Behavioral Dynamics and Systemic Risks of Generative AI Agents as Endogenous Economic Actors in Financial Markets

Peng Yifeng
yifengpeng@link.cuhk.edu.cn

November 28, 2025

Abstract

This research proposal systematically explores the paradigm shift initiated by considering Generative Artificial Intelligence (GAI)—specifically, large language model (LLM)-based agents—as **Endogenous Economic Agents** within financial markets, endowed with independent cognition, decision-making capabilities, and behavioral biases. Diverging from conventional research trajectories that treat AI merely as an ancillary analytical tool, this study is grounded in the theoretical premise of "**Homo Silicus**" (The Siliconized Human). We hypothesize that LLM agents, trained on vast corpora of human text, implicitly inherit human social preferences and cognitive limitations, leading to the emergence of novel market dynamics distinct from traditional algorithmic trading in high-frequency interactions. The proposed methodology involves designing a comprehensive **Cognitive Architecture** that incorporates modules for perception, memory, reflection, and decision-making. This architecture will be implemented in a microsecond-precision Continuous Double Auction (CDA) market environment to simulate the complex game-theoretic behaviors of tens of thousands of heterogeneous AI agents. The core objectives are to quantitatively assess: (1) micro-level **behavioral biases** (e.g., variations in loss aversion) exhibited by AI agents; (2) meso-level risks of **Algorithmic Collusion**; and (3) the macro-level mechanisms through which these agents can trigger systemic risks, such as a **Flash Crash**. This proposal not only offers a new experimental paradigm for computational finance but also provides regulatory bodies with empirical evidence, based on a "**Regulatory Sandbox**" approach, for policy formulation in the emerging era of "human-machine hybridization."

Contents

1 Introduction

1.1 The Rise of Agentic AI in Finance: From Tools to Actors

The evolution of financial markets is fundamentally a history of co-evolution between the cognitive capabilities and decision-making speed of trading entities. For a long period, the predominant paradigm in both academia and the financial industry defined Artificial Intelligence (AI) as a "tool" or an "**Oracle**" designed to assist human decision-making. Under this "**Predictive Paradigm**," the core mandate of models, whether traditional econometrics or deep learning-based sequential prediction algorithms (such as LSTM or Transformer), was confined to minimizing forecast errors using historical data—i.e., predicting returns, volatility, or default probabilities [10]. However, with the explosive breakthrough of Generative AI (GAI) and Large Language Models (LLMs), we are currently undergoing a historical leap from the Predictive Paradigm to the "**Agentic Paradigm**" [8, 26].

Current LLM agents have transcended the scope of mere information processing, exhibiting human-like capabilities in **Reasoning**, **Memory**, **Reflection**, and autonomous **Planning**. In financial practice, the emergence of autonomous trading frameworks, exemplified by models such as TradingGPT and FinMem [23, 31], signals AI's transformation from a passive analytical instrument into a **Goal-Directed** actor. These intelligent agents are not only capable of processing multi-modal, heterogeneous information—including financial news and corporate reports—but can also independently execute trading instructions on a microsecond timescale, based on specific risk appetites and strategic logic. Consequently, AI is no longer merely a "lens" for observing the market; it is becoming the very "**cell**" that constitutes the market itself. This ontological shift compels us to re-examine the micro-foundations of financial markets: When thousands of AI agents endowed with high-level cognitive abilities interact within the market, how will they reshape asset pricing efficiency, liquidity structure, and systemic stability? This has rapidly become the foremost question demanding an answer from the field of computational finance [16].

1.2 Theoretical Gap: The Absence of "Homo Silicus" in Market Microstructure Models

Despite significant advancements in Agent-Based Computational Economics (ACE) over the past three decades, existing market microstructure models still confront a notable theoretical vacuum when simulating AI-dominated markets. Traditional Agent-Based Models (ABMs) are typically constrained by a binary dilemma: the choice between "**Zero-Intelligence**" agents and those governed by "**Hard-Coded Rules**" [11]. Early ZI agents, lacking cognitive capacity, could only simulate the most rudimentary supply-demand matching. While subsequent Heterogeneous Agent Models (HAMs) introduced the interplay between fundamentalists and chartists, their behavioral rules remained static and pre-set, failing to capture adaptive and learning capabilities [4].

More recently, Deep Reinforcement Learning (DRL) agents have emerged with learning capabilities; however, their "**Black-Box**" nature renders the decision-making process opaque and often results in convergence towards homogeneous optimal strategies, making it difficult to replicate the rich tapestry of human irrationality observed in real financial markets [14]. More critically, existing models fail to incorporate the central concept of "**Homo Silicus**" (The Siliconized Human) proposed by John Horton—the idea that LLM agents implicitly internalize human social preferences, heuristic thinking, and even cognitive biases through their pre-

training process [18]. Current literature conspicuously lacks research that systematically maps this "**silicon cognition**" onto market microstructure models. It remains unclear whether these "Silicon Economic Agents," trained on vast amounts of human text yet executing at machine speed, will correct human irrationality, thereby enhancing market efficiency, or if they will precipitate novel forms of market failure through algorithmic collusion and herding behavior [5]. This missing transmission mechanism, linking "**micro-cognition to macro-emergence**", constitutes the core theoretical gap that this study seeks to bridge.

1.3 Research Questions

Building upon the aforementioned context, this study aims to explore the behavioral dynamics of AI agents as **endogenous economic entities** through the construction of a high-fidelity generative multi-agent simulation system. Specifically, this research focuses on the following core questions across three distinct levels:

- **RQ1 [Micro-Level]: Cognitive Biases and Heterogeneity in AI Agents.** Will LLM-based trading agents inherit or even amplify human behavioral finance biases (e.g., the disposition effect, recency bias)? Given different **Persona Calibrations**—such as agents initialized based on genuine investor psychological profiles (Big Five Personality Traits)—what kind of heterogeneity will their risk aversion coefficients and decision logic exhibit when faced with market shocks? Will they behave as "**hyper-rational**" machines or as "**anthropomorphized**" investors with specific prejudices [15, 30]?

- **RQ2 [Meso-Level]: The Emergence Mechanism of Tacit Algorithmic Collusion.** In a Continuous Double Auction (CDA) market characterized by information asymmetry and incomplete contracts, will autonomous AI agents with high-level reasoning capabilities spontaneously form **Tacit Collusion**? For instance, can they manipulate market liquidity by engaging in acts such as **Spoofing** (false order placement) or synchronously widening the bid-ask spread, all without explicit communication protocols [3]? Does a Multi-Agent Debate mechanism inhibit or accelerate the formation of collusion in this process [1]?

- **RQ3 [Macro-Level]: Systemic Risk and Regulatory Boundaries.** Can the market dynamics dominated by generative agents successfully replicate the **Stylized Facts** of real financial markets, such as volatility clustering, fat-tailed return distributions, and long memory? Will the homogeneous reaction of AI agents to external information shocks significantly increase the probability of a **Flash Crash**? Are existing market **Circuit Breakers** still effective in countering nanosecond-response AI herding behavior, or is there a need to design novel regulatory tools based on algorithmic behavioral characteristics [16]?

1.4 Significance of the Study

The academic merit and practical relevance of this research are manifested across three dimensions: theoretical innovation, practical application, and regulatory implication.

- **Theoretical Dimension: Inaugurating the Paradigm of "Machine Behavioral Finance".** This study transcends the limitations of traditional finance, which often models

agents as homogeneous rational actors or simple noise traders. We are the first to systematically incorporate **Generative Agents** capable of language understanding and reasoning into asset pricing models. By bridging the cognitive theories of behavioral finance with the Agent architecture of computer science, this research will catalyze the development of **Machine Behavioral Finance** as an emerging interdisciplinary field, thereby providing a new theoretical cornerstone for understanding market microstructure in the algorithmic economy.

- **Practical Dimension: Constructing a High-Fidelity Strategy Validation Sandbox.** The research will culminate in the development of a simulation platform, the **Agent Trading Arena**, calibrated using real Level 2 order book data. This platform overcomes the deficiency of traditional back-testing, which fails to simulate crucial phenomena such as **Price Impact** and **Reflexivity**. It offers financial institutions a "**Digital Twin**" market to test the performance of complex trading strategies in an adversarial environment, yielding substantial practical application value.

- **Regulatory Dimension: Enabling "Ex-Ante Regulation" and Agile Governance.** In the face of increasingly complex algorithmic trading risks, traditional "**ex-post accountability**" regulation is demonstrably lagging. The simulation environment provided by this research serves as a "**Regulatory Sandbox**" for supervisory authorities, allowing policymakers to simulate the impact of high-frequency AI trading on market stability, test the effectiveness of antitrust policies, circuit breaker thresholds, or AI conduct codes *before* actual crises occur. This provides both empirical evidence and technical tools for financial regulation in the age of "**human-machine hybridization**."

2 Literature Review

This chapter systematically reviews the interdisciplinary literature spanning computational finance, artificial intelligence, and behavioral economics. Its primary purpose is to clearly define the academic position of this research and to elucidate how Generative Artificial Intelligence can address the **cognitive gap** present in existing market simulation models.

2.1 Agent-Based Computational Economics (ACE): From ZI to RL

The evolution of financial market simulation models has unfolded across three principal stages, each characterized by a progressive enhancement in the cognitive capabilities of the agents.

The first stage focused on research employing **Zero-Intelligence (ZI)** agents. The pioneering work by Gode and Sunder demonstrated [11] that ZI agents, whose behavior is purely random yet constrained by a simple budget restriction, could achieve prices in a double-auction market that were close to the theoretical equilibrium. This finding established the classic view that "**Market Institution** dictates allocative efficiency more profoundly than individual rationality." However, ZI models inherently fail to account for complex non-equilibrium phenomena such as asset bubbles and market crashes.

The second stage involved traditional Agent-Based Models (ABM) that introduced agents with **Heterogeneous Rules**. Platforms such as the Santa Fe Artificial Stock Market (SF-ASM) [2] and the JLM Simulator [20] modeled the strategic interplay between fundamental analysts (*Fundamentalists*) and chartists (*Chartists*). Lux and Marchesi demonstrated that this heterogeneous interaction is crucial for generating financial **Stylized Facts**, such as **Volatility**

Clustering and **Fat Tails** [25]. Nevertheless, the behavior of agents in these models is typically based on static "**If-Then**" rules, lacking the necessary adaptability when confronted with novel market information.

The third stage is characterized by financial simulations based on **Deep Reinforcement Learning (DRL)**. Following the success of AlphaGo, researchers began utilizing DRL algorithms (e.g., PPO, DQN) to train agents to maximize returns within complex financial environments. Frameworks like FinRL have showcased the potential of RL agents in portfolio management and high-frequency trading [24]. However, existing literature highlights significant limitations of RL agents: they are essentially "**Black-Box**" optimizers, lacking interpretability, and often converge to homogeneous optimal strategies (**Model Monoculture**), resulting in simulation markets that lack the cognitive diversity and irrational noise characteristic of real human markets [6].

2.2 LLMs in Finance

The advent of Large Language Models (LLMs) has catalyzed a new paradigm centered on **Generative Agents**. Park et al. demonstrated the emergent capabilities of generative agents—including memory, reflection, and social planning—in the "Stanford Town" experiment [26]. Since 2024, the latest literature has begun attempting to migrate this architectural approach to trading scenarios, forming an early research cluster focused on "**LLM-based Trading Agents**."

- **TradingGPT [23]:** Introduced a Layered Memory and a **Chain-of-Thought (CoT)** mechanism, enabling the agent to process unstructured text such as financial news. Its innovation lies in the design of a **Multi-Agent Debate** mechanism, which significantly enhanced decision robustness by having subsidiary agents with bullish and bearish viewpoints debate to correct for potential hallucinations.

- **FinMem [31]:** Focuses on simulating the cognitive limitations of human investors. This study constructed a memory retrieval mechanism based on the Ebbinghaus forgetting curve, allowing the agent to dynamically adapt to market **Regime Shifts**. It successfully demonstrated that agents endowed with **Episodic Memory** outperform simple momentum strategies in long-term investing.

- **StockAgent [32]:** Shifted the research perspective from single-agent profitability to collective game theory. This framework simulated the interactions of multiple LLM agents in a paper trading environment and observed price discovery processes that diverged from traditional ABMs.

While these works validate the potential of LLMs in processing financial information, most have concentrated on the profitability backtesting of individual agents. They have largely overlooked the **Reflexivity** impact on market microstructure (e.g., liquidity, spreads, and depth) arising from large-scale agent interaction [28]. This research aims to specifically overcome this limitation, focusing instead on the system's macro-emergence properties.

2.3 Behavioral Finance & AI: Biases in LLMs

Another core dimension for considering AI as a subject of financial research is its potential for "**human-like**" cognitive biases. The "**Homo Silicus**" hypothesis, proposed by Horton

(2023) [18], posits that LLMs can function as powerful simulators of human economic behavior. Current empirical research presents two distinctly conflicting findings, which constitute a valuable academic debate:

On one hand, the study by Henning et al. [17] discovered that uncalibrated GPT-4 agents exhibited "**Hyper-Rationality**" in asset pricing experiments, tending to anchor prices near fundamental values and rarely generating bubbles. This finding suggests that generic LLMs may have undergone excessive **Reinforcement Learning from Human Feedback (RLHF)** alignment, potentially causing them to lose the capacity to simulate the genuine **Animal Spirits** of real markets.

On the other hand, research by Zhou & Ni [34] and Yang et al. [30] indicates that through specific **Persona Prompting**, LLMs can conspicuously display classic behavioral finance biases:

- **Loss Aversion:** Agents exhibit the **Disposition Effect**, manifesting as a tendency to prematurely sell profitable stocks while holding onto losing ones.
- **Herding Behavior:** In social network environments (such as simulated Twitter), LLM agents are highly susceptible to collective sentiment, leading to irrational momentum chasing and selling [22].

This contradiction underscores the criticality of **Calibration**: To accurately replicate authentic financial markets, we must not directly employ raw models, but rather construct a heterogeneous population of agents whose behavioral profiles align with the empirical distribution of real investor psychological characteristics.

2.4 Market Microstructure & Algorithmic Collusion

At the level of market microstructure, algorithmic collusion has emerged as a novel risk commanding high scrutiny from regulatory bodies (such as the SEC and FTC). Calvano et al. [5] first identified the phenomenon of **Tacit Collusion** in their research on AI pricing algorithms: AI agents, with no explicit communication and no hard-coded collusive instructions in their code, spontaneously learned to maintain supra-competitive prices solely through trial-and-error learning.

Dou & Goldstein [9], in their NBER working paper, further extended this finding to financial trading algorithms. They identified two micro-mechanisms of AI collusion:

1. **Price-Trigger Strategies:** Analogous to the "Grim Trigger Strategy" in game theory, AI agents learned to engage in aggressive retaliatory pricing against any competitor attempting to narrow the bid-ask spread.
2. **Over-Pruning Bias:** Termed "**Artificial Stupidity**," this refers to the phenomenon where AI agents prematurely abandon competitive strategies during the exploration phase, settling into a sub-optimal collusive equilibrium.

Furthermore, the 2010 **Flash Crash** event demonstrated that high-frequency algorithmic feedback loops are critical factors contributing to market fragility [21]. Danielsson et al. [7] cautioned that while LLM agents may not match the processing speed of traditional High-Frequency Trading (HFT) algorithms, their homogeneous interpretation of news sentiment (**Model Monoculture**) could trigger a novel, semantic-understanding-based **Herding Effect**, leading to an instantaneous liquidity drain. Existing literature has yet to systematically test this risk within an L2-level **Limit Order Book (LOB)** simulation environment [12, 16], a gap that this research endeavors to fill.

3 Theoretical Framework

This research establishes a closed-loop generative agent simulation framework by integrating theories from computational cognitive science, behavioral finance, and market microstructure. This chapter defines the agent's cognitive architecture (internal state), the methodology for persona calibration (source of heterogeneity), and the market interaction mechanism (external environment).

3.1 Cognitive Architecture for Financial Agents (CAFA)

To simulate trading entities endowed with "**human-like**" decision-making capabilities, this study proposes a specialized **Cognitive Architecture for Financial Agents (CAFA)**. This architecture transcends the simple "state-action" mapping of traditional Reinforcement Learning (RL) agents by introducing explicit memory and reasoning modules, enabling the agents to process unstructured information and exhibit adaptive behavior. CAFA comprises the following four core modules:

3.1.1 Perception Module: Multimodal Signal Processing

Agents do not directly observe the state space S_t; rather, they process multimodal information streams through a perception filter.

- **Textual Perception:** Utilizes a financial domain-fine-tuned **Embedding Model** (such as FinBERT or a dedicated LLM embedding layer) to process financial news, central bank announcements, and social media sentiment [19]. The system transforms the unstructured text D_t into a semantic vector v_{text}, and extracts features related to "sentiment polarity" and "event type."
- **Numerical & Visual Perception:** Agents not only receive Level 2 (L2) **Limit Order Book (LOB)** data but are also equipped with a visual encoder (such as a CLIP variant) to directly interpret Candlestick Charts, thereby identifying technical patterns like "**Head and Shoulders**" or "**Double Bottoms**," simulating the visual cognitive process of chartist traders.

3.1.2 Memory Module: Decay and Retrieval

Inspired by the "Generative Agents" architecture of Park et al. [26], CAFA introduces a hierarchical memory system to resolve the conflict between the limited Context Window of LLMs and the need for long-term market interaction. The memory stream encompasses **Working Memory** (current market state), **Episodic Memory** (historical trading and P&L experiences), and **Semantic Memory** (long-term accumulated trading rules).

The memory retrieval mechanism simulates the human law of forgetting. For a given query q (e.g., a sudden market news event), the retrieval score $Score(m, q)$ for a memory snippet m is defined as:

$$Score(m, q) = \alpha \cdot I(m) + \beta \cdot R(m, q) + \gamma \cdot e^{-\lambda(t_{now} - t_m)}$$

Where:

- $I(m)$ denotes the **Importance**, automatically annotated by the LLM during memory generation (e.g., memories of substantial losses are assigned high importance);

- $R(m, q)$ denotes the **Relevance**, measured as the cosine similarity between the query vector and the memory vector;
- $e^{-\lambda(t_{now}-t_m)}$ denotes the **Recency**, simulating the exponential decay of memory over time;
- α, β, γ are hyperparameters used to adjust the weights.

3.1.3 Reasoning Module: Chain-of-Thought and Reflection

This constitutes the core decision engine of CAFA.

- **Chain-of-Thought (CoT):** Prior to outputting a trading instruction, the agent must generate a natural language logical derivation (e.g., "Although the RSI indicator is overbought, given the recently released positive earnings report and the stack of buy orders in the LOB, I conclude that the upward trend is not over."). This significantly enhances the model's interpretability [29, 33].
- **Self-Reflection:** A "**Trading Diary**" mechanism is introduced. After the close of the trading day, the agent compares its expectations against the actual outcomes. Should a loss occur, the agent generates a reflective text stored in its Semantic Memory (e.g., "I underestimated the impact of the interest rate hike on technology stocks"), thereby allowing it to correct its strategy in subsequent decisions [27, 31].

3.2 The "Homo Silicus" Hypothesis and Persona Calibration

The research relies on the **Homo Silicus** (Silicon Human) hypothesis, which posits that LLMs implicitly internalize human social preferences and cognitive biases [18]. To preclude the systemic biases caused by **Model Monoculture**, it is imperative that we construct a highly heterogeneous population of agents.

3.2.1 Calibration via Survey of Consumer Finances (SCF)

This study proposes to use the **Survey of Consumer Finances (SCF)** data to conduct a "demographic mapping" of the agents. The specific procedure is as follows:

1. **Data Sampling:** Sample real investor individuals i from the SCF database and extract their attribute vector $\mathbf{x}_i$.
2. **Prompt Engineering Mapping:** Construct a mapping function $\mathcal{F} : \mathbf{x}_i \rightarrow \text{System Prompt}_i$. For example, a high-net-worth individual with low risk aversion is translated into the Prompt: "You are a 55-year-old veteran investor with $5 million in assets, primarily focused on long-term capital appreciation and insensitive to short-term volatility."
3. **Bias Injection:** Based on behavioral finance literature, cognitive biases will be explicitly injected into a specific proportion of agents (e.g., endowing 30% of retail agents with the **Disposition Effect**, and 20% with **Overconfidence**) to test the market's fragility under irrational exuberance.

3.3 Mechanism Design: Continuous Double Auction (CDA) & LOB Dynamics

To capture the microstructure characteristics inherent in high-frequency trading environments, this research models the market as a **Continuous Double Auction (CDA)** simulation, rather than simple end-of-day closing price backtesting.

3.3.1 Limit Order Book (LOB) Dynamics

The market state is completely described by the **Limit Order Book** $\mathcal{L}_t$, which contains the queue of Bids and Asks. The simulation engine adheres to the matching principle of **Price-Time Priority**.

- **Order Types:** Agents can submit Limit Orders, Market Orders, and Cancel Orders.
- **Market Impact:** Unlike conventional models that assume infinite liquidity, this system endogenously calculates **Price Impact**. Large Market Orders will "**eat through**" multiple layers of the LOB depth, resulting in an average execution price inferior to the best quote, thereby truly reflecting the friction costs of large-scale trading.

3.3.2 Latency and Asynchronous Interaction

To study algorithmic collusion and flash crashes, the time dimension is modeled as microsecond-level discrete events. The system introduces a heterogeneous latency parameter δ_i:

$$t_{execute} = t_{decision} + \delta_i + \epsilon$$

Where $\delta_{HFT} \ll \delta_{Retail}$. This **asynchronous interaction** mechanism allows us to simulate the speed advantage of High-Frequency Traders (HFTs) relative to ordinary LLM agents, and how this advantage might evolve into **Predatory Trading**.

4 Methodology

This chapter details the technical roadmap for constructing the high-fidelity "**Agent Trading Arena**." To resolve the spatio-temporal mismatch between the slow inference speed of Generative Agents and the microsecond-level matching required by financial markets, this research proposes an asynchronous, loosely coupled hybrid system architecture.

4.1 System Architecture: The "Agent Trading Arena"

To support high-frequency interaction among $N = 10,000+$ heterogeneous agents, we forgo traditional single-threaded backtesting frameworks in favor of a distributed architecture based on the **Actor Model**. The system comprises three core layers:

4.1.1 Hybrid Computational Layer (Python/Rust)

- **Agent Logic Layer (Python):** Given that mainstream LLM frameworks (such as PyTorch and LangChain) are rooted in the Python ecosystem, the agent's cognitive modules (Perception, Memory, Reasoning) will be executed in a Python environment. Each agent operates as an independent `AsyncIO` process, responsible for maintaining its own context window and memory retrieval.

- **Matching Engine Core Layer (Rust):** To accurately simulate real **Limit Order Book (LOB)** dynamics, the Matching Engine will be implemented in Rust. Rust's memory safety and zero-cost abstraction characteristics enable it to process order insertions, cancellations, and matching at a microsecond level, thus bypassing the Python **Global Interpreter Lock (GIL)** bottleneck.

4.1.2 Asynchronous Communication via RabbitMQ

Communication between the agents and the exchange is decoupled using a **Message Queue** (specifically RabbitMQ/ZeroMQ).

- **Uplink (Orders):** Agents send order messages, adhering to the **FIX Protocol** standard (in JSON format), to the `EXCHANGE_IN` queue.
- **Downlink (Market Data):** The exchange broadcasts L2-level LOB snapshots and execution information (**Tick Data**) to the `MARKET_DATA_PUB` channel.

This asynchronous architecture not only simulates network latency present in real trading but also allows us to horizontally scale the number of agents across a distributed cluster.

4.2 Simulation Engine Dynamics

The central task of the simulation engine is to generate endogenous price paths and liquidity characteristics, moving beyond the simple replay of historical data.

4.2.1 Matching Logic and Continuous Double Auction (CDA)

The market utilizes the **Continuous Double Auction (CDA)** mechanism. At any given time t, the **Limit Order Book** $\mathcal{L}_t$ comprises the set of Bids $\mathcal{B}_t$ and the set of Asks $\mathcal{A}_t$. Matching strictly adheres to the principle of **Price-Time Priority**. A trade is executed immediately when a newly arriving buy order b_{new} satisfies $P(b_{new}) \geq \min_{a \in \mathcal{A}_t} P(a)$; the remaining portion then enters the set $\mathcal{B}_t$. This mechanism ensures that the micro-foundation of market clearing is consistent with major exchanges like NASDAQ and NYSE [11].

4.2.2 Stochastic Latency Modeling

To investigate algorithmic collusion and the advantages held by high-frequency trading, we introduce a heterogeneous latency model. The delay Δt_i between the agent i's instruction sending time t_{send} and the exchange processing time t_{proc} is modeled as:

$$\Delta t_i = \delta_{net} + \delta_{compute} + \epsilon_t$$

Where:

- δ_{net} represents the network transmission latency. For market-making agents hosted on high-frequency servers, $\delta_{net} \sim \text{Exp}(\lambda_{fast})$; for ordinary retail agents, $\delta_{net} \sim \text{Exp}(\lambda_{slow})$.
- $\delta_{compute}$ represents the LLM inference latency, which is dependent on the number of tokens generated.
- ϵ_t is the stochastic jitter.

This modeling approach allows us to precisely quantify how the "**speed advantage**" is converted into an "**information advantage**."

4.2.3 Endogenous Market Impact

In contrast to traditional backtesting, which assumes infinite liquidity, this system endogenously computes **Market Impact**. For a market order of size Q, the actual average execution price $\bar{P}$ depends on the depth distribution of the LOB:

$$\bar{P}(Q) = \frac{1}{Q} \int_0^Q P(v)\, dv$$

Where $P(v)$ is the price of the v-th unit of liquidity consumed in the order book. This feature necessitates that agents weigh trading speed against **slippage cost**, which should lead to the emergence of advanced strategies such as **Order Splitting**.

4.3 Agent Calibration: From Survey Data to Prompts

To construct a "**Silicon Society**" whose distribution aligns with the real world, we will calibrate the agents using data from the **Survey of Consumer Finances (SCF)** [18].

4.3.1 Demographic Mapping Pipeline

We map each sample household h from the SCF dataset to an independent LLM agent. The mapping function $\Phi : R^d \to \mathcal{T}$ transforms numerical features into a natural language Prompt template $\mathcal{T}$:

> *Template:* "You are a [Age] year old investor with a net worth of $[Net Worth]. Your risk tolerance is [Risk Level] (based on SCF Question X3014). You work in the [Industry] sector. Currently, the market news indicates..."

4.3.2 Risk Tolerance Calibration

Risk preference is categorized into four levels (ranging from "unwilling to take any risk" to "willing to take substantial risk for high returns"). We will utilize **Few-Shot Prompting** to fine-tune the LLM, ensuring that the risk aversion coefficient γ exhibited in the agent's asset allocation decisions matches the implied coefficient derived from the SCF data [13].

4.4 Data Sources

This study will employ multimodal data to both drive and validate the simulation environment:

- **LOBSTER (Level 2 Market Data):** High-precision **LOB Reconstitution** data from NASDAQ will be used as the "seed" environment for the simulation, providing the initial liquidity state and the prior distribution of order flow arrival rates.
- **FinGPT (Financial News Sentiment):** The open-source FinGPT dataset will be leveraged to provide a stream of financial news aligned with a historical timeline. Agents will read this news and generate sentiment signals to drive their trading decisions.
- **Synthetic Data via CTGAN:** To stress-test extreme scenarios (such as the 2010 Flash Crash), we will use a **Conditional Tabular Generative Adversarial Network (CTGAN)** to produce synthetic extreme order flow data.

5 Experimental Design

This research will adhere to the logical chain of "**Micro-Validation** $\rightarrow$ **Macro-Emergence** $\rightarrow$ **Policy Intervention**," designing three core experiments. All experiments will be conducted within the aforementioned *Agent Trading Arena* simulation environment.

5.1 Experiment 1: Micro-validation of Behavioral Biases

Objective: To verify whether the **Persona Calibrated** LLM agents successfully replicate classic behavioral biases of human investors at the individual system level, particularly the **Disposition Effect**.

Experimental Setup:

- **Subjects:** Initialize $N = 500$ LLM agents calibrated based on different SCF samples (differentiating between "**Conservative**" and "**Aggressive**" types).
- **Environmental Input:** Agents are fed a **Synthetic Price Path** containing both positive and negative shocks, ensuring that all agents face the identical market history.
- **Task:** Agents must decide at each time step whether to "**Hold**," "**Sell to Realize Gain**," or "**Sell to Cut Loss**."

Metrics and Hypotheses: We adopt the classic metric proposed by Odean (1998), calculating the Proportion of Gains Realized (PGR) and the Proportion of Losses Realized (PLR):

$$PGR = \frac{\text{Realized Gains}}{\text{Realized Gains} + \text{Paper Gains}}, \quad PLR = \frac{\text{Realized Losses}}{\text{Realized Losses} + \text{Paper Losses}}$$

- **Hypothesis H_1:** For generic LLM agents without specific prompt engineering, their behavior tends towards rationality, i.e., $PGR \approx PLR$.
- **Hypothesis H_2:** Agents injected with a "**Loss Aversion**" persona prompt will exhibit a significant **Disposition Effect**, meaning $PGR > PLR$ (prematurely selling winners and holding losers), and this difference will be statistically significant (t-test, $p < 0.01$).

5.2 Experiment 2: Emergence of Stylized Facts

Objective: To investigate whether a market composed of generative agents can spontaneously exhibit the **Stylized Facts** of real financial markets, thereby validating the model's effectiveness as a macro-simulator.

Experimental Setup:

- **Control Group (Baseline):** The market is composed of 100% **Zero-Intelligence (ZI)** agents, who are subject to budget constraints but place orders randomly.
- **Experimental Group (Homo Silicus):** The market consists of 10% Fundamentalist agents, 40% Momentum Trading agents (Chartists), and 50% Noise Traders, with the Momentum agents possessing **CoT-based** trend reasoning capabilities.
- **Procedure:** The simulation is run for 10,000 time steps, recording the minute-by-minute closing price P_t and the logarithmic return $r_t = \ln P_t - \ln P_{t-1}$.

Key Validation Metrics: We will compare the statistical characteristics of the return series generated by the two market groups:

1. **Fat Tails:** Calculate the **Kurtosis** of the return distribution. Real markets typically satisfy $K > 3$. We hypothesize that $K_{exp} \gg K_{control} \approx 3$.
2. **Volatility Clustering:** Examine the Autocorrelation Function (ACF) of the absolute returns $|r_t|$. If clustering effects are present, $Corr(|r_t|, |r_{t-\tau}|)$ will remain significantly positive for larger values of τ.
3. **Long Memory:** Calculate the **Hurst Exponent** H using R/S analysis. If $0.5 < H < 1$, it indicates that the market exhibits persistent trends and does not follow a random walk.

5.3 Experiment 3: Systemic Risk & Regulation via Sandbox

Objective: To simulate extreme risk scenarios within a "**Regulatory Sandbox**" and test the effectiveness of novel market manipulation detection and regulatory tools in AI-dominated markets.

5.3.1 Scenario A: Tacit Algorithmic Collusion

Context: Theory predicts that in oligopolistic Market Maker (MM) markets, AI may discover that "maintaining a high bid-ask spread" is a Nash Equilibrium through trial-and-error learning (Dou et al., 2025).

- **Setup:** Only five large LLM Market Maker agents are retained, retail agents are removed, and they are engaged in a long-term game.
- **Detection Mechanism:** Monitor the evolution of the **Bid-Ask Spread**. Collusion is inferred if the spread is significantly higher than the perfectly competitive level, without any evidence of explicit communication.
- **Policy Test:** Introduce a **Zero-Knowledge Proof-based Regulation** (ZKP-based Regulation), requiring Market Makers to prove that their quoting algorithms do not contain specific "retaliatory trigger" logic, and observe whether this disrupts the collusive equilibrium.

5.3.2 Scenario B: Flash Crash and Circuit Breakers

Context: The 2010 **Flash Crash** demonstrated that the homogeneous reaction of high-frequency algorithms can instantaneously deplete liquidity.

- **Stress Test:** At time $t = 5000$, introduce a massive exogenous negative news shock (generated via FinGPT as extremely pessimistic macro-economic news).
- **Herding Metric:** Use the **LSV (Lakonishok, Shleifer, and Vishny)** measure to monitor the real-time consistency of buy and sell directions:

$$LSV_t = \left| \frac{B_t}{B_t + S_t} - P_t^{buy} \right| - E_t \tag{1}$$

- **Circuit Breaker Intervention:** Compare the **Time to Recovery** across three regulatory environments:

- *No Intervention (Laissez-faire)*
- *Traditional Circuit Breakers (Traditional CB):* Trading is halted for 15 minutes if the index falls by $> 7\%$.
- *AI-Specific Limits:* Restrictions on the **Order-to-Trade Ratio** or the imposition of a dynamic **Tobin Tax** on aggressive momentum strategies.

6 Expected Contributions and Timeline

This research aims to establish the empirical foundation of "**Machine Behavioral Finance**" through a closed-loop trajectory of "Build $\rightarrow$ Simulate $\rightarrow$ Validate." The expected outcomes are not only intended to expand academic theory but also to provide open-source tools of industrial strength and concrete policy recommendations for regulation.

6.1 Expected Contributions

The core contributions of this project can be summarized across the following three dimensions (3P: Papers, Platform, Policy):

6.1.1 1. Theoretical Innovation: Founding "Machine Behavioral Finance"

- **Reconstruction of Micro-Foundations:** For the first time, this study introduces Large Language Model (LLM)-based "**Homo Silicus**" (Silicon Economic Agents) into market microstructure research, quantitatively assessing the causal impact of "**Persona Calibration**" on asset pricing.
- **Explaining Algorithmic Collusion:** We will unveil the specific path dependencies through which autonomous AI agents evolve **Tacit Collusion** via trial-and-error learning, even without explicit communication. This challenges existing antitrust law standards for determining "**intent**."

6.1.2 2. Methodological Contribution: The Open-Source "Agent Trading Arena"

- **High-Performance Simulation Benchmark:** We will release the first open-source simulation platform based on a `Rust/Python` hybrid architecture, capable of supporting high-frequency interaction among tens of thousands of LLM agents. This platform endogenously incorporates **Limit Order Book (LOB) Dynamics** and network latency, solving the critical limitations of traditional backtesting regarding the simulation of "**Price Impact**" and "**Reflexivity**."
- **Standardized Dataset:** We will provide an *Agent Persona Repository*, a standardized **Benchmark** calibrated using **SCF (Survey of Consumer Finances)** data, to assist the academic community in studying heterogeneous agents.

6.1.3 3. Practical & Regulatory Implications: AI Governance Sandbox

- **Regulatory Sandbox Mechanism:** We will propose an algorithmic regulatory framework based on **Zero-Knowledge Proofs (ZKP)**, allowing regulatory bodies to verify whether trading algorithms contain predatory strategies without compromising the intellectual property of the code.

- **Policy White Paper:** A Policy Brief will be drafted for regulatory bodies such as the SEC or CFTC, recommending adjustments to the trigger thresholds and recovery logic of **Circuit Breakers** in AI-dominated high-frequency markets.

6.2 Research Timeline

This research project is planned for completion within **12 to 15 months**, divided into three phases: infrastructure construction, simulation experiment execution, and results analysis and dissemination.

6.2.1 Phase 1: Infrastructure & Calibration (Months 1-4)

Focus: System Development & Data Engineering

- **Month 1 (Core Engine):** Develop the high-performance **Matching Engine** based on Rust, implementing the logic for LOB order insertion, cancellation, and execution; set up the RabbitMQ message queue for asynchronous communication between the Python Agent and Rust Engine.
- **Month 2 (Agent Architecture):** Build the **CAFA Cognitive Architecture** using LangChain, implementing the Perception-Memory-Reflection modules; integrate FinBERT and CLIP models to handle multimodal (text/chart) inputs.
- **Month 3 (Data Pipeline):** Clean and process LOBSTER L2 high-frequency data as the market environment "seed"; utilize FinGPT to process historical financial news streams; generate **Synthetic Crash Data** for extreme scenario testing.
- **Month 4 (Persona Calibration):** Process SCF survey data, use **Prompt Engineering** to construct the persona descriptors for $N = 10,000$ heterogeneous agents, and conduct small-scale Turing tests to validate the consistency of their risk preferences.

6.2.2 Phase 2: Simulation & Experimentation (Months 5-8)

Focus: Running Experiments & Iterative Tuning

- **Month 5 (Exp 1 - Micro Validation):** Run single-agent experiments to test the **Disposition Effect** in agents with different Personas when facing gains and losses; calibrate the loss aversion coefficient λ.
- **Month 6 (Exp 2 - Macro Emergence):** Launch the full-scale market simulation. Compare the statistical characteristics of the "**Zero-Intelligence Agent Market**" versus the "**LLM Agent Market**" to validate the emergence of Volatility Clustering and Fat Tails.
- **Month 7 (Exp 3 - Systemic Risk):** Simulate extreme scenarios (e.g., replicating the **2010 Flash Crash**). Measure the **Herding Intensity** of the AI agent population and the speed of liquidity depletion following a sudden negative news shock.
- **Month 8 (Regulatory Sandbox):** Introduce intervention variables such as **Circuit Breakers** and **Transaction Taxes**; run **Counterfactual Simulations** to evaluate the impact of various regulatory policies on market recovery efficiency.

6.2.3 Phase 3: Analysis, Writing & Dissemination (Months 9-12)

Focus: Data Analysis & Output Production

- **Month 9 (Data Analysis):** Process TB-scale simulation log data. Use econometric methods (such as VAR, GARCH models) to analyze the causal relationship between agent behavior and market prices.
- **Month 10 (Drafting Paper 1):** Draft a paper on "**Micro-Behavioral Biases in AI Agents**," targeting submission to a Fintech special issue of the *Journal of Finance (JF)* or the *Review of Financial Studies (RFS)*.
- **Month 11 (Drafting Paper 2):** Draft a paper on "**Algorithmic Collusion and the Regulatory Sandbox**," targeting submission to the AI for Finance Workshop at *NeurIPS* or *ICML*.
- **Month 12 (Thesis & Open Source):** Finalize the PhD thesis proposal/mid-term report; organize code documentation and publicly release the *Agent Trading Arena* as **Open Source** on GitHub.

References

[1] Kushal Agrawal, Verona Teo, Juan J Vazquez, Sudarsh Kunnavakkam, Vishak Srikanth, and Andy Liu. Evaluating llm agent collusion in double auctions. *arXiv preprint arXiv:2507.01413*, 2025.

[2] W Brian Arthur, John H Holland, Blake LeBaron, Richard Palmer, and Paul Tayler. Asset pricing under endogenous expectations in an artificial stock market. In *The economy as an evolving complex system II*, pages 15–44. CRC Press, 2018.

[3] Alessio Azzutti, Wolf-Georg Ringe, and H Siegfried Stiehl. Machine learning, market manipulation, and collusion on capital markets: Why the" black box" matters. *U. Pa. J. Int'l L.*, 43:79, 2021.

[4] William A Brock and Cars H Hommes. Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic dynamics and Control*, 22(8-9):1235–1274, 1998.

[5] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolo, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.

[6] Jon Danielsson, Robert Macrae, and Andreas Uthemann. Artificial intelligence and systemic risk. *Journal of Banking & Finance*, 140:106290, 2022.

[7] Jon Danielsson and Andreas Uthemann. Artificial intelligence and financial crises. *Journal of Financial Stability*, page 101453, 2025.

[8] Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. Large language model agent in financial trading: A survey. *arXiv preprint arXiv:2408.06361*, 2024.

[9] Winston Wei Dou, Itay Goldstein, and Yan Ji. Ai-powered trading, algorithmic collusion, and price efficiency. *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper, The Wharton School Research Paper*, 2025.

[10] Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European journal of operational research*, 270(2):654–669, 2018.

[11] Dhananjay K Gode and Shyam Sunder. Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *Journal of political economy*, 101(1):119–137, 1993.

[12] Martin D Gould, Mason A Porter, Stacy Williams, Mark McDonald, Daniel J Fenn, and Sam D Howison. Limit order books. *Quantitative Finance*, 13(11):1709–1742, 2013.

[13] John E Grable and Ruth H Lytton. Assessing the concurrent validity of the scf risk tolerance question. *Journal of Financial Counseling and Planning*, 12(2):43, 2001.

[14] Ben Hambly, Renyuan Xu, and Huining Yang. Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3):437–503, 2023.

[15] John Hartley, Conor Brian Hamill, Dale Seddon, Devesh Batra, Ramin Okhrati, and Raad Khraishi. How personality traits shape llm risk-taking behaviour. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21068–21092, 2025.

[16] Ryuji Hashimoto, Takehiro Takayanagi, Masahiro Suzuki, and Kiyoshi Izumi. Agent-based simulation of a financial market with large language models. *arXiv preprint arXiv:2510.12189*, 2025.

[17] Thomas Henning, Siddhartha M Ojha, Ross Spoon, Jiatong Han, and Colin F Camerer. Llm trading: Analysis of llm agent behavior in experimental asset markets. *arXiv preprint arXiv:2502.15800*, 2025.

[18] John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

[19] Allen H Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023.

[20] Bruce I Jacobs, Kenneth N Levy, and Harry M Markowitz. Financial market simulation. *Journal of Portfolio Management*, pages 142–152, 2004.

[21] Andrei Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017.

[22] Tong Li, Hui Chen, Wei Liu, Guang Yu, and Yongtian Yu. Understanding the role of social media sentiment in identifying irrational herding behavior in the stock market. *International Review of Economics & Finance*, 87:163–179, 2023.

[23] Yang Li, Yangyang Yu, Haohang Li, Zhi Chen, and Khaldoun Khashanah. Tradinggpt: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance. *arXiv preprint arXiv:2309.03736*, 2023.

[24] Xiao-Yang Liu, Hongyang Yang, Qian Chen, Runjia Zhang, Liuqing Yang, Bowen Xiao, and Christina Dan Wang. Finrl: A deep reinforcement learning library for automated stock trading in quantitative finance. *arXiv preprint arXiv:2011.09607*, 2020.

[25] Thomas Lux and Michele Marchesi. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397(6719):498–500, 1999.

[26] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[27] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

[28] George Soros. *The alchemy of finance*. John Wiley & Sons, 2015.

[29] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[30] Yuzhe Yang, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. Twinmarket: A scalable behavioral and social simulation for financial markets. *arXiv preprint arXiv:2502.01506*, 2025.

[31] Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Jordan W Suchow, Denghui Zhang, and Khaldoun Khashanah. Finmem: A performance-enhanced llm trading agent with layered memory and character design. *IEEE Transactions on Big Data*, 2025.

[32] Chong Zhang, Xinyi Liu, Zhongmou Zhang, Mingyu Jin, Lingyao Li, Zhenting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, et al. When ai meets finance (stockagent): Large language model-based stock trading in simulated real-world environments. *arXiv preprint arXiv:2407.18957*, 2024.

[33] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[34] Yuhang Zhou, Yuchen Ni, Xiang Liu, Jian Zhang, Sen Liu, Guangnan Ye, and Hongfeng Chai. Are large language models rational investors? *CoRR*, 2024.