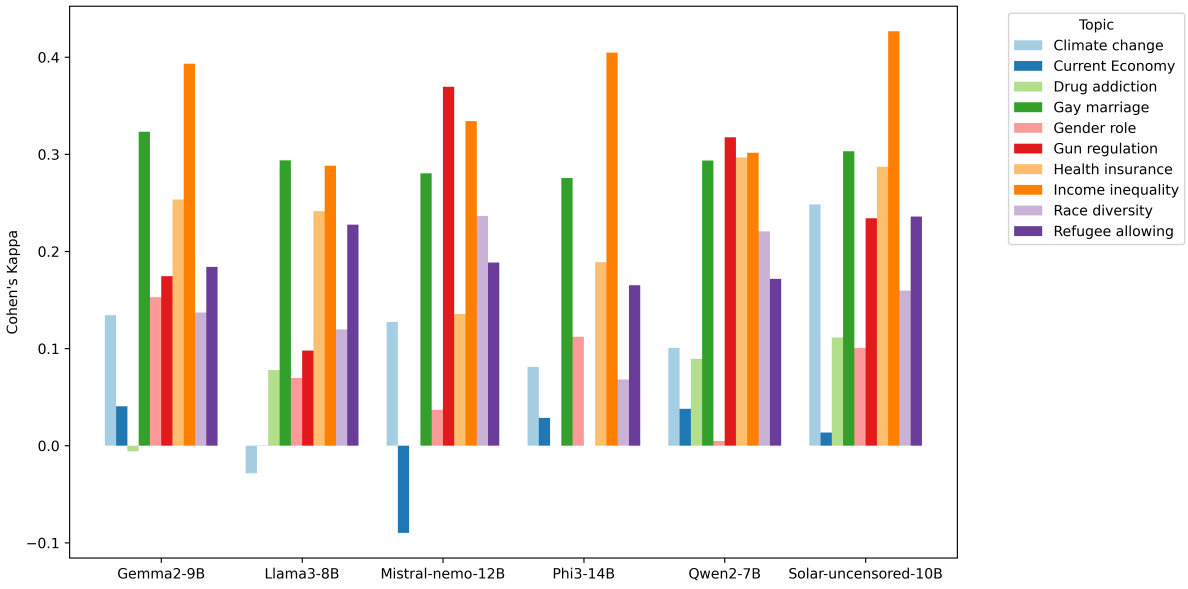

Figure 11: Cohen's Kappa grouped by model.

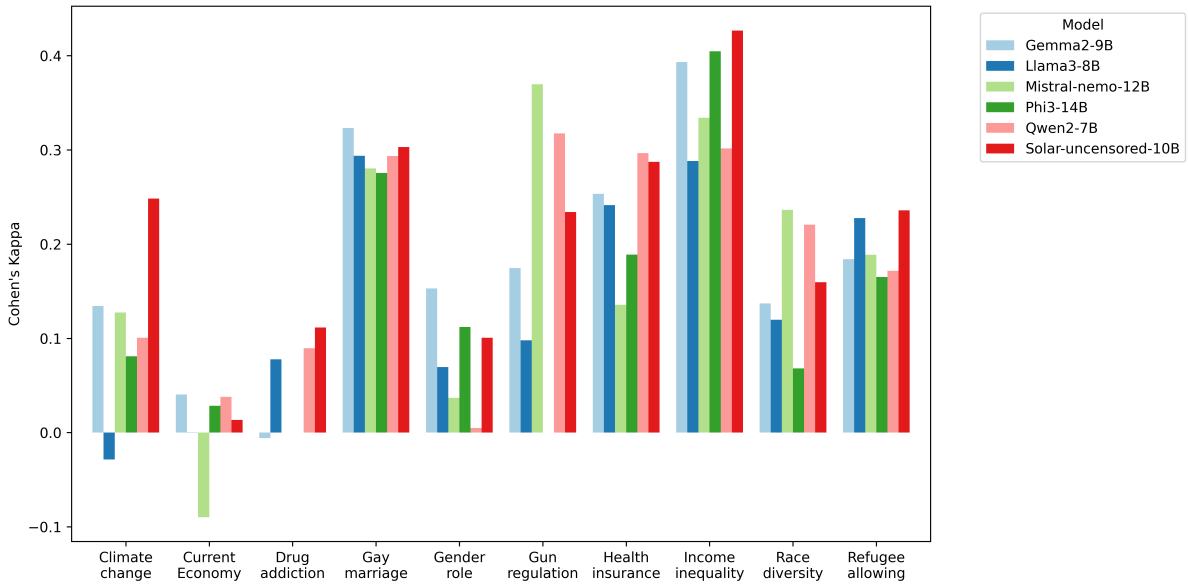

Figure 12: Cohen's Kappa grouped by topic.

The results in Table 6 shows that **Phi3** has the highest average kappa, with strong scores in *Income inequality* and *Health insurance*. This indicates good agreement with human responses, especially on socio-economic topics. **SolarU** also performs relatively well, particularly on *Income inequality*. On the other hand, **Llama3** and **MistralN** have the lowest kappa averages, indicating weak agreement across most topics. Looking that the topics, *Income inequality* has the highest average kappa, showing it may be a topic where LLMs can capture response patterns reliably. On the opposite side, *Current Economy* has the lowest average kappa, indicating poor agreement with human responses and possibly reflecting the complex, variable nature of economic opinions.

Interpreting the magnitude of Cohen's Kappa is not straightforward, as it is highly context-dependent and influenced by factors like response distribution and class imbalance. Kappa adjusts for agreement expected by chance, which means that in cases of imbalanced response categories, the score can be artificially lowered even if two raters (or models) frequently match on dominant choices. Consequently, what constitutes "good" or "strong" agreement varies by field and task complexity, as different research domains have different standards for acceptable agreement levels.

Furthermore, Kappa is often more informative as a relative measure rather than an absolute one. When used to compare agreement across different models or scenarios, Kappa can indicate which conditions yield higher alignment. Therefore, rather than relying solely on its numeric thresholds, Kappa should be interpreted in relation to the

specific context, response patterns, and the inherent difficulty of the task, offering a nuanced view of agreement quality.

7.5 Model-Specific Observations

- **Qwen2**: Shows a balanced performance across all metrics. It's neither the best nor the worst performer in any one metric, suggesting a good average capacity to simulate human responses.
- **Llama3**: Performs poorly on all metrics, with particularly low accuracy and kappa scores, suggesting limitations in capturing nuanced human opinion distributions.
- **Gemma2**: Shows moderate performance but has low kappa values, particularly struggling with agreement in sensitive topics like *Current Economy*.
- **SolarU**: Performs very well in all metrics. It is the best performing model overall.
- **MistralN**: Performs very well in all metrics, similarly to **SolarU**, but with a slightly higher JDS.
- **Phi3**: It shows good accuracy, bu performs poorly on both JSD and kappa scores. This is significant, since this is our largest model, with 14B parameters.

7.6 Topic-Specific Observations

- **Best Aligned Topics**: *Gay marriage* and *Drug addiction* consistently show high accuracy and low JSD across models, suggesting they may be easier for LLMs to simulate accurately. *Income inequality* shows high kappa, indicating good agreement across models, likely due to clearer, more distinct response patterns.
- **Challenging Topics**: *Current Economy* has low accuracy, high JSD, and almost negligible kappa scores, highlighting the difficulty of modeling human responses in complex and fluctuating economic situations. *Race diversity* and *Gun regulation* have high JSD and low kappa, showing that opinion diversity and polarization might hinder model alignment with human responses.

7.7 General Remarks

In summary, **SolarU** emerge as the strongest model across all metrics. In second place, we have **MistralN**, performing well in all metrics, but showing a significant lower kappa score than SolarU. All the other models underperform in one, two, or in the case of **Llama3**, all metrics. These results shows that model parameter size is not the deciding factor, and that model censorship probably plays a significant role when models are use to simulate human opinions. More experiments are planned to further investigate these behaviors and to determine which parameters affect them.

8 Concluding Remarks

Thesis conclusions (future work).

Statements and Declarations

Competing Interests

The authors declare no competing interests.

Acknowledgements

This work was made possible thanks to grants provided by *Universidade Presbiteriana Mackenzie*.

References

[1] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[2] D. Acemoglu and A. Ozdaglar. Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, 1:3–49, 2011.

[3] G. V. Aher, R. I. Arriaga, and A. T. Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

[4] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[5] V. Ashish. Attention is all you need. *Advances in neural information processing systems*, 30:I, 2017.

[6] R. Axelrod. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration: Agent-Based Models of Competition and Collaboration*. Princeton university press, 1997.

[7] N. M. Barrington, N. Gupta, B. Musmar, D. Doyle, N. Panico, N. Godbole, T. Reardon, and R. S. D'Amico. A bibliometric analysis of the rise of chatgpt in medical research. *Medical Sciences*, 11(3):61, 2023.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[9] Y.-S. Chuang, A. Goyal, N. Harlalka, S. Suresh, R. Hawkins, S. Yang, D. Shah, J. Hu, and T. T. Rogers. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618*, 2023.

[10] P. Cisneros-Velarde. On the principles behind opinion dynamics in multi-agent systems of large language models. *arXiv preprint arXiv:2406.15492*, 2024.

[11] P. Clifford and A. Sudbury. A model for spatial conflict. *Biometrika*, 60(3):581–588, 1973.

[12] J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, J.-N. Eckardt, N. G. Laleh, C. M. L. Löffler, S.-C. Schwarzkopf, M. Unger, G. P. Veldhuizen, et al. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141, 2023.

[13] S. De Marchi and S. E. Page. Agent-based models. *Annual Review of political science*, 17(1):1–20, 2014.

[14] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.

[15] M. H. DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121, 1974.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[18] R. Durrett, J. P. Gleeson, A. L. Lloyd, P. J. Mucha, F. Shi, D. Sivakoff, J. E. Socolar, and C. Varghese. Graph fission in an evolving voter model. *Proceedings of the National Academy of Sciences*, 109(10):3682–3687, 2012.

[19] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

[20] J. M. Epstein. Agent-based computational models and generative social science. *Complexity*, 4(5):41–60, 1999.

[21] J. M. Epstein and R. Axtell. *Growing artificial societies: social science from the bottom up*. Brookings Institution Press, 1996.

[22] A. Flache, M. Mäs, T. Feliciani, E. Chattoe-Brown, G. Deffuant, S. Huet, and J. Lorenz. Models of social influence: Towards the next frontiers. *Jasss-The journal of artificial societies and social simulation*, 20(4):2, 2017.

[23] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical methods for rates and proportions*. john wiley & sons, 2013.

[24] N. E. Friedkin and E. C. Johnsen. Social positions in influence networks. *Social networks*, 19(3):209–222, 1997.

[25] S. Galam. Majority rule, hierarchical structures, and democratic totalitarianism: A statistical approach. *Journal of Mathematical Psychology*, 30(4):426–434, 1986.

[26] S. Galam. Minority opinion spreading in random geometry. *The European Physical Journal B-Condensed Matter and Complex Systems*, 25:403–406, 2002.

[27] N. Gilbert and P. Terna. How to build and use agent-based models in social science. *Mind & Society*, 1:57–72, 2000.

[28] N. Gilbert and K. Troitzsch. *Simulation for the social scientist*. McGraw-Hill Education (UK), 2005.

[29] B. Golub and M. O. Jackson. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010.

[30] V. Grimm and S. F. Railsback. Individual-based modeling and ecology. In *Individual-based modeling and ecology*. Princeton university press, 2013.

[31] I. Grossmann, M. Feinberg, D. C. Parker, N. A. Christakis, P. E. Tetlock, and W. A. Cunningham. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109, 2023.

[32] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

[33] Ö. Gürcan. Llm-augmented agent-based modelling for social simulations: Challenges and opportunities. *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 134–144, 2024.

[34] R. Hegselmann. Bounded confidence revisited: What we overlooked, underestimated, and got wrong. *Journal of Artificial Societies and Social Simulation*, 26(4), 2023.

[35] R. Hegselmann, U. Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, 5(3), 2002.

[36] W. Huang, X. Ma, H. Qin, X. Zheng, C. Lv, H. Chen, J. Luo, X. Qi, X. Liu, and M. Magno. How good are low-bit quantized llama3 models? an empirical study. *arXiv preprint arXiv:2404.14047*, 2024.

[37] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[38] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023.

[39] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mixtral of experts, 2024.

[40] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[41] C. M. Macal. Everything you need to know about agent-based modelling and simulation. *Journal of Simulation*, 10(2):144–156, 2016.

[42] S. Masoudnia and R. Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42:275–293, 2014.

[43] J. H. Miller and S. E. Page. Complex adaptive systems: an introduction to computational models of social life. *(No Title)*, 2008.

[44] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, et al. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15, 2021.

[45] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[46] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[47] T. C. Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186, 1971.

[48] K. Sznajd-Weron and J. Sznajd. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11(06):1157–1165, 2000.

[49] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[50] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[51] U. Wilensky and W. Rand. *An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo*. MIT press, 2015.

[52] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[53] www.the decoder.com. Gpt-4 architecture, datasets, costs and more leaked. `https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/`, 2023. Accessed: 10 19, 2024.

[54] H. Xia, H. Wang, and Z. Xuan. Opinion dynamics: A multidisciplinary review and perspective on future research. *International Journal of Knowledge and Systems Science (IJKSS)*, 2(4):72–91, 2011.

[55] A. Yadav. A comprehensive review on large language models: Exploring applications, challenges, limitations, and future prospects. *Advancing Software Engineering Through AI, Federated Learning, and Large Language Models*, pages 17–39, 2024.

[56] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[57] Q. Zha, G. Kou, H. Zhang, H. Liang, X. Chen, C.-C. Li, and Y. Dong. Opinion dynamics in finance and business: a literature review and research opportunities. *Financial Innovation*, 6:1–22, 2020.

[58] S. Zuev. solar-10.7b-instruct-v1.0-uncensored, Dec 2023.