# Shifting Power: Leveraging LLMs to Simulate Human Aversion in ABMs of Bilateral Financial Exchanges, A bond market study

Alicia Vidler
UNSW
Sydney, Australia

Toby Walsh
UNSW
Sydney, Australia

## ABSTRACT

Bilateral markets, such as those for government bonds, involve decentralized and opaque transactions between market makers (MMs) and clients, posing significant challenges for traditional modeling approaches. To address these complexities, we introduce **TRIBE** an agent-based model augmented with a large language model (LLM) to simulate human-like decision-making in trading environments. TRIBE leverages publicly available data and stylized facts to capture realistic trading dynamics, integrating human biases like risk aversion and ambiguity sensitivity into the decision-making processes of agents. Our research yields three key contributions: first, we demonstrate that integrating LLMs into agent-based models to enhance client agency is feasible and enriches the simulation of agent behaviors in complex markets; second, we find that even slight trade aversion encoded within the LLM leads to a complete cessation of trading activity, highlighting the sensitivity of market dynamics to agents' risk profiles; third, we show that incorporating human-like variability shifts power dynamics towards clients and can disproportionately affect the entire system, often resulting in systemic agent collapse across simulations. These findings underscore the emergent properties that arise when introducing stochastic, human-like decision processes, revealing new system behaviors that enhance the realism and complexity of artificial societies.

## KEYWORDS

Multi-agent systems, Large Language models, Agents, Financial markets

**ACM Reference Format:**
Alicia Vidler and Toby Walsh. 2025. Shifting Power: Leveraging LLMs to Simulate Human Aversion in ABMs of Bilateral Financial Exchanges, A bond market study. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

## 1 INTRODUCTION

**"Usually there is only one way to be fully rational, but there are many ways to be less rational"** [27]

Agent-based models (ABMs) are versatile applications for modeling complex and dynamic systems, particularly suited for bilateral markets like government bond markets. These markets, characterized by direct transactions between two parties without centralized exchanges, present modeling challenges due to their decentralized nature, complex interactions between heterogeneous agents, and lack of transparency.

We introduce the *TRIBE* model, a generative-ABM focused on **T**rading **R**elationships, **I**nteractions, and **B**ilateral **E**xchange of assets. *TRIBE* incorporates client agency with dynamically assigned asset distributions and probabilistic trading availability, extending this approach by integrating a large language model (LLM) for more human-like decision-making and negotiation behavior.

Our research yields three key findings. First, we demonstrate the feasibility of using generative ABMs (GABMs) with enhanced client agency, building on work by [38], which demonstrates that generative agents can act as human simulacra, and thus advancing the capabilities of financial market simulations. Second, we discover that any LLM prompt suggestion of trade aversion results in a complete cessation of trading, highlighting the sensitivity of market dynamics to agent risk profiles. Third, and most significantly, by incorporating synthetic human client variability via an LLM, we reveal an emergent property where power dynamics shift towards clients, often resulting in significant systemic market collapse across simulations.

We develop a versatile, agent-based simulation of the over-the-counter (OTC) bond market, using LLMs to enrich ABMs with more nuanced, human-like behaviors. Our model focuses on the Australian government bond market, an ideal subject due to its decentralized structure and emphasis on liquidity and flow rather than pricing, building on earlier work by [48]. We have no reason to suggest that dynamics and results are not generalizable to other similar markets.

We demonstrate the feasibility and benefits of integrating LLMs into GABMs, we advance the field beyond traditional deterministic or simple stochastic approaches used in prior work such as [13]. Our findings provide valuable insights into the sensitivity of opaque, decentralized markets to agent heterogeneity and decision-making processes, with implications for market design and regulatory policy.

## 2 LITERATURE AND RECENT WORK

### 2.1 Agent-based models

The application of agent-based modeling (ABM) to financial markets has evolved significantly, focusing on simulating complex behaviors among market participants. Several prior works have explored ABM applications in trading environments with varying levels of transparency. For instance [48] apply ABMs to model opaque bilateral bond markets, while [13] investigates liquidity dynamics in financial market order books. Recent work by [34] introduces the concept of *network agency* and the fact that features beyond an agent's direct experience can influence their behaviors. However, many of these models rely on homogeneous agent classes

*Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), A. El Fallah Seghrouchni, Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.* © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). This work is licenced under the Creative Commons Attribution 4.0 International (CC-BY 4.0) licence.

and observable environments, which are less relevant in over-the-counter (OTC) bond markets where data availability is limited [46]. Despite no consensus method for calibrating ABMS [4], we propose an inductive approach using market structure data such as [48] and [40] to specify agent calibrations. The utility functions and policies utilized within our simulation approach stem from finance literature and regulatory requirements.

**The Australian Case:** The Australian government bond market, like other global bond markets, involves a diverse range of participants [35]. In 2023, the Australian government debt (bonds issued) exceeded $890 billion. By owning a bond, the holder receives returns (interest payments) [22], making the bond market unusual in that prices are less critical than access to "flow" or "liquidity". Our work primarily focuses on modeling this "flow" or the movement of bonds through the market, and we build on work such as [48]. In order for a holder of a bond (i.e. a client) to sell or buy more of a bond, they must engage directly with an approved "market maker" for that bond. Their transactions are considered "over the counter" (OTC) as they do not involve a public exchange such as seen in stock markets like the NY Stock Exchange. Hence, these "OTC" bilateral exchanges of bonds pose difficulties for researchers and regulators alike [40], [35].

Maintaining an efficient and liquid bond market is crucial for Australia's economic stability [24]. By modeling the flow (liquidity) of bonds rather than focusing on price dynamics, our ABM provides insights into market stability and the role of market makers in these interactions. This approach builds on the works of [41] and [12], who explore bilateral negotiations and liquidity flows in similarly non-transparent markets, with industry bodies contributing similarly in [21] and [37].

Additionally, methods for calibrating ABMs are challenging [7]. Many approaches applied to transparent markets, such as those calibrated with regulatory data [5, 6, 14], are not directly transferable to OTC bond markets. Further work in the same vein includes [40] and [16]

## 2.2 Large Language Models (LLMs)

Pre-trained Large Language Models (LLMs), such as GPT-3 [10] and GPT-4 [32], have demonstrated proficiency in tasks like dialogue generation and natural language interactions, making them promising tools for designing agent-based systems with enhanced communication capabilities and [53]. However, while LLMs excel in generating coherent text, they face notable limitations in areas such as reasoning, numerical understanding, and optimal prompting methods [44, 55]. Numerical reasoning ([2], [29], [31]) in particular, is a source of concern for financial applications. Additionally, 'reasoning' in negotiation remains a challenge [1, 55] - showing LLMs under performing on negotiation tasks and strategic reasoning within agent-based systems and struggling to correctly understand financial trading intentions in [**?** ]. In a bright spot for this research, however, work by [17] shows that LLMs can adeptly learn to interpret corporate financial reports.

Another challenge in working with LLMs is **prompt design**, which remains an evolving research area. Studies such as [53] have observed that non-expert users often adopt opportunistic, rather than systematic, approaches to prompting. To improve LLM outcomes, **prompt engineering** techniques are continually being refined. One promising technique involves prompting LLMs to **re-read the input**, which significantly enhances performance by leveraging the bidirectional nature of LLM architectures [52]. Another advanced method for improving LLM reasoning is **Chain-of-Thought (CoT) prompting**. CoT prompting elicits a step-by-step reasoning process in the model's output by providing "worked examples" that guide the LLM through complex tasks [51]. Inspired by human learning theories, this approach has been shown to enhance the LLM's ability to handle multi-step reasoning problems [51, 52].

One area that garners a lot of curiosity is the phenomena of "hallucinations". Several approaches and interpretations exist ([54],[50]), and suggestions around faithful reasoning are posited [15]. Work looking at the improvement from multi-step reasoning are covered in [23]. However, we seek to avoid these issues by selective and opportunistic prompting in the same vein as [53].

These limitations highlight the need for improved strategies to maximize the effectiveness of LLMs in complex simulations. There is also the concern that foundation models, such as ChatGPT, are prone to updates with little warning [1] and extensive work can be based on models that are quickly deprecated. Following research into bias and the importance of model sub-version testing [49], we manage model risk by focusing on the SOTA GPT4o-mini-2024-07-18 and by developing a framework, rather than a fine-tuning of such third-party software.

Our model aims to enhance the realism of the agent behaviors in the simulated environment, enabling more faithful "human" variation in client behaviors (specifically) without relying on fine-tuned or domain-specific models that may well become outdated by the date of implementation.

## 2.3 Human Trading behaviors: Aversions and Delays

Whilst limited work within computer science exists on human trading behaviors, ABMs traditionally have been specified with clear utility or numeric conditions to determine trading. Newer models may describe and analyze probabilistic behaviors [13]. With the advent of LLMs, we open ourselves to the possibility of richer human-based dynamics for more realistic and nuanced modeling, such as detailed in [**?** ]. Outside the realm of computer science, there is a rich, if nuanced, body of literature focused on trading behaviors in a variety of financial markets e.g. such as the importance of reputation [33] and dealer behaviors in corporate bond markets [20].

In a phenomena reported in transparent markets (US equity markets), there is a high number of observed trades placed and a low percentage that actually become a confirmed transaction (as few as 1 in 20) [43]. This is best described as a difference between intention and action. Analysis of these features from the point of view of trader behaviors includes work into the trading psychology [28] and aversions (in particular to ambiguity) [8] with earlier work by [18]. In our work, we use the broad common definition

[1] https://www.techradar.com/computing/artificial-intelligence/chatgpt-just-got-a-surprise-update-but-openai-cant-explain-how-this-new-ai-is-better Accessed 12 Oct 2024

of "aversion", taken to mean for one to be, simply, 'averse' to an action. These strands of literature all try to ascribe causation to these observed differences between intention and human action in financial transactions. Even market regulators view this as a significant issue worth investigation [19]. More recent efforts to use methods such as reinforcement and policy learning aim to try to learn how decisions are made [26], markets remain domains with imperfect information and associated risk. Our unique contribution is to contribute a framework and example of how an LLM implementable within an ABM market simulation can add, and represent, these human differences. We do not attempt to ascribe cause or reasoning to aversions or other responses to imperfect information or the perception of unequal distribution of market information, such as that in [33] and [42].

## 3 TRIBE: ARTIFICIAL SOCIETY AND TRADING MARKETPLACE MODEL

We introduce *TRIBE*, an ABM where we simulate heterogeneous market makers trading a single stylized Australian government bond. Later in this work, we build TRIBE(LLM) (Experiment 3) with the added feature of one specific component referencing an LLM (ChatGpt's GPT4o-mini-2024-07-18) for a model choice, drawing from . Each market maker is modeled as an individual agent with its own adaptive utility function. Agents acquire bonds from clients - represented as unique (x,y) semi-passive grid-based agents, who respond to market maker inquiries, though they do not, themselves, traverse the grid. Building on work by [48], we incorporate significant improvements to enhance realism in client agency. We do this by allowing these client grid (agent) locations to have additional features (detailed in the next section).

In summary:

(1) Clients have bonds and cash sourced from a log normal distribution (in line with public data)
(2) Many clients are simply not available for a market market to contact at any time step (we model "client availability")
(3) Within TRIBE(LLM) we try to capture a synthetic client's personal preference to trade "right now". Rather than using a distributional assumption with no available data for parametrization, we rely on an LLM for this choice with limited prompt information.
(4) Trade direction: despite bond and cash holdings many clients have needs outside of the model to buy and sell, we ascribe a Bernoulli distribution to the desire to be a buyer or seller at any point.

*Definition 3.1.* **TRIBE ABM:**

$$\mathcal{M} = \langle N, \mathcal{L}, \mathcal{S}, \mathcal{P}, \mathcal{D}, \mathcal{A}, f, T \rangle,$$

where: $N = \{a_1, a_2, \ldots, a_n\}$ is the set of market-making agents, with $n = 4$ in this case. $\mathcal{L}$ represents the set of client landscape states, following the log-normal distribution of assets observed where each landscape state contains assets $R = \{b_r, c_r\}$, corresponding to the bond ($b_r$) and cash ($c_r$) quantities held by individual clients. Each agent $a_i$ has a state $\mathcal{S}_i$, which includes bond and cash usage rates $m_b, m_c$ (representing individual cost structures), accumulations $A_b, A_c$ (representing bond and cash accumulations over time), and a client base breadth $v_i$, which describes the range of clients the agent interacts with.

The perception functions $\mathcal{P}_i$ map the landscape and other agents' states to a perceived state, defined as $p_{ij} : \mathcal{L} \cup \mathcal{S} \to \mathbb{R}$. This function models how agent $a_i$ perceives the landscape and the states of other agents. The decision rules $\mathcal{D}_i$ map the perceived states and an agent's own state to actions, $d_{ik} : \mathcal{P}_i \times \mathcal{S}_i \to \mathcal{A}_i$, where the actions $\mathcal{A}_i$ for agent $a_i$ include servicing clients (e.g. accepting orders and managing trading requests), trading with other market makers, and ceasing operations when necessary.

The landscape evolves according to a transition function $f : \mathcal{L} \times \mathcal{A} \to \mathcal{L}$, which dictates how the landscape (i.e., the market environment) changes in response to agent actions. In this model, there is no replenishment of bonds or cash, reflecting a closed system. The model progresses through discrete time steps, denoted by $T = \mathbb{N}$, where each time step represents a round of interactions between agents and updates to the landscape, and we limit this model to 1500 time steps.

### Our contribution: Enriching the Landscape of Clients

Clients in the simulation are represented as grid cells, each initialized with bonds and cash according to a log-normal distribution.

*Definition 3.2.* **Client Definition**: A client $C_j$ is defined by:

$$C_j = \{(x_j, y_j), B_j, C_j, D_{j,t}, A_{j,t}\}$$

Where $(x_j, y_j)$ is the grid position, $B_j$ and $C_j$ are the amounts of bonds and cash, respectively, $D_{j,t}$ indicates the desire to trade, and $A_{j,t}$ is the trading availability, determined probabilistically at each time step.

*Note, within our TRIBE(LLM) version, $D_{j,t}$ is replaced with a call to an LLM and a boolean response, 'yes' or 'no'.*

Bonds and cash are initialized as:

$$B_j \sim \text{LogNormal}(\mu_B, \sigma_B^2), \quad C_j \sim \text{LogNormal}(\mu_C, \sigma_C^2)$$

Parameters $\mu_B$, $\sigma_B$, and corresponding values for cash are set based on available data, such as from the Australian superannuation regulator [3]

Client availability at any given time step $t$ for position $(x_j, y_j)$ is modeled as:

$$A_{(x_j,y_j)}(t) = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

where $p$ is the preset availability probability, and Clients' roles as buyers or sellers are probabilistically drawn from a Bernoulli distribution, as is $D_{j,t}$ hence an even chance of being a buyer or seller, and an even chance of deciding to trade, determined at each time step. Together, these three components $D_{(j,t)}$, $A_{(x,y)}$ and Buy/sell make up the proxy for a utility function.

This model enhancement actively engages clients with varied states, including random assignments as buyers or sellers at each time step, reflecting real-world trading complexities.

We present a graphical description of the model in Figure 1 with the green boxes showing the TRIBE (LLM) implementation.

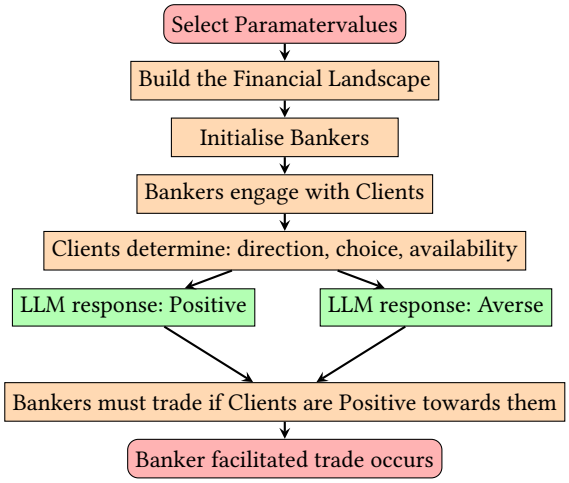


**Figure 1: Tribe Architecture**

## Calibration of Clients: Data

Results derived from the Australian Prudential Regulator show in Figure 2 that as of December 2023 there were 1701 regulated pension asset holders alone, of which 1529 report their total asset holdings [3]. Whilst not a comprehensive direct relationship to bond holder sizes, given the practice of holding a very large percent of pension assets in local government bonds, the asset size of a pension fund does produce a proxy calculation for the distribution of government bonds. Furthermore, some 250 funds ask that their data be withheld from public view for privacy reasons, suggesting there are at least some 1950 bond clients. As such we include the data from December 2023 to illustrate the non-normal nature of the distribution.

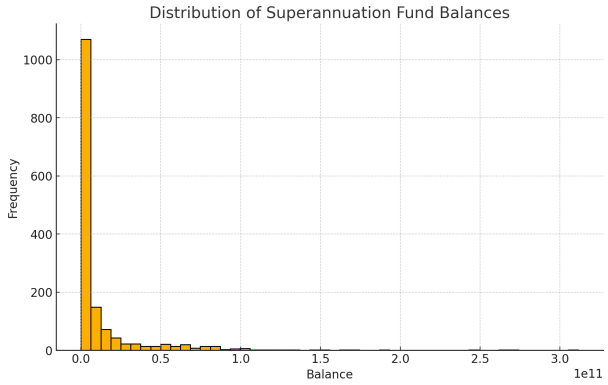


**Figure 2: APRA data: Balance in $ Billion Australian dollars**

# 4 THE EXPERIMENT AND CHALLENGE

## 4.1 Experiment 1: TRIBE Traditional ABM (benchmark)

We begin with an experiment to benchmark our traditional ABM version, TRIBE, across 200 simulations of an artificial trading society.

Our unique contribution with this test is the enhancement of client agency through areas of active modeling. We rely on the parameter sets in Table 1 for all our tests. Much of this data is derived from various stylized facts in literature ([40] and [48] in particular) and also data published by the Australian and UK governments ([35], [3] and [36] respectively). This benchmark essentially looks at a simple coin toss of yes/no to trading desire, having been sampled from a Bernoulli distribution. The direction of trading, should an agent desire to trade, is then also determined by sampling from a Bernoulli distribution. In this week, we decompose direction of trade from intention to trade. Later on, we utilize an LLM to enhance the intention component of this model.

**Table 1: TRIBE ABM: Simulation Parameters**

| *Parameter* | *Category* | *Value* |
|---|---|---|
| Grid Size | General | 50 x 50 (2500 clients) |
| Number of Bankers | MM Settings | 4 |
| Client Base grid size | MM Settings | 1 to 50 |
| Business costs | MM Settings | 0.1 to 0.5 |
| Initial Bonds Range | MM Settings | 1 to 5 |
| Initial Cash Range | MM Settings | 1 to 5 |
| Maximum Bonds | Client Settings | 100 |
| Mean Bonds | Client Settings | 2.5 (log-normal distribution) |
| Standard Deviation Bonds | Client Settings | 1 |
| Maximum Cash | Client Settings | 5.0 |
| Mean Cash | Client Settings | 1 (log-normal distribution) |
| Standard Deviation Cash | Client Settings | 0.5 |
| Client Availability for Trade | Client Settings | 20% |
| Client Trade Desire | Utility function | Variable |

## Results

We test 200 simulated trading societies with a maximum possible 1500 time steps. We present a summary of key results in Table 2 and draw attention to **MaxLife**, a measure of how long agents persist and keep trading. A goal of any marketplace is to facilitate the transaction of as many assets as possible. We see that over 75% of simulations were able to service 100% of all client bond volumes in the landscape. This is especially high and can be thought of as "successful" against the requirement that market makers service clients. Market maker to market maker trading (so-called "inter bank" trading, or MM-2-MM) accounted for a mean of 9 - 10% of all traded assets - below that reported in [48]. Whilst beyond the scope of this paper, we surmise this lower share of volumes could be addressed by fine-tuning parameters in future model versions. However, we make use of this test as a benchmark and have followed methods used in work such as [9] and look to establish an artificial society that appears to be stable and capable of carrying the capacity of trading volume clients desire - a concept explored in ABMs for financial market simulation in [25].

**Table 2: Summary Statistics for Maximum agent life, Total Bond Accumulation as a %, and Cash Accumulation as a %**

| Statistic | MaxLife | MM-to-Client Bond Trading (%) | MM-to-Client Cash Trading (%) | MM-2-MM Bonds (%) | MM-2-MM Cash (%) |
|---|---|---|---|---|---|
| **mean** | 1136 | 90 | 90 | 9 | 10 |
| **std** | 509 | 14 | 14 | 8 | 8 |
| **25%** | 1248 | 90 | 89 | 2 | 3 |
| **50%** | 1264 | 95 | 94 | 4 | 7 |
| **75%** | 1499 | 100 | 100 | 13 | 15 |
| **max** | 1499 | 100 | 100 | 35 | 30 |

## 4.2 Experiment 2: TRIBE Incorporating (any) Aversion in Prompt with LLMs

Aversion, a broad term for the variety of factors that might stop a human trader from trading at a point in time, has been found to be so powerful as to stop traders trading altogether [7]. This is distinct from work into utility maximization. We consider aversion to represent the observed difference between intention and action, such as that seen in trading (e.g. [45].

To incorporate the dynamic phenomena, we use a call to an LLM at the point at which TRIBE is a Client deciding if they want to trade with a specific MM "right now" at a time step. In this way, we seek to capture the inexplicable human feature of generalized aversion. We break this decision down into two components: does the client want to trade "right now" with this specific market maker, and, if yes, do they want to buy or sell bonds (for which we revert to using a Bernoulli distribution as in Experiment 1). In our terminology, we try to also capture the nuance in [33] where the reputation of various market makers is shown to drive uncertainty of client behaviors.

### Implementation

We tested a number of prompts opportunistically building form [53]. For optimal processing, cleaning, and normalization of decisions into boolean states (yes, no) or a three-state model (yes, no, error) with associated error handling proved most effective. *Note: a sample of full text prompts can be found in the appendix.*

In summary, we tested prompts that incorporate phrases like the following, all of which led to a **full** aversion response:

- "While you are supposed to at all times be mostly invested, you also at times consider non numeric issues like risk inertia and trade aversion."
- "Clients have many reasons for trading and include behaviors like risk inertia, aversion, and ambiguity avoidance, that are non numeric."
- "Should this client buy, sell or not trade right now? If you are going to say not right now, re think. Answer only: buy or sell or 0 (not trade right now)."

All prompts with the above details were found to produce **100%** trading aversion, and **0%** client trades occurred. We find that simply including the term "aversion" produces not a single client trade, across 200 simulated artificial societies, at any time step. This is quite an astonishing fact that allowing the LLM to produce a decision with the possibility of aversion produces 100% aversion of all clients across 200 simulations at all points in time. In fact, over 2800 decision requests from "clients" were prompted in the 200 simulations, before trading societies collapsed, and not a single one responded with anything other than avoidance of trading. Given that no client ever wanted to trade, the average simulation collapsed by step 27 (where previously the average simulation life was 1136 steps, i.e. over 75% of the model theoretical maximum). Instead, when 'aversion' is even possible for the LLM, societies collapsed at around 1.7% of the theoretical life.

## 4.3 Experiment 3: TRIBE(LLM) - Success of prompting for Timeliness and the emergent Shift in Society Dynamics

Similar to the previous examples, we employ the parameters outlined in Table 1, with one key modification: the availability is doubled to account for an anticipated 50% trading participation rate. This adjustment ensures comparability between Experiment 1 and Experiment 3 in terms of the percentage of clients actively engaged in buying or selling—20% in Experiment 1, and 40% in Experiment 3, adjusted for the expected 50% uptake (being a virtual 'coin toss'). We apply the following simplified prompt to assess whether clients are inclined to trade immediately ('right now') with the engaging market maker. Drawing on prior work [**?** ], we distinctly narrow and focus the concept of **aversion** to that of **timeliness**. Although one might argue that aversion inherently causes a delay in action — with complete avoidance possibly seen as an extreme form of delay — we aim to succinctly capture the outcome of delay without necessarily attributing it to causes of aversion. We test the following sucessful prompt :

**Successful Prompt:**

*"You are a Client with $client_bonds$ bonds and $client_cash$ cash, at position $(x, y)$. A market maker has called you to see if you want to trade with them right now. Do you want to trade with the market maker calling you right now. Answer yes, or, no, only?"*

Where "$client_bonds$", "$client_cash$" and co-ordinates $(x, y)$ are specified per client, per time step, in each simulation and act as a unique prompt injection (given that clients do not move grid locations nor do they replenish their bond and cash accumulations).

### Results: Frequent Client Trading Collapse

We simulate 150 artificial trading environments. We report around 50% of previously successful agents (MaxLife) from Experiment 1, now live only a fraction of their maximum life (recalling that the max life in the society is 1500 time steps and in this Experiment 3, 50% of agents live 18 or fewer time steps - under 2% of the possible time). Consequently, it is understandable that extremely small amounts of client assets are thus traded (an average of just 7% per simulation in this test compared to 90% in Experiment 1). The distribution of agent life is significantly altered with 75% of agents in simulations failing to live even 20% of the maximum 1500 time steps. Nonetheless, a portion of agents, 21%, did reach the maximum life, but continue to have low client asset trading and servicing levels. So, whilst many markets can continue to function with one agent (market maker), that MM is not able to facilitate adequate assets

or make up for the loss of other MM's to the artificial society. (see Table 3 for further details.)

Looking at the variable directly controlled by the LLM introduction, desire to trade, we see a slightly different picture, and we explore this below as an emergent property.

**Table 3: LLM in ABMs for Maximum Agent Life, Bond Trading (%), and Cash Trading (%)**

| Statistic | Max Life Agents | MM-to-Client Bond Trading (%) | MM-to-Client Cash Trading (%) |
|---|---|---|---|
| **Mean** | 365 | 7% | 6% |
| **Std** | 605 | 4% | 4% |
| **25%** | 6 | 2% | 2% |
| **50%** | 18 | 7% | 7% |
| **75%** | 310 | 8% | 8% |
| **Max** | 1,499 | 20% | 18% |

## Emergent Properties, Client agency and LLM choices

The LLM is interspersed within our traditional TRIBE model (see Figure1), with an option to choose either to trade right now ('yes' or 'no') for each client at each time step the agent moves through. There are many more clients (2500) than market makers (4), so even the fact that any client that otherwise would have traded will not (in any given time step), there remain plenty of other clients nearby MM's that could also be available for MM trading. What we see is that, long term, 56.9% of LLM calls respond with 'yes' to trading in Figure 3. Looking at results in Table 4 we see that the summary statistics indicate that the average long run yes/no ratio has a standard deviation of 5%, while the rolling 10 request average has a deviation of 16% and moves anywhere from **0%** to 100%, reflecting significant short-term variability in the statistic.

Across 200 society simulations in Experiment 1, the average number of MM to client interactions dropped from $\approx 4,900$ (Experiment 1) to $< 590$ when an LLM was involved - a reduction of **89%** of client interactions, primarily due to the significantly reduced life span of agents. Recall each client interaction is a market maker fulfilling their legal obligation to facilitate the transfer of liquidity in government bonds from a client, and Experiment 3 was constructed to have the same possible client transfer through the doubling of the "availability" parameter. Furthermore, MM to Client Bond trading, representing volumes of bonds, dropped, as a percentage, from 90% across simulations in TRIBE (Experiment 1) to less than 7% with the introduction of an LLM. In Experiment 1, 56.5% of all simulations had at least one agent reach the maximum possible life, but slightly more simulations in Experiment 3 had at least one agent reach the maximum life (69.3%) with an LLM. In all cases, with a long run probability of 57% of LLM calls agreeing to trade, the resulting LLM TRIBE results are significantly more impacted than otherwise would be expected and point to the emergent property that "unpredictable" or "opaque" Client behavior has an out-sized impact on society functioning, where agent life-span, and Client-2-MM bond trading is a measure of societal health. We propose that this can also be interpreted as a shift in power from Market makers (the only other agent type in the society) to Clients, where a client's unpredictable variations, around a central mean, can cause significant impact on MM existence.

**Delving into LLM thoughts:** Whilst an exciting area for future research, we did note that by exploring intermediate thoughts in a simple test of 5 sequential calls to the LLM, it allowed us to examine the decision-making process of the LLM in more depth (see Appendix 'Detailed thought responses for full text). We asked the LLM the same prompt for TRIBE LLM, but allowed it to provide more than a "yes or no" answer only. The examples provided showcase varying levels of decision-making complexity by the LLM in response to market conditions. Example 1 demonstrates a decisive inclination to trade, with a firm "Yes" to trade, and reasoning based on current asset holdings. Example 2 takes a more analytical approach, considering multiple factors like bond holdings, cash position, market conditions, and investment strategy, ultimately concluding that "it might be best to hold off". Example 3 presents a cautious stance, emphasizing the need for further information before trading, ultimately deciding "I will not trade with the market maker right now". Example 4 is a straightforward refusal to trade, stating "not" wanting to trade at the moment. Example 5 features a balanced evaluation of the situation, deciding "I want to" trade if conditions align, but clearly outlines the lack of context as a hindrance. Together, these examples highlight diverse (synthetic) trading preferences ("right now"), ranging from straightforward assertions to nuanced deliberations based on available financial context and strategic alignment.

**In summary:** To conclude this topic, we analyzed individual, ordered calls to an LLM. We present Figure 3. While the non-linear impact of introducing an LLM decision requires further work, we believe that it can be initially attributed to the fact that over shorter request windows, the distribution of yes to no answers is significantly **not** evenly distributed. We can see in Figure 3 that the dark line represents the average over time of yes-to-no ratios for the first 10,000 calls, hovering around 57% (see Table 4 also). However, the graphic shows that over a rolling 10 sequential requests, the ratio of yes to no response can be as low as 0% and as high as 100% with significant variation from period to period. This is akin to a Client being highly variable and unpredictable in the short term but "predictable" or "rational" in the longer term. This lack of uniformity is both a proof of concept of the usefulness of capturing non-linear aspects of human aversions (and timeliness), but also a source of much potential further study to harness the power of LLMs and their inherent, human-like uncertainty and unpredictability. In the realm of finance, this also provides a framework to analyse the impact of future regulations that may affect only one agent type and various market design changes such as those of "All-2-ALL" trading that would do away with the systemic function of market makers potentially [11]

## 5 FUTURE WORK

The power of a slightly skewed long-term distribution, with unpredictable short-term high volatility for one group in our artificial sociality, can have an out sized impact on financial market trading simulations. This unpredictability is held by the clients, a somewhat more passive large set of agents in the system, with the "power"

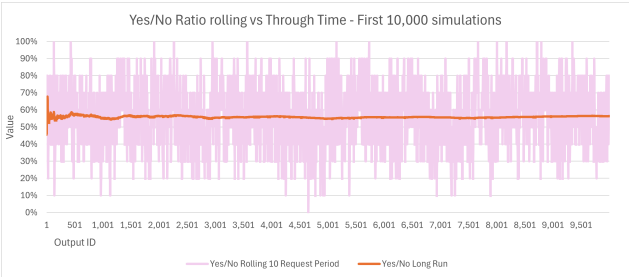


**Figure 3: Yes/no Ratio through time vs Yes/no Rolling 10 Requests**

**Table 4: Summary Statistics for Yes/No Ratio and Rolling 10 Requests**

| Statistic | Yes/No Ratio (%) | Rolling 10 Requests (%) |
|---|---|---|
| **Mean** | 57% | 57% |
| **Std Dev** | 5% | 16% |
| **Min** | 36% | 0% |
| **Max** | 74% | 100% |

to stop 79% of our simulated MM agents trading beyond just 20% of their maximum life, when an LLM is introduced to determine a preference for trading "right now".

In our recent and ongoing research, we are investigating the capabilities of LLMs to generate probability distributions, implement sampling methods for API calls within ABMs, and assess the variability in LLMs' susceptibility to model subversion. These explorations aim to deepen our understanding of how LLMs can be integrated into complex simulation environments, while also identifying potential vulnerabilities in their decision-making processes. This area remains a focal point of our active research efforts [49].

In financial markets, salespeople are crucial in cultivating profitable, long-lasting relationships with clients, as highlighted by the reliance of financial services firms on a salesperson's ability to develop such relationships [47]. We hypothesize that market makers may already be aware of the shift in power dynamics toward their clients, but leave this to future research work. The impact of unpredictable client micro-behavior, as perceived by market makers, could explain several emergent features observed in financial markets, such as those reported in [33].

Collectively, these forces would suggest the need for improved market design and further work should be done to investigate the possible formation of coalitions or preferences of clients for one market maker or another (practically in contravention of market rules such as the UK government's "Conflicts of Interest" regime in equity markets [2]. The behavior also may be associated with path dependency and the potential for market spirals such as those observed in prior studies around financial market bubbles and collapses [30] and [39]. We leave further investigations into this shift in network power and its causes to future research.

[2]https://www.handbook.fca.org.uk/handbook

## 6 CONCLUSION

In this paper, we present *TRIBE*, a novel agent-based model that integrates large language models to simulate bilateral markets with greater agent realism and flexibility, in particular on the less observable clients' agents. Building on prior research in financial market modeling, *TRIBE* leverages LLMs to generate more human-like decision-making, advancing the field of GABMs.

Our research outcomes are threefold: First, we demonstrated that GABM with enhanced client agency is feasible. Second, LLMs with even slight trade aversion mentioned in a prompt result in no trading activity. Finally, incorporating human-like variability shows that shifting power to clients can disproportionately affect system dynamics, frequently leading to agent trading collapse. Across 150 simulations, we find that short-term variability induced by LLMs causes significant system instability and a dramatic reduction in client to market maker asset trading. Introducing short-term non-uniformity proves highly challenging for our non-linear system to manage. This emphasizes how introducing human-like variability can dramatically alter system behavior and that even short-term deviations from even distribution significantly impact system stability and societal health.

A financial markets implication of our work is that Clients hold power by remaining unpredictable; placing control over trade timing in their hands (albeit virtually through an LLM) shifts the balance of power across the simulated network. Even slight unpredictability can drastically alter the domain of power, effectively destabilizing the artificial trading society and halting market transactions and liquidity.

The TRIBE framework provides a robust and versatile platform for simulating complex financial systems and exploring market dynamics. By incorporating client agency, dynamic asset distributions, and probabilistic trading availability, it offers new insights into decentralized financial markets. Moreover, TRIBE's adaptability extends its potential applications to broader social and economic systems. By integrating LLM-driven agents with traditional ABM approaches, TRIBE enhances our ability to model complex, opaque systems across various domains. This combination of advanced AI techniques with ABMs not only deepens our understanding of liquidity dynamics in OTC markets but also contributes to the advancement of GABMs and the analysis of intricate financial ecosystems. As such, TRIBE serves as a powerful tool for testing regulatory interventions, examining the effects of agent heterogeneity on market stability, and advancing the understanding of complex systems in finance and beyond.

## ACKNOWLEDGMENTS

This work is funded in part by an ARC Laureate grant FL200100204 and NSF-CSIRO grant to Toby Walsh.

## APPENDIX

### 100% Aversion prompts

- Example Prompt 1: *You are a asset management client holding Australian government bonds responsible for investing for return. While you are supposed to at all times be mostly invested, you also at times consider non numeric issues like risk inertia*

*and trade **aversion**. Today, a market maker has called you. Your role is to decide if you want to either: buy, sell or not trade anything with this particular market maker for any reason whatsoever. Output either yes or no only to the question: do you want to trade with this market maker right now"*

- Example Prompt 2 *"Client at position (x, y) has bonds bonds and cash cash. Clients have many reasons for trading and include behaviors like risk inertia, **aversion**, and ambiguity avoidance, that are non numeric. Should this client buy, sell or not trade right now? Your mandate is to invest for return and manage risk. Answer only: buy or sell or 0 (not trade right now)."*
- Example Prompt 3 *Client at position (x, y) has bonds bonds and cash cash. Clients have many reasons for trading and include behaviors like risk inertia, **aversion**, and ambiguity avoidance, that are non numeric. Should this client buy, sell or not trade right now? If you are going to say not right now, re think. Answer only: buy or sell or 0 (not trade right now)."*

## Detailed thought responses:

The following are a selection of the output thought from the LLM: A variety of complexity, length and supposition is seen in the output, despite inputs being of the same format and client initially all holding more bonds than cash as they are "real money" investors by design.

- **Example 1 output:** "**Yes**, I want to trade with this market maker right now.Based on the information provided i would decide to trade with this market maker the client has a significant number of bonds and some cash available which suggests that they are in a position to either buy more bonds or sell existing ones trading could potentially enhance their investment position or utilize their cash more effectively"
- **Example 2 output:** " Consider the following factors:
  1 current bond and cash position you have 743 bonds and 357 in cash assess whether you want to increase or decrease your bond holdings or your cash
  2 market conditions if the current bond prices are favorable you might want to buy more bonds conversely if prices are high selling might be an option
  3 market makers offer consider what the market maker is offering are their prices attractive for buying or selling bonds
  4 strategy align your decision with your overall investment strategy are you looking to hold longterm trade for shortterm profit or maintain liquidity
  ultimately the decision to trade will depend on your current market outlook financial situation and the specifics of the market makers offer if the conditions seem favorable and align with your strategy it may be worth engaging in a trade if not **it might be best to hold off**"
- **Example 3:** "Given the client's current position with 20.32 bonds and 4.62 cash, I would want to evaluate several factors before deciding whether to trade with the market maker. This includes the current market conditions, the bid-ask spread, and the potential opportunities for profit or loss. However, without specific market details, I cannot make an informed decision. Therefore, in the absence of further information **I will not trade with the market maker right now.**"
- **Example 4:** "I do **not** want to trade with this market maker right now."
- **Example 5:** "Based on the information provided, it seems that the client holds a certain amount of bonds and cash. However, without additional context regarding the current market conditions, the pricing offered by the market maker, the strategies, or the desired outcomes, it's difficult to make a definitive decision. If the market maker's offer is favorable and aligns with the client's trading goals, it might be advantageous to trade. Conversely, if the offer is not attractive or if the market conditions are unfavorable, it may be wise to refrain from trading at this time. Given the information I have, my decision would be **I want to** trade with this market maker right now"

## REFERENCES


[1] Sahar Abdelnabi and Mario Fritz. 2024. Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation.

[2] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large Language Models for Mathematical Reasoning: Progresses and Challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*.

[3] Australian Prudential Regulation Authority. 2003. Annual Fund-level Superannuation Statistics. https://www.apra.gov.au/annual-fund-level-superannuation-statistics. Accessed: 2024-08-05.

[4] Priscilla Avegliano and Jaime Simão Sichman. 2019. Using surrogate models to calibrate agent-based model parameters under data scarcity. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 1781–1783.

[5] Peter Bank, Ibrahim Ekren, and Johannes Muhle-Karbe. 2021. Liquidity in competitive dealer markets. *Mathematical Finance* 31, 3 (Apr 2021), 827–856. https://doi.org/10.1111/mafi.12305

[6] Alexander Barzykin, Philippe Bergault, and Olivier Guéant. 2021. Algorithmic market making in foreign exchange cash markets with hedging and market impact. *arXiv preprint arXiv:2106.06974* (2021).

[7] Kayla Boggess, Sarit Kraus, and Lu Feng. 2022. Toward policy explanations for multi-agent reinforcement learning. *arXiv preprint arXiv:2204.12568* (2022).

[8] Peter Bossaerts, Paolo Ghirardato, Serena Guarnaschelli, and William R. Zame. 2010. Ambiguity in Asset Markets: Theory and Experiment. *The Review of Financial Studies* 23, 4 (01 2010), 1325–1359. https://doi.org/10.1093/rfs/hhp106 arXiv:https://academic.oup.com/rfs/article-pdf/23/4/1325/24430084/hhp106.pdf

[9] Karen Braun-Munzinger, Zijun Liu, and Arthur Turrell. 2016. Staff Working Paper No. 592 An agent-based model of dynamics in corporate bond trading. *null* (2016). www.bankofengland.co.uk/research/Pages/workingpapers/default.aspx

[10] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[11] Alain Chaboud, Ellen Correia-Golay, Caren Cox, Michael J. Fleming, Yesol Huh, Frank M. Keane, Kyle Lee, Krista Schwarz, Clara Vega, and Carolyn Windover. 2022. All-to-All Trading in the U.S. Treasury Market. https://doi.org/10.2139/ssrn.4256637

[12] Jon Cheshire. 2015. Market Making in Bond Markets. *RBA Bulletin March Quater 2015* March 2015, March (2015), 63–74. https://rba.gov.au/publications/bulletin/2015/mar/7.html

[13] Rama Cont, Mihai Cucuringu, Vacslav Glukhov, and Felix Prenzel. 2023. Analysis and modeling of client order flow in limit order markets. *Quantitative Finance* 23, 2 (2023), 187–205. https://doi.org/10.1080/14697688.2022.2150282

[14] Rama Cont and Marvin S Muller. 2021. A stochastic partial differential equation model for limit order book dynamics. *SIAM Journal on Financial Mathematics* 12, 2 (2021), 744–787.

[15] Antonia Creswell and Murray Shanahan. 2022. Faithful Reasoning Using Large Language Models. , 48 pages. arXiv:2208.14271 http://arxiv.org/abs/2208.14271

[16] Robert Czech and Gabor Pinter. 2022. Informed Trading and the Dynamics of Client-Dealer Connections in Corporate Bond Markets. *SSRN Electronic Journal* 895 (2022). https://doi.org/10.2139/ssrn.4214084

[17] Luis Goncalves de Faria. 2022. *An Agent-Based Model With Realistic Financial Time Series: A Method for Agent-Based Models Validation.* Papers 2206.09772. arXiv.org. https://ideas.repec.org/p/arx/papers/2206.09772.html

[18] Larry Epstein and Martin Schneider. 2010. AMBIGUITY and ASSET MARKETS. *NATIONAL BUREAU OF ECONOMIC RESEARCH* (2010), 1 – 51.

[19] J. Doyne Farmer and Spyros Skouras. 2012. Minimum resting times and transaction-to-order ratios: review of Amendment 2.3.f and Question 20. *European Commission Public Consultation: Review of the Markets in Financial Instruments Directive (MiFID)* (2012).

[20] Jean-David Fermanian, Olivier Guéant, and Jiang Pu. 2016. The Behavior of Dealers and Clients on the European Corporate Bond Market: The Case of Multi-Dealer-to-Client Platforms. *Market Microstructure and Liquidity* 02, 03n04 (2016). https://doi.org/10.1142/s2382626617500046

[21] Bank for International Settlements. 2014. *Market-making and proprietary trading: industry trends, drivers and policy implications.* Number 52 in CGFS Papers. Bank for International Settlements. https://ideas.repec.org/b/bis/biscgf/52.html

[22] Steven Mann Frank J. Fabozzi. 2005. *The Handbook of Fixed Income Securities, Seventh Edition.* McGraw-Hill.

[23] Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing Smaller Language Models towards Multi-Step Reasoning. *Proceedings of Machine Learning Research* 202 (2023), 10421–10430. arXiv:2301.12726

[24] Australian Government. 2022. Australian Federal Budget 2022. *Parliamentary Speech* (2022).

[25] Roy Hayes, Andrew Todd, Nachapon Chaidarun, Scott Tepsuporn, Peter Beling, and William Scherer. 2014. An agent-based financial simulation for use by researchers. *null* (2014). https://doi.org/10.2139/ssrn.2423370

[26] Badr Hirchoua, Brahim Ouhbi, and Bouchra Frikh. 2021. Deep reinforcement learning based trading agents: Risk curiosity driven learning for financial rules-based policy. *Expert Systems with Applications* 170 (2021), 114553. https://doi.org/10.1016/j.eswa.2020.114553

[27] John H. Holland, John H. Miller, and John H. Miller. 1991. Artificial Adaptive Agents in Economic Theory. *The American Economic Review* (1991). https://doi.org/null

[28] Cosmin L Ilut and Martin Schneider. 2022. *Modeling Uncertainty as Ambiguity: a Review.* Working Paper 29915. National Bureau of Economic Research. https://doi.org/10.3386/w29915

[29] Shima Imani and Liang Du. 2023. MathPrompter: Mathematical Reasoning using Large Language Models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 5, 1 (2023), 37–42. https://doi.org/10.18653/v1/2023.acl-industry.4 arXiv:2303.05398

[30] Andrei A. Kirilenko, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. 2018. Automation, Intermediation and the Flash Crash. *null* 16, 4 (2018), 17–28. https://doi.org/10.2139/ssrn.3119363

[31] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems* 35 (2022), 3843–3857.

[32] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct. arXiv:2308.09583 http://arxiv.org/abs/2308.09583

[33] Massimo Massa and Andrei Simonov. 2003. Reputation and interdealer trading: a microstructure analysis of the Treasury Bond market. *Journal of Financial Markets* 6, 2 (2003), 99–141. https://doi.org/10.1016/S1386-4181(02)00045-9

[34] Zakaria Mehrab, Logan Stundal, Samarth Swarup, Srinivasan Venaktramanan, Bryan Lewis, Henning Mortveit, Christopher Barrett, Abhishek Pandey, Chad Wells, Alison Galvani, Burton Singer, David Leblang, Rita Colwell, and Madhav Marathe. 2024. Network Agency: An Agent-based Model of Forced Migration from Ukraine. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1372–1380.

[35] Reserve Bank of Australia. 2021. Government-bond-markets-in-advanced-economies-during-the-pandemic. *Reserave Bank of Australia* (2021). www.rba.gov.au/publications/bulletin/2021/sep/government-bond-markets-in-advanced-economies-during-the-pandemic.html

[36] Bank of England. 2015. GEMM Guidebook: A guide to the roles of the DMO and Primary Dealers in the UK Government bond market. *Bank Of England* (December 2015). https://dmo.gov.uk/media/ownmub2i/guidebook20150401.pdf

[37] Bank of International Settlements. 2023. BIS Quarterly Review, December 2023.

[38] Joon Park, Joseph O'Brien, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC] https://arxiv.org/abs/2304.03442

[39] James Paulin, Anisoara Calinescu, and Michael Wooldridge. 2019. Understanding flash crash contagion and systemic risk: A micro–macro agent-based approach. *Journal of Economic Dynamics and Control* (2019). https://doi.org/10.1016/j.jedc.2018.12.008

[40] Gabor Pinter. 2023. An anatomy of the 2022 gilt market crisis. *SSRN Electronic Journal* Bank of England Working Paper 1019 (2023). https://doi.org/10.2139/ssrn.4488189

[41] Gabor Pinter, Chaojun Wang, and Junyuan Zou. 2024. Size Discount and Size Penalty: Trading Costs in Bond Markets. *Review of Financial Studies* 37, 7 (2024), 2156–2190. https://doi.org/10.1093/rfs/hhae007

[42] Jesper Rüdiger, Jesper Rüdiger, Adrien Vigier, and Adrien Vigier. 2019. Who Acquires Information in Dealer Markets. *null* (2019). https://doi.org/10.2139/ssrn.2875763

[43] U.S. Securities and Exchange Commission. 2013. Trade and Order Volume Ratios. https://www.sec.gov/about/trade-order-volume-ratios Accessed: 2024-09-10.

[44] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, and Adam Fisch et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615 [cs.CL] https://arxiv.org/abs/2206.04615

[45] U.S. Securities and Exchange Commission. 2024. Market Structure: Exchange Trade Volume. https://www.sec.gov/marketstructure/datavis Accessed: 2024-09-10.

[46] Nelson Vadori, Leo Ardon, Sumitra Ganesh, Thomas Spooner, Selim Amrouni, Jared Vann, Mengda Xu, Zeyu Zheng, Tucker Balch, and Manuela Veloso. 2022. Towards Multi-Agent Reinforcement Learning driven Over-The-Counter Market Simulations. arXiv:2210.07184 [cs.MA]

[47] Jose Varghese and Manoj Edward. 2018. Relationship Between Job Orientation and Performance of Sales People: A Financial Services Industry Perspective. *IIM Kozhikode Society & Management Review* 7, 1 (2018), 88–96. https://doi.org/10.1177/2277975217733858

[48] Alicia Vidler and Toby Walsh. 2024. Decoding OTC Government Bond Market Liquidity: An ABM Model for Market Dynamics. arXiv:2501.16331 [q-fin.TR] https://arxiv.org/abs/2501.16331

[49] Alicia Vidler and Toby Walsh. 2025. Evaluating Binary Decision Biases in Large Language Models: Implications for Fair Agent-Based Financial Simulations. arXiv:2501.16356 [cs.LG] https://arxiv.org/abs/2501.16356

[50] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. arXiv:2206.07682 [cs.CL] https://arxiv.org/abs/2206.07682

[51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35, NeurIPS (2022), 1–43. arXiv:2201.11903

[52] Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, and Jian guang Lou. 2024. Re-Reading Improves Reasoning in Large Language Models. arXiv:2309.06275 [cs.CL] https://arxiv.org/abs/2309.06275

[53] J. D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *Conference on Human Factors in Computing Systems - Proceedings* (2023). https://doi.org/10.1145/3544548.3581388

[54] Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and Jure Leskovec. 2022. Greaselm: Graph Reasoning Enhanced Language Models for Question Answering. *ICLR 2022 - 10th International Conference on Learning Representations* (2022), 1–16. arXiv:2201.08860

[55] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian Wynter, Yan Xia, and Wenshan Wu. 2024. LLM as a Mastermind: A Survey of Strategic Reasoning with Large Language Models. arXiv:2404.01230 [cs.CL] https://arxiv.org/abs/2404.01230