Agentic Intelligence in Financial Markets: A Multi-Agent System with Embedded LLMs for Bilateral Fixed Income Trading

Alicia Vidler

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Computer Science and Engineering

Faculty of Engineering

The University of New South Wales

31 March 2025

Thesis Title and Abstract | **Declarations** | Inclusion of Publications Statement | Corrected Thesis and Responses

ORIGINALITY STATEMENT

☑ I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

COPYRIGHT STATEMENT

AUTHENTICITY STATEMENT

☑ I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.

Originality Statement, Copyright and Authenticity Statement

Thesis Title and Abstract | Declarations | **Inclusion of Publications Statement** | Corrected Thesis and Responses

Candidates including publications in their thesis are instructed to select either Category 2 or Category 3. If the thesis is a combination of Category 2 and 3, they must select Category 3 and provide the details.

Here is a summary of the 3 available categories:

Category 1	The candidate's thesis contains no publications, either published or submitted for publication.
Category 2	Some of the work described in the candidate's thesis has been published, but they have not included the entire publication(s) in place of any chapter(s). Instead, the work has been rewritten/edited/expanded to incorporate it into the thesis, and this has been clearly acknowledged within the thesis.
Category 3	The candidate's thesis contains one or more publications that have been incorporated **in place of** a chapter/s (whether published or submitted for publication). All the following statements are true for each publication they are using in this way: • The research was completed **during the candidate's current HDR candidature**. • The candidate contributed **more than 50%** to the publication. • The candidate is listed as the **primary author**. • The publication is **not restricted** by any contracts or third-party agreements. • The chapter is included **word-for-word** as it appears in the publication (e.g. the **author-accepted version**).

☑ **Category 2:** The candidate has declared that **some of the work described in their thesis has been published, but they have not included the entire publication(s) in the place of any chapter(s). Instead, the work has been rewritten/edited/expanded to incorporate it into the thesis, and this has been clearly acknowledged within the thesis**.

A short statement on where this work appears in the thesis and how this work is acknowledged within chapter/s:

> My work consists of 4 peer reviewed papers. "Decoding OTC Government Bond Market Liquidity: An ABM Model for Market Dynamics" was accepted and is pending publication in IEEE and is the basis of Chapter 3. "TraderTalk: An LLM Behavioural ABM applied to Simulating Human Bilateral Trading Interactions" was peer reviewed and published in IEEE International Conference on Agents (ICA) Pages: 164-167 and is the basis of the Part A of Chapter 4. Part B of Chapter 4 contains the work " Evaluating Binary Decision Biases in Large Language Models: Implications for Fair Agent-Based Financial Simulations" peer reviewed and presented at AAAI 2025 Workshop on AI for Social Impact: Bridging Innovations in Finance, Social Media, and Crime Prevention. The final two technical chapters, 5 and 6 incorporate the peer reviewed work "Shifting Power: Leveraging LLMs to Simulate Human Aversion in ABMs of Bilateral Financial Exchanges, A bond market study" accepted at AAMAS 2025 as an extended abstract.

Candidate's Declaration

I declare that I have obtained written permission for the reproduction of all third-party material included in this thesis.

Inclusion of Publications Statement

Abstract

This thesis explores the integration of agent-based models (ABMs) with contemporary artificial intelligence inference methods – specifically, Large Language Models (LLMs) – to develop what has recently become known as 'Agentic AI', or otherwise, 'Agentic Intelligence'. The incorporation of agent-based model methods with LLMs holds great promise while raising many questions. Modelling complex dynamic financial trading networks, such as bilateral bond markets, is a challenging problem. In this thesis, I develop ABMs and incorporate LLMs, applying both to this particular problem domain. I examine the potential of such integration, including the challenges of using LLMs for modelling bilaterally traded fixed income markets, where the combination of limited data and complex human decision-making creates unique challenges.

My research reveals three key findings. Firstly, agent-based modelling demonstrates that multi-agent systems can effectively simulate strategic behaviours and autonomous decision-making in partially observable bilateral bond markets. My results reveal thresholds in the state of bilateral market stability, and demonstrate that market functioning depends more on agent diversity and cost structures than on the number of participants. These results validate the ability of agent architectures to capture complex market feedback mechanisms while providing quantitative insights for market design.

Secondly, LLMs, as components of Agentic Intelligence, demonstrate significant limitations in strategic decision-making, particularly in generating random responses and handling binary trading choices, exhibiting systematic biases and "aversion patterns" during complex decisions. Through systematic testing of LLM models in financial decision-making contexts, I demonstrate significant variations in output distributions between model versions and model sub-versions. Analysis reveals that methods of LLM sampling can critically impacts output and results – and I introduce nomenclature to describe both. I successfully integrate both in the TraderTalk model I present, demonstrating that LLMs can realistically simulate trading interactions without domain-specific training. Challenges remain in coordinating agent turn-taking and output interpretation; observed discrepancies between trading intentions and execution enhance the model's realism by mimicking human trader observed behaviours.

Thirdly, I produce a bespoke ABM with an integrated LLM, the TRIBE(LLM) Model. This model provides a demonstration of viability and empirical substantiation for Agentic

AI within complex financial market simulations. I showcase integrations of LLMs with agent-based models which enhance market simulations through more sophisticated client behaviour modelling. While LLM integration enriches agent decision-making processes, it also reveals critical system sensitivities – notably, even minor trade aversion encoded in LLM requests can trigger complete market cessation. The complexity of variation in LLM binary decision-making produces distributions that are not well replicated with probabilistic or deterministic sampling – opening the way for LLMs to provide more "human-like" unpredictability than current ABM methods. This integration models a shift of market power dynamics toward clients and introduces stochastic human-like variability that can lead to systemic market cessation, revealing emergent properties not captured solely by traditional ABM approaches alone.

These findings advance the theoretical understanding of multi-agent system design, detail and discuss challenges of LLM integration – along with suggested implementation remedies, and finally, provide practical insights for modelling complex financial markets. My research demonstrates that whilst the combination of these technologies offers promising capabilities, successful implementation requires careful consideration of their respective strengths and limitations, particularly in harnessing the potential of Agentic AI to enhance agent autonomy in modelling market dynamics.

Acknowledgement

This research have been supported and enabled by my supervisor, Professor Toby Walsh, and by a group of beloved close friends and family, and I thank them all dearly. When I began my doctoral studies, Professor Walsh advised our cohort that “life happens when you do a PhD”, and indeed during the course of this research I have experienced far more “life” than I ever thought possible. It has been my goal to study my field of research for more years than the average PhD candidate has walked this earth. In pursuit of my dreams, I am deeply grateful to my family – Shai and Ethan for their unbounded love, support, and belief in me. I thank Dr. Belinda Middleweek for her eternal friendship, her support, and her unwavering encouragement along the way. Without all of these, this PhD would not exist.

If life is a team sport, I am also lucky enough to have some fantastic teammates – many of whom volunteered to proofread various portions of my work over the years. In no particular order I would like to thank: Lena, Josephine, BMW, Ashley, Vanessa, Sarah, Christine, Warren, Margot, Warwick and Dave. I thank you all for your time, energy and honest comments.

I am profoundly thankful to Professor Toby Walsh for his academic guidance, broad subject matter expertise, infectious passion for AI, and steadfast support whilst a lot of “life happened” to me. Thank you, Toby, for believing in me and this research. I would like to thank my academic colleagues and collaborators for their engagement and insights. At UNSW, I am grateful to Professor Haris Aziz, Dr. Edward Lee, and Dr. Matthew Olckers. I am particularly thankful to Professor Gal Kaminka at Bar Ilan University for hosting me as a visiting scholar and for his guidance throughout my visit. I also thank Professor Sarit Kraus at Bar Ilan University for her insights. At the University of Oxford, I thank Professor Michael Wooldridge, Professor Anni Calinescu, Dr. Arnau Quera-Bofarull, and Dr. Nicholas Bishop for creating an intellectually stimulating environment. At the Technion, thanks to Professor Reshef Meir for insightful discussions. Additionally, I am thankful to Liquidity Cube for their industry collaboration on bond market research.

This work is funded in part by an ARC Laureate grant FL200100204

Finally, I dedicate this thesis to Max, and to finally becoming an Engineer.
May his memory be a blessing.

Publications and Presentations

UNSW is supportive of candidates publishing their research results during their candidature, as detailed in the UNSW. Thesis Examination Procedure. I have had the opportunity to publish several peer-reviewed pieces of work during my doctoral studies. The papers that form the foundations of specific areas included in this thesis are as follows.

List of publications: peer reviewed

- Alicia Vidler, Toby Walsh: **Decoding OTC Government Bond Market Liquidity: An ABM Model for Market Dynamics** (IEEE 2025 Symposium Series on Computational Intelligence, To appear - in proceedings). Available at: https://arxiv.org/abs/2501.16331.
- Alicia Vidler, Toby Walsh: **TraderTalk: An LLM Behavioural ABM applied to Simulating Human Bilateral Trading Interactions** (2024 IEEE International Conference on Agents (ICA) Pages: 164-167).
- Alicia Vidler, Toby Walsh: **Shifting Power: Leveraging LLMs to Simulate Human Aversion in ABMs of Bilateral Financial Exchanges, A bond market study**. (AAMAS 2025, To appear - In proceedings).
  Available at: https://arxiv.org/abs/2503.00320.
- Alicia Vidler, Toby Walsh: **Evaluating Binary Decision Biases in Large Language Models: Implications for Fair Agent-Based Financial Simulations** (AAAI 2025 Workshop: AI for Social Impact: Bridging Innovations in Finance, Social Media, and Crime Prevention).
  Available at: https://arxiv.org/abs/2501.16356.

List of publications (not included in this thesis)

During the course of my doctoral studies I have had the opportunity to work with academic collaborators looking at, and modelling, a variety of dynamic complex systems and environments. The following research is not directly contained in this thesis:

- Alicia Vidler, Toby Walsh: **Non cooperative Liquidity Games and their application to bond market trading**, (arXiv preprint 2024), available at: https://arxiv.org/abs/2405.02865.
- Alicia Vidler, Toby Walsh: **Playing games with Large language models: Randomness and strategy**, (arXiv preprint 2025), available at: https://arxiv.org/abs/2503.02582v1.
- Alicia Vidler: **Recommender Systems in Financial Trading: Using machine-based conviction analysis in an explainable AI investment framework**, (arXiv preprint 2023), available at: https://arxiv.org/abs/2404.11080.
- Cassandra Crowe, Belinda Middleweek, Laura Ryan, Alicai Vidler and Bronwen Whiting: **The role of gender in promotion rates in the Australian Finance Industry**, (arXiv preprint 2024), available at: https://ui.adsabs.harvard.edu/abs/2024arXiv240914384C/abstract.
- Matthew Olckers and Alicia Vidler and Toby Walsh: **What Type of Explanation Do Rejected Job Applicants Want? Implications for Explainable AI**, (IJCAI 2022 Workshop on Explainable AI), available at: https://arxiv.org/abs/2205.09649.

List of invited talks and tutorials

As interest has grown in AI applications, most notably in areas of financial markets, I have had the privilege of being invited to present tutorials and presentations at the following venues during my studies:

- EASSS 2023 Invited Tutorial "AI for Financial Trading. Autonomous agents and agent based systems applied to Bond Markets: Can we build a better market using AI?".
- AAMAS 2024 Invited Tutorial "The emerging area of agent based models applied to Bond markets".
- IJCAI 2024 Invited Tutorial "AI for Capital Markets – Agent based models applied to Bond Markets: Can we build a better market using ABM's?".
- Technion University Game Theory Seminar, 2024, Invited presentation: "Liquidity games a new type of repeated Bayesian games to model financial capital markets".
- 1st Technion Behavioural Data Science Symposium 2023. Invited talk entitled: "Intimate AI - Its anything but artificial. A definition. Virtual agents and ethical AI development".

Contents

List of Figures

List of Tables

Abbreviations

ABM	Agent-based model
AF	Argumentation Framework
BOE	Bank of England
COT	Chain of thought
GABM	Generative agent-based model
LLM	Large Language Model
MM	Market maker
OTC	Over–the–counter
OTR	Order to trade ratio
RBA	Reserve Bank of Australia

Chapter 1

Introduction

Agentic AI methods are at the forefront of multi-agent collaboration and design [1]. Yet, 'Agentic AI' and 'Agentic Intelligence' remains a relatively new term which is still largely being defined by the AI community [2]. While agent-based models and multi-agent systems have been researched for well over a decade [3], agentic models using artificial intelligence are more recent. They are generally understood to mean artificial intelligence models and systems that exercise or exhibit agency including computer systems that include work flows using agents [2]. One recent definition describes these models and systems as "autonomous agents that, given high-level instructions, plan, use tools, and carry out multiple, iterative steps of processing"[1]. Throughout this thesis I will use Agentic AI and Agentic Intelligence interchangeably.

In this thesis, I focus on simulating agent behaviour to study emergent phenomena in complex systems (namely bond markets), which I refer to as an agent-based model approach (ABM). Agent-based models employ computational methods to simulate interactions between various agents [3–5]. A closely related field, Multi-agent systems (MAS), can be thought of as the broad category of "systems composed of multiple interacting computing elements, known as agents" [3]. The integration of ABMs with Large Language Models

[1] `https://www.deeplearning.ai/the-batch/issue-253/`, accessed 25 Jan 2025.

(LLMs) holds promise for developing models with LLM enhanced strategic autonomy. In this thesis, I make use of MAS to approach integrating LLMs into ABMs. These ABM-LLM models can be termed GABMs or Generative Agent-based models [6]. Collectively, I will refer to these models as examples of Agentic Intelligence, or simply as Agentic AI.

As Agentic AI transitions from theory to practice, the integration of ABMs and LLMs represents one potential advancement in multi-agent models in the direction of Agentic AI, and forms the core of this thesis. While LLMs excel at generating human-like responses, they face fundamental challenges in tasks such as reasoning and making decisions. This is evident in their suboptimal performance in producing even simple random sequences [7–9]. Meanwhile, ABMs contribute established methodologies for modelling strategic decision-making and adaptive learning processes. The appeal of these methods is evident in commercial adoption – in November 2024, OpenAI announced the development of an autonomous agent-based version of their LLM, reportedly called "Operator" [10].

To explore the integration of LLMs into ABMs, this work focuses on financial markets, particularly bilaterally traded fixed income markets -— often referred to as bond markets, with government bonds forming a key subset. These markets present an ideal testing ground for agent-based AI methods. These markets are large and complex, yet operate with relatively few key central transacting agents (market makers) and often without a central clearing exchange, creating a market with limited observability and partial transparency. Government bonds are crucial to monetary functioning, public services, and retirement incomes in countries such as the United States, the UK, and Australia. However, trading remains bilateral, sometimes even conducted via phone calls. In these "over-the-counter" markets, an official market maker must typically participate in each transaction. These markets are systemically important – the UK government bond market, for example, has fewer than 20 market-making firms at the time of writing, and is ranked as the world's fifth largest bond market in 2023 [2].

Meanwhile, the COVID-19 pandemic accelerated the issuance of government bonds world-

[2]https://www.weforum.org/stories/2023/04/ranked-the-largest-bond-markets-in-the-world/, accessed 5 Jan 2025.

wide. In response to these expanding markets, central banks such as the Reserve Bank of Australia [11], Bank of Canada [12], Bank of England [13] and the Federal Reserve Bank of New York [14] have placed renewed focus on improving government bond market functioning, stability, and liquidity supply [15].

Recognising these significant market shifts, in July 2024, the Australian Council of Financial Regulators[3], the coordinating body for Australia's main financial regulatory agencies, launched an investigation citing "considerable changes in the size and structure of the Australian bond and repo markets in recent years" [15]. Against this backdrop of regulatory concern [11], this thesis examines how agent-based models might bring transparency to bilateral bond markets while uncovering emergent features from modelling a stylised market environment commonly found in Australia, Canada and the UK. To enhance modelling fidelity, LLMs are explored as a way to fill gaps in human behaviour representation, ultimately presenting a unified ABM-LLM framework that captures the dynamics of both market makers and client participants.

In addressing the engineering challenges of applying ABMs with embedded LLMs to bond market modelling, I explore several emergent properties of LLMs: their capabilities to replicate random and non-random distributions, their challenges in decision-making compared to human traders, and how their indecision can model societal aspects of bond market behaviour through the possible replication of human uncertainty.

1.0.1 Research question

This thesis investigates facets of multi-agent systems and collaboration between agent-based modelling and integrated large language models, creating agentic AI systems with direct application to bilateral fixed income markets. The research explores three core technical questions:

[3] https://www.cfr.gov.au/about.html

Research Question 1: How effectively can agent-based modelling architectures capture complex strategic behaviours in partially observable market environments? Through the development of a bespoke multi-agent model focussing on market maker and client interactions, I explore whether these approaches can meaningfully represent dynamic trading environments. This question addresses both the theoretical foundations of market modelling and the practical challenges of simulating environments with limited observable data. My research reveals several emergent properties, particularly around agent interactions and market stability.

Research Question 2: What limitations emerge when using LLMs to model strategic decision-making in multi-agent systems? To answer this broad question, I focus on examining LLM response distributions and decision boundaries within a very narrow strategic trading scenario. This tests the LLM's capability to replicate desired distributions and approximate human reasoning in such specific contexts. I report insights on LLM behaviour, particularly their suitability for modelling both deterministic and stochastic decision processes in agent-based frameworks and financial trading contexts. Stochasticity and unpredictability are important features of market trading strategies to avoid predatory trading "even though traders observe prices, they cannot infer other traders' actions" [16]. Randomisation can help prevent competitors from detecting patterns that they could otherwise exploit [17]. Given the results of LLM behaviour, I explore the challenges in autonomous decision making and negotiations between (LLM) agents in a limited virtual trading environment. In particular, I identify patterns of LLM "aversion" that emerge during complex trading decision tasks, aligning with observed patterns in US equity markets trading.

Research Question 3: I question whether AI built from the integration of LLMs can enhance traditional agent-based systems in meaningful ways. I apply agentic methods to bilateral trading markets and explore the use of LLMs within a broader ABM. Through the development of the TRIBE model (Trading Relationships in Bilateral Environments – an ABM comprised of an LLM), I present a framework that combines these distinct computational approaches. I detail the power held by client participation rates within stylised market simulations. I note a number of emergent properties and vulnerabilities of simulated markets. In further incorporating LLMs into the TRIBE model, I examine how LLMs can augment agent decision processes while maintaining the computational efficiency of traditional agent-based simulations. I note a number of computational challenges involved in this integration and the results of increased variance, experienced by simulations relying on LLM agents' decisions rather than probabilistic specifications.

These questions progressively examine the computational foundations of integrated agent-based models, moving from core agent-based architecture design through system integration approaches and finally, empirical analysis of combined model capabilities (GABMs). In addressing these questions, I aim to advance the theoretical understanding of basic agentic AI model design, provide implementable solutions for complex multi-agent simulations utilising LLMs, and ultimately, advance our ability to model complex financial markets.

1.0.2 Thesis structure and methodology

This thesis is structured around the three key areas that make up multi-agent collaboration methods. Each main chapter explores a dedicated aspect of the topic, including several bespoke models and testing methodologies appropriate for different modelling approaches. Following this chapter, Chapter 2 reviews the literature both recent and relevant to the three central aspects of agent-based approaches in this thesis. ABMs, LLMs and the domain-specific aspects of bond markets relevant to modelling. Chapter 3 is the first technical chapter and details the development and analysis of a bespoke ABM for bond markets, focussing on detailed modelling of one class of agents. Chapter 4 examines the integration of LLMs into agent-based models, with particular focus on their capabilities and limitations in replicating random processes and human decision-making abilities in narrow trading contexts. Chapter 4 also looks at technical aspects of integrating LLMs within ABMs such as API sampling and its implications for integration.

Chapter 5 presents the TRIBE model, augmenting work conducted in Chapter 3 (expanding to include limited client autonomy, or 'agency') and advancing to a unified ABM framework applied to bond market modelling. Chapter 6 extends the TRIBE model to include LLM integrations to represent individual client choices, step by step, producing the TRIBE(LLM) model. Chapter 7 concludes with key findings and implications for future research. Appendix A provides a comprehensive background on the mechanics and terminology of the bond market. Appendix B provides additional details specific to Chapter 3 and ABM landscape initialisation and client resource allocation. Appendices C and D contain additional LLM and prompt-specific testing data.

Chapter 2

Literature Review

I explore multi-agent methods, using a combination of agent-based engineering with contemporary artificial intelligence and inference methods. Large language models (LLMs) have seemingly become ubiquitous since first garnering user attention in 2022, and have gone on to establish themselves as the “most influential and widely adopted AI technology to date” [18]. Although questions remain about their fundamental capabilities to replicate human decisions [19], recent developments such as OpenAI’s “Operator” [10] suggest promising directions for integrating LLMs with agent-based approaches. This thesis examines the potential of such integration, applied to the domain of bilaterally traded fixed income markets, where the combination of limited data and complex human decision-making creates unique modelling challenges.

To explore this topic comprehensively, the research in this chapter covers three key areas: Firstly, agent-based modelling approaches and their strengths and limitations, Secondly, the specific challenges and opportunities in bond market trading, especially the complex dynamic environment, and thirdly, the characteristics of LLMs that enable agency. I conclude with a brief discussion of other methods considered whilst researching this work.

Please note: Additional literature that addresses specific technical aspects of LLMs and agent-based modelling systems appears only in individual chapters, where that material is unique to that chapter's implementation. For interested readers keen to better understand the bond market basics and features discussed throughout, please refer to Appendix A, where a short primer on bond markets can be found.

2.1 Agent-based modelling

Agent-based models (ABMs) employ computational methods to simulate interactions between various agents [3–5]. They have a relatively long history in computer science, with early concepts established in the 1940s (cellular automata by John von Neumann and Stanislaw Ulam). The realisation of ABM's potential became more widespread as computational power increased. Uses of ABMs in areas such as social science and economics were advanced as early as the 1970s with Thomas Schelling's famous segregation model, which demonstrated emergent social dynamics from simple agent-based rule sets and turn-taking. Additional work in the field include: [20], [21], [22], [23] and [24] to name but a few sources.

Major advances in ABMs applied to economics were pioneered by Axtell and Epstein in [25] and their Sugarscape model, where the authors apply ABMs to model the growth of artificial, complex societies. Their simulation captures a society where agents and their environment interact following a set of rules. Indeed, the original work is entitled "Growing Artificial Societies, Social Science from the Bottom Up". Published in 1996, it has been a widely adopted text on practical modelling and forms the basis of many later research efforts. Notable iterations of the [25] framework and its MESA extension (an early open-source Python library for agent-based modelling) are examined in [26–29]. The robustness and extensive research into [25], coupled with the extensive documentation and open-source nature of MESA, which extends Axtell and Epstein's original work, provides a reliable and accessible platform for my research. I build upon this work in Chapter 2 of this thesis.

Additionally, applications of heterogeneous and homogeneous multi-agent systems have been advanced both on small scales and large scales, including across the areas of game theory, negotiation and fairness [3,30–33] – to mention just a few works. Work in a similar vein includes [34–36].

The area of agent reasoning and communication, and contributions to the design of intelligent agents and logical frameworks for multi-agent cooperation, is an entire field in its own right (see works by [33], [37], [38] for examples). Extensions to automated negotiation and decision-making in environments of uncertainty, [39] blends game theory and AI to model complex, adaptive interactions (see [40], [41] to name just a few). Research on the impact of goals within ABMs includes [42] and a somewhat related area of regulation of (autonomous agents) is extensively covered in [43] as just one example. The ability to let agents 'play out' scenarios is crucial for reproducing complex dynamic systems [44]. Work has expanded to autonomous bidding agents in multi-agent systems [45–47]. Simulation using ABMs have proven powerful in its ability to replicate dynamics between interdependent and adaptive agents, and in facilitating the reproduction of complex dynamic systems [44,48–54].

My contribution to the field is by applying ABMs to a partially decentralised marketplace characterised by numerous sources of goods and relatively few agents. Furthermore, it combines this approach with the Adaptive Market Hypothesis (AMH), introduced in [55]. AMH focusses on the interaction between market participant behaviour and design characteristics. Although AMH is a substantial field in its own right, it posits that market efficiency is a dynamic process that evolves over time. This recognition of dynamic processes is explored in later work [56]. For the purposes of my research, the reasoning behind agent behaviour (rational or otherwise) is somewhat moot. What has been incorporated into my model development is the recognition that agents evolve and change within a market structure over time, irrespective of underlying motivations.

ABM design and parametrisation

Designing an ABM involves significant parametrisation of agents. The work of [3], [57], and [58] explores agent-based approaches, focussing on logic-based methods for agent behaviour, known as "belief, desire, intention" (BDI). These concepts can often translate to a process of analysis, planing, and action. However, specifying agent features presents a research challenge.

In work by [59] the trade-off between model refinement and parameter calibration is discussed, detailing three approaches: black-box calibration (with other work proposing concepts, such as goal-plan trees [60]), structural model refinement, and computer-based systematic calibration. However, [59] proposes a "white box" calibration approach, focussing on a top-down configuration to highlight the features that have the greatest impact on the model results. This approach reduces the number of sub-models to calibrate, relying on known collective agent behaviour outcomes. I have incorporated these principles, specifically "white-box" methods, into the models built and described in this thesis to provide a framework for abstracting heterogeneous model relationships, thereby facilitating more efficient model calibration. Other work developed on [59] includes [61], [62] and Prometheus – a method of ABM design [63] with goal orientation at its core, for example.

Calibration of ABMs

No consensus exists on the best method to calibrate an ABM [64]. The complexity of environments often precludes observing agents' utility functions or response functions [65]. Machine learning has been applied to estimate calibrations within ABMs across various systems [66], [67]. In finance, this has proven successful in foreign exchange trading [68], [69] however, these approaches require significant amounts of data, incompatible with the bilateral bond markets. The existing body of work often benefits from features of the application domain (e.g., high frequency foreign exchange markets in the case of [69]), which allows the reduction of agents into only a few classes, homogeneous in utility function and form, and mutually exclusive ("Liquidity Providers" or "Liquidity

Takers" [69]). Furthermore, previous work depends on either a highly observable market environment or a unique environment description due to regulatory requirements or other functions [70–72]. Other techniques for calibrating ABMs in finance include employing machine learning techniques (such as reinforcement learning, or robust optimisation with uncertain data [73]) and the creation of "surrogate" models. Although surrogates can improve computational efficiency by reducing parameter dimensionality [74], they are found to not help initial parameter calibration [64].

Validation of ABMs

During several decades of research into ABM techniques, several questions remain regarding the validation and testing of models based on ABM techniques. Work by [75] defines validation as "the process of determining if a model adequately represents the system under study for the model's intended purpose" and details that indeed there is "no universally accepted approach" but proposes nine methods: "data analytics, docking, empirical validation, sampling, visualization, bootstrapping, causal analysis, inverse generative social science, and role-playing". Many of these approaches are employed throughout this thesis. Early work by [76] detailed multi-agent system tests such as "unit, agent, integration, system, and acceptance". Different approaches to code testing and computer code validation methods are detailed in [62] and also [61]. In [77] practitioners and academics in the field of applied ABMs discuss the progress toward establishing 'credible' methodologies that meet peer acceptance standards. The authors draw on their various backgrounds and expertise to detail the challenges in the space – including the ability to explain results to non-technical users and how to validate the necessary model assumptions implicit within such ABM frameworks (such as goal adherence [78]). Some suggestions across the relevant literature include:

- **Statistical tests**: The validation of ABMs for finance is focused on statistical similarities between real-world data (such as profit, loss, stock market returns etc) and computer simulation output [79]. Key tests include the distribution of returns (non-normal), the lack of auto-correlation in returns, and volatility clustering of returns. Although the study demonstrates that two out of these three tests have been sufficiently passed, it highlights the diminishing profitability as the number of high-frequency trading agents increases. Other proposed methods include the element distance ratio, which measures the distance between two model structures to assess credibility [80].

- **Credibility**: [48–50] and earlier work by [53] detail the key challenge of credibility for an ABM. There is a consensus that the credibility of ABMs relies on a model's ability to be explainable through repeatable results. The application of ABMs typically focusses on complex, nonlinear systems, where explainable features rarely emerge from direct linear relationships in the modelling process. Instead, these features often arise from intricate interactions and feedback loops within the system where "emergent behaviour... cannot simply be inferred as a linear function of agent attributes" [81] (page 276). [48], p31, highlight that the "freedom of design also includes heterogeneity of agent populations or the environment", which significantly expands the calibration possibilities available to researchers. Furthermore, the authors detail the range of complex agent interactions that can be effectively modelled within this framework".

- **Explainable AI and Model Validation**: The validation of ABMs through explainability for building user trust in modelled systems is explored in [82]. Work continues on building design-focussed explanatory models, or rather models with explanatory aspects built into their design [83] and [84] to name a few. Other approaches involve surrogate (low-dimensional) models with fewer parameters [64]. Heterogeneous agent-based models (HAMs) have demonstrated effectiveness in replicating stylised facts from financial data series and explaining market anomalies such as bubbles, crashes, and sources of market "chaos" [85]. However, surrogate models typically rely on markets adhering to normality assumptions and require training through multiple simulated pricing scenarios. Although the COVID-19 pandemic demonstrated the utility of large-scale ABMs for policy simulations [80, 86], assuming that individual calibration errors would average out, this approach proves less effective for smaller agent populations, where homogeneity assumptions may oversimplify market dynamics [48, 87]. The distinctions between ABMs and equation-based simulations, explored by [88] and [89], become particularly relevant when validating small-agent ABMs in regulated market contexts, where specific behavioural patterns require credible and explainable simulation outcomes. In my research, I aim to advance the area of small agent ABM validation, focussing on specific, regulated market behaviours to provide credible and explainable simulation outcomes. I variously make use of a number of the methods suggested in [75] throughout.

2.1.1 Market mechanism research

Mechanism design, a subset of game theory, serves as "a tool to analyse decentralised systems and their properties" [90], examining how mechanisms implement social choice functions under rational participant behaviour. Unlike equity markets, which feature Walrasian mechanisms and established auction processes [91], bond markets lack centralised exchange mechanisms and are not composed solely of price-taking agents. Research by [92] explores the design of mechanisms in unique marketplaces, demonstrating how market outcomes can be influenced through design and collaboration, potentially approaching market manipulation when applied to financial markets. Additional research in this field [93–98] examines various aspects of mechanism design. Further research by [99] showed how technology has transformed traditional human-based market design, noting subtle changes in the flow of information to market participants through electronic mediums.

2.1.2 Agent-based computational finance

Initial research into agent-based computational finance integrated complexity theory and agent-based approaches to financial markets, [100] and earlier [101], with a focus on agents following strategies to forecast stock prices. By allowing flexibility in agent selection, researchers showed improved outcomes but highlighted the "relative infancy" of heterogeneous agent modelling [100]. They detail the empirically calibrated inconsistency between micro-agent behaviours and macro-level coordination imposed by some models, showcasing that agents and macro-level environments may be at odds. This thesis will investigate the potential for negative juxtaposition by testing various hypotheses of micro-changes and macro-level outcomes. A key focus of my work is indeed the micro level environment that is heavily regulated in financial markets, and the macro-level environment which is the most observable market aspect.

The market topology, which includes structure rules and participant conventions, is explored by [21] and [102], with applications to the corporate bond market detailed in [103]. [104] pioneered agent-based approaches using reinforcement learning models, with [105] and [106] contributing insights on coalition formation and information transmission. Similar work looked at equity market microstructure analyses [34–36, 107] and utilised ABMs to explore market topology and dynamics [21, 102–106, 108]. The Securities and Exchange Commission (SEC) research report, [109], examined applying ABMs to model dynamics of equity markets, focussing on transparency and minimum pricing increments. In [110], the authors demonstrate that an "agent based simulation model can accurately capture the key characteristics of the nearest months E-Mini S & P 500 futures market". This finding, along with work by [111], [112], supports the application of ABM to dynamic financial markets as digital replicas of market dynamics. [108] explored over-the-counter (OTC) corporate bond markets using machine learning methods to study dealer behaviour. I extend these approaches to model market maker (also often referred to as 'dealers') and client dynamics in the government bond market context.

The applicability of ABMs to financial markets is discussed in [79], and [112], which aims to demonstrate the flexibility and adaptability of ABMs in this context. The authors provide a method for equity market modelling, showing in fact the results of the carrying capacity of equity markets for particular trading strategies, such as those employed by market makers. Given that market makers play a defined role common to many markets, including bond markets, the authors' work offers valuable insights into how results from model implementations such as mine may be interpreted. They go on to introduce an agent type, the "intelligent" market makers, who choose their net position of securities to hold over the day". This approach emphasises the risk that such traders face from large price moves and their tendency to avoid trading with "informed" traders, those who trade in a persistent price direction. This phenomenon is referred to as "adverse price selection", where one type of agent systematically prefers to avoid trading with another type following specific trading patterns. I look to extend this work by exploring how these interactions and preferences can impact market dynamics further.

Energy market models and agent-based approaches: Energy markets have emerged as a prominent domain for ABMs – given their distinct characteristics that align well with agent approaches. The market structure, which features clearly defined roles (generators, distributors, consumers) with individual decision-making capabilities, naturally maps to the agent paradigm [113, 114]. ABMs have proven particularly effective in capturing the multi-layered interactions in energy markets, including modelling stability [115].

At the micro level, agents represent individual market participants with heterogeneous strategies, which through decentralised decision-making and complex supply-demand interactions produce emergent macro-level market behaviours [116]. The availability of granular data and clear regulatory frameworks has allowed detailed validation of these agent-based approaches [117, 118], but also captures complex interactions between supply and demand [119, 120] and [121]. This provides a demonstration of how ABMs can bridge the micro-macro gap in trading markets.

Multi-agent reinforcement learning: Approaches to model high frequency market trading (the limit order book) have been recently advanced [69], effectively categorising market participants, such as those in foreign exchange markets, as either liquidity providers or liquidity consumers. This approach models the utility functions of participants through iterative learning processes, demonstrating the potential to uncover the underlying behavioural patterns through repeated market simulations. Building on work by [122], the research uses a dual agent framework that captures the essential dynamics of market participants. Through continuous interaction with historical market data, the model demonstrates the ability to discover latent behaviours and utility functions that emerge from these two distinct types of agents. Similar work is reported in [72]. This thesis contributes to this area by expanding the number of heterogeneous agents specifically used in a market simulation.

Flash Crash analysis – an ABM specific example for market crisis analysis: The US equity markets experienced a "flash crash" on 6 May 2010. This extreme market event provides a detailed case study of a market under stress and agent behaviour within it. Market regulators reported [123] on the events, highlighting two crucial findings: In trading markets, market liquidity is not equivalent to market volume, and during periods of volatility, firms implement stricter risk controls, reducing both risk exposure and market-making functions. Market participants reported how rapidly changing perceptions of the market influenced their behaviour, with their risk aversion tendencies potentially reducing firm-specific risk, but collectively increasing market-wide risk. A particularly significant observation from [123] (page 5) notes:

> *Many over-the-counter ("OTC") market makers who would otherwise internally execute as principal a significant fraction of the buy and sell orders they receive from retail customers (i.e. "internalizers") began routing most, if not all, of these orders directly to the public exchanges where they competed with other orders for immediately available, but dwindling, liquidity.*

Relying on the electronic nature of the market and exchange-based participant categorisation by "style", [58] and [57] developed an ABM examining the price impact function in equity markets during this well-documented crisis event. They detail real-world verifiable actions, particularly how market makers' behavioural changes can impact the broader dynamic system. In particular, the risk-averse behaviour described in [123] aligns with regulatory requirements and the guidance of the banking regulators, suggesting that actions that appear detrimental to the wider market may in fact have been necessary actions on the part of trading firms, in order for them to fulfil their legal obligations to protect customer funds. **This dichotomy provides a key driver for my later work on client behaviour changes and market stability**.

2.2 Bond market research

2.2.1 Bond pricing

Research on bond market design remains scarce. The literature in the area of bonds appears to have focussed on pricing products [124–130], to quote just a few of the papers in the field. Furthermore, the literature reporting alternative pricing methods using operator functions is also advanced, including work by this author [131], [132] and [133]. Unlike stock markets driven by supply and demand, the prices of government bonds primarily respond to central bank interest rate policies, inflation expectations [127], and periodic interest payments. Although bond prices inversely relate to market-expected yields and can price in potential rate changes, they largely reflect current rate policies, which is information that is publicly available. For more detailed discussion on how bond pricing works, please see Appendix A.

When holding a bond, which provides interest returns, access to bond market "flow" or "liquidity" becomes crucial, as returns come from holding the asset. Bond market design research has tended to focus on a related market, that of the corporate bond market [103] or derivative markets such as bond futures [134]. Market-specific work was carried out by [135] in a highly specialised financial product market for Contingent Capital Callable bonds. Work has also been carried out on smaller markets such as the supply of wood [136]. Although financial market products may appear similar, there are distinctly different.

2.2.2 Bilateral markets

Bilateral markets are characterised by trading that occurs between two parties. Only the two parties to the transaction know that a trade has taken place until it is required for one party to disclose the trade details to a third party (such as a regulator, central bank or central clearing house [15]), typically shortly after the close of a trading day.

In contrast to bilateral trading, most people will be familiar with multi-lateral markets such as the New York Stock Exchange or the London Stock Exchange. In multi-lateral markets, transactions typically occur through an official exchange firm (e.g. the London Stock Exchange is a business owned by LSEG[1]) and these exchanges are characterised by a central clearing house and, or, another centralised mechanism for reconciling trades between multiple parties. A common characteristic of exchange-traded markets include "standardized contracts that have been defined by the exchange" [137] and are typically regulated by governments and regulatory bodies [15, 138].

Unlike multi-lateral markets, bilateral markets such as the government bond markets studied here, have one party to every transaction that must be a regulated "market maker" with specific obligations and characteristics as defined in each country's regulatory framework, including reporting the transaction to the issuing government [139]. However, the bilateral nature of bond market trades prevents direct observation of agents' utility functions and policies [65]. This hinders the application of machine learning techniques that have worked well in calibrating ABMs for observable markets like foreign exchange [66, 68, 69]. Such approaches typically require features absent in bilateral bond markets: homogeneous agent classes, observable environments, or comprehensive regulatory data [70–72].

I have focussed my work on analysing bilateral government bond trading in markets whose historical design is linked to that of the former British Empire. This primarily includes Australia, the United Kingdom, Canada, and New Zealand. A key feature among these markets is that they do not have a central clearing house exchange [11–13] and are therefore also not open to the public [11, 15]. There are of course other design models, such as the European Central Bank's process of issuing bonds on behalf of member states, and the US market, where large volumes are cleared through official exchanges [140]. Given the historical ties of the former British Empire markets and their shared market design, there

[1] https://www.lseg.com/en

is no reason to suppose that any one market (e.g. Australia) differs significantly from another (such as the United Kingdom or Canada). However, the US market represents a more stock exchange-like system [141], with the market for bonds issued by the European Central Bank falling somewhere between with both methods in operation [140]. In 2023, approximately 53% of the trading in Euro-government issued bonds occurred on a centralised platform [142]. For more information on the role of central banks, see Appendix A.

Market micro-structure and trading venues:

The lack of publicly available data for empirical validation is a key challenge in simulating bilateral markets like bond markets [143]. Although macro trading data indicates that domestic market makers traded between themselves on average 27% of all secondary market bond turnover in Australia since 2016[2], granular details on individual trades and participant strategies remain obscured. Bilateral transactions are mediated through a handful of electronic messaging platforms [144]. Recent efforts to launch a joint venture that centralises the collection of data on bond market transactions [145] underscore the possibilities for future research and the regulatory demands for transaction visibility [146].

Whilst bond markets have no unified trading venue, adding to data collection challenges, they do trade across fragmented and overlapping technical platforms. These facilitate the bilateral nature of bond trading and include global firms such as TradeWeb, Bloomberg, LiquidityCube, ICAP, MarketAxess, LiquidNet, and Yieldbroker. These platforms act as facilitators, essential for agent trading without a clearing bank or similar entity. In May 2023, significant industry consolidation was announced. TradeWeb acquired Yieldbroker [3], while ICAP purchased LiquidNet. Furthermore, Bloomberg, TradeWeb, and MarketAxess launched a joint venture with the goal of centralising transaction data collection [4].

[2] https://www.aofm.gov.au/data-hub

[3] https://www.afr.com/chanticleer/bond-market-players-call-time-on-6-trillion-aussie-exchange-20230525-p5db5n

[4] https://www.tradeweb.com/bloomberg-marketaxess-and-tradeweb-sign-joint-venture-agreement

Despite significant amounts of bonds being traded OTC across fragmented platforms, global markets are increasingly facing pressure from governments to move to central clearing and electronic trading venues [14, 109, 147]. This shift brings enhanced transparency and aligns with the work of [148]. As regulation tightens, stronger rules around transparency are required [148], and regulators express growing concerns about bond market structures [11–14, 149]. Regulatory changes impact future market designs both positively and negatively [150].

In the future, it may indeed be possible for researchers to model fine-grained data and transaction level information, something that remains lacking at the time of writing.

Bond market participants, liquidity and crisis events:

Bond markets comprise various participants, including retail banks as market makers (interchangeably referred to as 'dealers'), superannuation funds as debt purchasers, and hedge funds as clients [143]. Despite initiatives like YieldBroker, Australia, for example, lacks a dominant trading venue for government bonds, resulting in an opaque, decentralised, and fragmented market structure.

Although futures contracts that reference government bonds trade on exchanges such as the Australian Stock Exchange, offering potential model calibration data [85], these derivatives possess distinct characteristics, including cash flow and leverage features, that make them unsuitable for understanding underlying bond liquidity. Futures contracts, which derive value from changes in asset price without conferring ownership, differ fundamentally from bonds held by 'real money' investors seeking direct ownership[5]. Recent crisis events highlight market vulnerabilities. The 2022 UK gilt market crisis [143] demonstrated market fragility and risks from concentrated market makers, mirroring disruptions in U.S. Treasury markets [146, 151]. These events revealed particular vulnerabilities in OTC markets where liquidity providers struggled during extreme volatility. A growing body of literature is examining the efficiency of markets and the potential for "chaos" to

[5]https://www.asx.com.au/markets/trade-our-derivatives-market/overview/interest-rate-derivatives/bond-derivatives

be a distinct market feature [122] [152], adding to the complexity of market microstructure analysis.

The **liquidity** of OTC government bond markets plays a critical role in the functioning of modern financial systems, influencing both market stability and the implementation of economic policies. Liquidity, in this context, refers to the ease with which bonds can be bought or sold within a reasonable time frame, or rather *"liquidity is the ability to trade when you want to trade"* [153]. Despite its importance, market liquidity in OTC bond markets remains difficult to quantify and predict, particularly during periods of stress when liquidity can evaporate suddenly, triggering broader financial instability [143].

To quote former US Federal Reserve Governor, Kevin Warsh, "Liquidity is confidence"[6].

The importance attached to the well functioning of markets for bonds can be seen by the Reserve Bank of Australia when it states:

> *"The cost and ease with which governments obtain funding in primary markets can be affected by conditions in secondary markets for government debt. Dealers act as intermediaries within the secondary markets and between the primary and secondary markets, providing liquidity to maintain smooth trading conditions. Dysfunction in the secondary markets can make it difficult or costly for market participants to buy and sell securities, which can in turn limit the ability or willingness of participants in primary markets, particularly dealers, to buy newly issued securities. In addition, bond yields in secondary markets influence the cost of issuing government debt in primary markets".* [154].

To avoid confusion, it is helpful to note that "primary markets" refer to the initial offering of government debt by a government to clients and occurs only once (ever) for each bond issuance. Primary markets can be thought of as a version of an IPO for a bond – they only occur once for each bond ever. All activities thereafter, such as clients buying or

[6]https://www.federalreserve.gov/newsevents/speech/warsh20070305a.html

selling these bonds or perhaps governments repaying the bonds earlier by buying back their own bonds, constitutes activity in the "secondary" market for bonds. Throughout this document, I will deal exclusively with the secondary market for bonds and will simply refer to "markets" with the understanding that this is specifically the "secondary" bond market.

Market stability:
The World Bank defines financial market stability as *"the absence of system-wide episodes in which the financial system fails to function"* [155], with the International Monetary Fund going so far as to define the stability of certain financial markets, such as that of the UK system as a "global public good" [156].

Maintaining an efficient and liquid bond market is crucial for Australia's economic stability [157]. As discussed above, bond prices largely reflect current rate policies, which is information that is publicly available, and therefore bond prices are largely known [158] (for more detailed discussion on how bond pricing works, see Appendix A). However, the main concern of regulators and issuing governments is the liquidity present within the market for their government bonds [11, 14, 15, 141, 154]. In this thesis, I explore modelling the liquidity of bonds rather than focussing on price dynamics, creating an ABM which provides insights into market stability and the role of market makers in these interactions. This approach builds on the work of [159] and [160], who explore bilateral negotiations and liquidity flows in similarly nontransparent markets, with industry bodies contributing in [161] and [162]. Market stability serves as the key requirement for agents that transfer liquidity (bonds and cash) through the artificial society.

While exploring market stability, significant contributions have been made through social choice literature in related areas, particularly in argumentation frameworks (AF's) [163–166] which form possible interpretations of the style of bilateral negotiations taking place in real life markets. The work on AFs with incomplete information [167], [168] presents possibilities for analysing bond market interactions. However, while these approaches merit further investigation by others in the field, my research adopts an ABM, and uses simulation to capture the dynamic complexities of market behaviour.

There are many other aspects of the bond market beyond the scope of this work; bond pricing, prudential regulation, inflation, and central bank interest rate decisions, to name but a few. Whilst all of these are interesting topics, they are outside the scope of relevant market information for the purpose at hand. The interested reader is directed to Appendix A, where I cover some of these areas pertinent to the specific topic at hand more fully. Other extensive resources include books such as [137] and [158].

2.3 Large Language Models (LLMs)

2.3.1 Short history

LLMs represent a significant advancement in natural language processing, built on transformer architectures that utilise self-attention mechanisms to process sequential data (see [169] for more details). Pre-trained LLMs, such as GPT-3 [170] and GPT-4 [171], 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. These models, trained in vast corpora of text, employ multiple transformer layers to capture complex linguistic patterns, contextual relationships, and semantic meanings (see [172] for a comprehensive introduction). However, they may be acting as a "stochastic parrot" [19], defined as "systems that haphazardly stitch together sequences of linguistic forms observed in their vast training data according to probabilistic information about how these forms combine, but without any reference to meaning" (p.617).

Embeddings (see [173]) are a crucial component of GPT models and are a numerical representation of text. However, while fundamental to their operation around text, a detailed discussion of embeddings falls beyond the scope of this thesis, and they do not directly impact the model designs and implementations presented herein.

2.3.2 Agency in LLMs

Whilst pre-trained LLMs, such as GPT-3 [170] and GPT-4 [171], have demonstrated proficiency in tasks such as dialogue generation and natural language interactions, a fundamentally different question is asked of an LLM within an ABM. Unlike text and analysis needs, ABMs rely on *agency*, where an agent is "capable of *independent* action on behalf of its user or owner" [3] (p5). That is to say, agents are imbued and defined by the requirement to make a decision in some frame or context [3]. Many such implementations ultimately require a binary choice from the agent: to move now, or not, "yes" or "no". Although LLMs offer promising capabilities as decision-making agents, early attempts to introduce an LLM as an agent [174] found that granular agent decisions were technologically unfeasible to connect on a per agent basis within large ABMs.

However, Concordia [6] was able to integrate LLMs in ABMs to create Generative Agent-Based Models (GABMs). In this framework, the authors reported that agents show the ability to "apply common sense" and "act reasonably" within simulated environments [175]. Recent studies [176, 177] reveal that current LLMs perform poorly in negotiation tasks and strategic reasoning within agent-based systems. Recent work by [178] explores LLM agents in the modelling of transportation systems, recommending evaluation tests using behavioural soundness and benchmarking system-level performance against theoretical information.

Unlike previous work [179] which designs generic agents using LLMs, I focus on design features essential for heterogeneous agents in a multi-agent system. I build on recent scholarship that proposes analysing the agency within LLMs to evaluate how ABMs utilising large language models may affect decision-making [180].

2.3.3 Limitations of LLMs

LLMs excel in generating coherent text in various domains [181], yet face notable limitations in areas such as reasoning, numerical understanding, and optimal prompting methods [177, 182], [183] and [184]. Numerical reasoning [185], [186], [187] in particular is a source of concern for financial applications. To mitigate some of these limitations, Retrieval-Augmented Generation (RAG) frameworks have emerged, combining the generative capabilities of LLMs with external knowledge retrieval systems (see [188] for further discussion).

Additionally, LLM reasoning in negotiation tasks remains a challenge [176, 177] – showing LLMs performing poorly on negotiation tasks. However, work by [189] shows that LLMs can learn to interpret corporate financial reports with great skill. Working with LLMs via **prompt design** remains an evolving research area. Studies such as [190] have observed that non-expert users often adopt opportunistic, rather than systematic, approaches to prompting. One area that garners a lot of curiosity is the phenomenon of "hallucinations", when a model produces false or misleading information, presented as facts and often times with great conviction [191]. Several approaches and interpretations exist [192, 193], and suggestions about faithful reasoning are posited [194], [195]. Work looking at the improvement from multi-step reasoning is covered in [196]. I address these issues with selective and opportunistic prompting in the same vein as [190]. I leave it to other researchers to develop methods that can more broadly describe the prompting space without overburdening analysis engines.

A note on the proliferation of foundational LLM models: at the time of writing, a veritable explosion in foundational models is underway. Recent work by Deepseek [197] calls into question the "large" in LLMs. Numerous smaller models such as Bert [198], Llamma [199] and Mistral [200], to name a few, exist and can be used in smaller or more specific tasks, including being fine tuned and locally installed on accessible (laptop) hardware. In this thesis, research has focussed on the most powerful, so-called "state-of-the-art" models for two reasons; firstly the largest body of research has focussed on this group of models, and secondly, they are generally considered to be the benchmark for highest quality reasoning and output [201]. This advantage will likely prove short-lived as the rapidly evolving field soon renders such distinctions obsolete, though I leave this to future researchers to explore.

2.3.4 Prompt engineering

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 [202]. Another advanced method for improving LLM reasoning is **Chain-of-Thought (CoT) prompting**. The CoT prompt elicits a step-by-step reasoning process in the model output by providing "worked examples" that guide the LLM through complex tasks [203]. Inspired by human learning theories, this approach has been shown to enhance LLMs abilities to handle multi-step reasoning problems [183, 202, 203].

These limitations underscore the need for robust strategies in deploying LLMs within complex simulations. A significant challenge in this field is the dynamic nature of foundation models, which undergo frequent updates that can substantially alter their behaviour and capabilities [7]. This volatility creates a model risk where research based on specific versions may quickly become deprecated. This model risk is managed by focussing on the SOTA GPT40-mini and developing a framework-based approach rather than relying on fine-tuning of third-party models. This approach aligns with emerging research integrating LLMs with agent-based frameworks [6, 175, 204], a direction that has shown considerable promise [1]. In financial applications specifically, significant advances have been made, from Bloomberg's pioneering work in LLM fine-tuning [205] to recent breakthroughs in financial statement analysis [206].

LLMs encompass numerous fascinating research areas, including few-shot versus zero-shot learning, world simulation vs decision resolution, the impact of temperature settings on agent behaviour, and the trade-offs between model size and performance (see [204] for a comprehensive survey). These topics, particularly those relevant to the implementation of ABMs with LLMs, will be explored in subsequent chapters.

[7] 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.

2.4 Other related research areas

Before narrowing my focus to agent-based methods enhanced by LLMs, I also considered several complementary strands of inquiry. These included classical game theory and equilibrium analysis, as well as graph neural network approaches to model trading relationships, all of which provided valuable insights into strategic decision making and market dynamics. I briefly detail these areas below:

Game Theory:
Games of strategy cover many forms, from early work pioneered by [207] such as single shot, pure and mixed strategies, to extensions including repeated games [208].

A central idea in game theory is the notion of equilibrium [209], an outcome where neither player can unilaterally deviate and improve their payoff. Probabilistic notions explored in Bayesian games [210] extend these ideas to include the role of probability in equilibrium. The application to agent-based systems [3] brings forward concepts of Pareto optimality and agent coalition formation, particularly relevant to computer-based game playing [32]. Other aspects of game theory include work that touches on multi-agent systems (including embodied AI, such as robots) including [211], dynamic game theory [212] and research into rational robotic swarms [213].

Strategic games often lack a consistent "pure strategy" and instead rely on mixed strategies, where players randomly choose between different pure strategies according to specific probabilities. This raises important questions when integrated with LLMs. In work beyond this thesis, I continue to explore the ability of LLMs to learn optimal mixed strategies, given their known limitations in replicating random variables [214]. Although LLMs have proven useful in social science experiments and societal simulation [201], their potential application to game theory, particularly in repeated games and strategic decision making, presents unique challenges.

Taken together, the thesis focusses on ABM methods with and without incorporated LLMs, rather than direct game-theoretic analysis. The combination of systems with many heterogeneous agents and limited visibility on agent rationality suggested that an agent-based approach with direct modelling would be more tractable. In research work not contained in this thesis [215], I have continued to develop game-theoretic approaches to simplified bond market transactions, and use LLMs in simple game-like setups [183].

Graph methods:
Given that trading entities can be modelled as nodes in a graph, I briefly explored graph neural networks, building on the work of [216]. Currently, markets of interest have limited available data. In the future, as additional market data could be incorporated, techniques that use graph representations and matching market makers' relationships, such as those suggested by [217], offer promising avenues for further investigation.

2.5 Summary

This chapter reviews literature spanning agent-based modelling, bond market structure, and large language models, highlighting key opportunities and limitations in each domain. Although ABMs excel in simulating market dynamics, and LLMs show promise in natural language tasks, both face distinct challenges in modelling bilateral bond markets. The bilateral nature of bond markets, combined with their inherent opacity and relationship-driven dynamics, creates unique challenges that traditional (data-hungry) modelling approaches struggle to address in isolation. This gap presents an opportunity to explore how, firstly, ABMs with limited calibration might model a bilateral market, and, furthermore, how LLMs might augment agent-based models by incorporating human-like factors (such as stochastic parrots [19]) which can increase the fidelity of simulating trading environments.

Importantly, the approach I present does not rely on fine-tuned or domain-specific LLMs that risk over-fitting and becoming obsolete by the time of implementation. In future research, it will be interesting to consider how fine-tuned models (that is, models with domain-specific training) may improve or decrease results. Instead, here I choose to develop a framework that leverages the complementary strengths of ABMs and LLMs while accounting for their respective limitations – particularly in areas such as numerical reasoning and strategic decision-making.

Chapter 3

Bespoke ABM of Market Maker agents and Clients: Bilateral Markets

This chapter encompasses work from the peer-reviewed publication:
Vidler, A., Walsh, T. (2025). "Decoding OTC Government Bond Market Liquidity: An ABM Model for Market Dynamics". In *Proceedings of the IEEE Conference on Computational Intelligence for Financial Engineering and Economics (CiFEr-25 – To Appear).*

Decoding OTC government bond Market liquidity and dynamics

In this chapter, I devise a multi-agent simulation employing small-scale meta-heuristics to model the opaque, bilaterally traded, government bond market. My approach captures the unique over-the-counter (OTC) bank-to-bank trading dynamics between Market Makers (MM), facilitated among a limited number of agents, without the use of price mechanisms.

The model focusses on the dynamics of liquidity and stability in the trading of government bonds, particularly in concentrated markets like those found in Australia and the UK. Through designing and implementing a bespoke ABM, I report results for over 400 simulated trading epochs. Each trading epoch can at most last 1500 model time steps, within each time step each agent can, if conditions permit, move. I test key hypotheses around improving market stability, focussing on the effects of agent diversity, business costs, and client base size. I report emergent properties; greater agent diversity enhances market liquidity and that reducing the costs of market-making can improve overall market stability. The model offers insights into computational finance by simulating trading without price transparency, highlighting how micro-structural elements can affect macro-level market outcomes. This research contributes to the evolving field of computational finance by employing computational intelligence techniques to produce a market simulator to better understand the fundamental mechanics of government bond markets, providing actionable insights for both academics and practitioners.

3.1 Introduction

One can neatly conceive of a bond market as a set of agent-based interactions, representing the dynamics of bond market trading using autonomous agents. I introduce this concept to include participants such as MMs (who are agents in this model), the trading environment (i.e. the trading landscape or market place which encompasses clients), and interactions to simulate and analyse the factors influencing trading of bonds between agents.

To address this, I build an ABM designed to simulate the micro-level interactions of key participants (clients and MMs) within the OTC government bond market. ABMs are particularly suited to this challenge, as they allow the examination of complex, decentralised systems where individual behaviours and local interactions give rise to emergent phenomena at the macro scale [3].

3.1.1 Market participants

Global bond markets, including those of former British Empire nations, such as Australia, involve diverse participants acting as agents: banks as trading facilitators and MMs, institutional investors (as debt purchasers for longer term investments), and trading firms like hedge funds. [218]. Unlike centralised equity (stock market) exchanges, bond trading occurs over-the-counter across fragmented electronic platforms. The Australian bond market exemplifies this decentralised landscape, lacking a dominant trading venue and operating through opaque bilateral transactions facilitated by the four major retail banks designated as MMs [144], while being systemically important to both Australia and the global trading landscape [154].

Within this framework, several distinct types of participants interact under established frameworks and market conventions and can be categorised into three key types:

1. **Issuer and market regulator**
   Central banks (e.g., Bank of England, Reserve Bank of Australia) issue government bonds and regulate the issuance mechanism. They auction bonds to regulated market making banks (UK: 18, Australia: 19, Canada: 15 [1]). This primary market issuance is distinct from secondary market trading [2].

2. **Market maker**
   Regulated financial firms with legal obligations to make markets in government bonds [218]. These firms (usually regulated banks) must be available during market hours to buy or sell bonds. Their systemic importance stems from the requirement that clients trade through them, creating a bilateral, decentralised trading environment. In the Australian market for bonds, similar to many other markets, MMs operate under principles of neutrality and independence. They are expected to function without exhibiting preferences for specific clients or forming coalitions with other

[1] Online information accessed 10 July 2024.

[2] https://www.bankofcanada.ca/markets/government-securities-auctions

market participants, as required by regulations in markets such as Australia's [3]. MMs must prioritise the "best functioning" of the marketplace [218]. As such, I do not model MM strategic coalition formation, which can be contrary to policy [219]. These assumptions simplify the model by avoiding complex information exchange dynamics and fit within the regulated bond trading environment.

3. **Clients**

   End purchasers of government bonds, ranging from small non-profits to institutional investors, pension funds, hedge funds, and corporate finance divisions. Clients cannot trade directly with each other in most bilateral markets and are commonly classified as long-term investors (e.g., pension funds) or short-term trading firms [143]. In this work broadly, I do not distinguish between the two categories. Although the model in Chapter 5 and 6 examines the impact of client-specific behaviours on market dynamics, it deliberately avoids ascribing particular behaviours to traditional investor categories due to a lack of data and transparency. Instead, this thesis analyses how these agent behaviours affect aggregate market outcomes regardless of which investor type exhibits them – a methodological choice that acknowledges the regulatory environment in which all participants, regardless of classification, are subject to equivalent oversight and compliance requirements.

In practice, MMs and clients generally operate under expectations of fair dealing that discourage counterparty favouritism in transaction decisions (see [15, 138, 139, 218, 219]). While formal regulations offer principles-based guidance, the day-to-day implementation often relies on market convention where participants interact according to established industry practices.

Note: Within the ABM presented in this chapter, MMs will be modelled as agents but clients are modelled as locations within a landscape, where clients are agnostic as to which MM they engage with. As discussed in the literature review, primary markets

[3] Corporations Act, 2001 – Australia, Section "Fair and Efficient Order Handling Policy" Section 912A(1)(a).

represent the initial issuance of bonds (analogous to an IPO), occurring only once per bond. All subsequent trading activity constitutes the secondary market. This thesis focuses exclusively on secondary bond markets, with 'markets' referring specifically to secondary bond trading unless otherwise stated.

3.1.2 Research challenges

A particular challenge for research on bilateral bond markets is that data in this field is sparse and bond trade level data are not freely available [143]. To address this, I model the dynamics between MMs, the data for which are partially available to the public, isolating their trading behaviours and liquidity provision through agent-to-agent trading and client interaction. In addition to liquidity, the related concept of market stability is important [155, 156]. Through an ABM approach, my aim is to provide novel insights into the mechanisms that drive liquidity fluctuations and market resilience. Researchers are also studying the inherent uncertainty in financial markets and the potential for "chaos" to be a distinct market feature [122] [152]. I incorporate this perspective into the work presented here.

This chapter is structured as follows. I describe preliminary details on markets and related concepts. The hypotheses that I seek to test using an ABM are described below. Next, I formally introduce the ABM in section 3.3, including the definitions of key components and assumptions. The results of more than 400 simulated trading epochs are presented to answer the four hypotheses presented. Finally, I conclude with a discussion of the limitations of my approach, later work in this thesis, and future academic research into emergent properties.

3.1.3 Preliminaries

Bond Markets: Bond market structures broadly fall into two groups: bilateral trading and multi-lateral (exchange-based trading) where participants can trade with multiple counterparties simultaneously. Government bond markets trade OTC and are characterised by their bilateral trading structures, which pose unique challenges to understanding and ensuring market stability and liquidity.

I focus on bilateral trading markets for government bonds, of which there are several global examples (e.g. Australia, UK, Canada). These markets share a common ancestry (United Kingdom), common structural characteristics, and similar numbers of market-making agents, typically less than 20 agents. Exploring complex adaptive financial trading environments through multi-agent based simulation methods presents a novel approach within the realm of applied computer methods for financial markets research.

A short background on artificial societies: Using the definition from [220], an artificial society is an "agent-based, computer-implemented simulation model of a society or group of people, usually restricted to their interaction in a particular situation". Common across economics and social sciences, these virtual societies enable in-silico representation of agents in controlled environments, offering model designers significant flexibility in implementation.

Recent work by [221] introduces the concept of *network agency* or the fact that features beyond an agent's direct experience can influence their behaviours. While many of these models rely on homogeneous agent classes and observable environments, they are less relevant in OTC bond markets where data availability is limited [69]. In this thesis I model market structure and agents inductively and use stylised facts for related markets to specify agent calibrations [143]. The utility functions and policies used within the simulation approach stem from finance literature and regulatory requirements [11,154,218].

3.1.3.1 The Australian case study:

The Australian government bond market, like other global bond markets, involves a diverse range of participants [154]. At roughly half the size of UK or Canadian government bond markets, it shares essential structural features with these markets due to their common

British market design heritage, including similar debt ratings and governmental structure. In 2023, the Australian government debt (issued bonds) exceeded $890 billion.

The current structure of the Australian bond market is under critical examination by local regulatory authorities, and an active review is underway by the Council of Financial Regulators [11, 15]. Key concerns rest around the fact that "currently, there is no central counterparty (CCP) that clears transactions in the Australian bond and repo markets". This means that clearing occurs bilaterally and reflects the complex web of transactions that occur between participants [15]. The lack of a central counterparty (i.e. the lack of a clearing house) produces a decentralised trading framework. The impact of this 'complex web' is detailed below and a key impetus for this research thesis. The roles, responsibilities, and behaviours of the participants are covered by a number of regulatory bodies [138] and Australian Office of Financial Management (AOFM).

Maintaining an efficient and liquid bond market is crucial for Australia's economic stability [157]. Despite its importance, the AOFM only publishes high-level macro-trade data [154]. Given the shared characteristics with the UK and Canadian markets, the insights from this analysis can be generalised to these similar jurisdictions [222].

By owning a bond, the holder receives returns (interest payments) [158], making the bond market unusual in that prices are less critical than access to "flow" or "liquidity". 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 MM, such as the four major retail banks designated as MMs in Australia [144, 154]. Their transactions are considered over–the–counter as they do not involve a public exchange such as seen in stock markets like the NY Stock Exchange. This approach is based on the work of [159] and [160], who explore bilateral negotiations and liquidity flows in similarly non-transparent markets, with industry bodies contributing similarly in [161] and [162].

Also important is the balance between agent-to-client servicing and inter-agent trading, which forms the foundation of the market's network viability (and is partially observable through real-world data) [4]. In this study, I adopt the percentage of inter-agent trades as my primary metric, serving as a proxy for both liquidity and stability in the simulated bond market. This choice reflects the established relationship between these two characteristics in fixed income markets and is observable from reputable publicly available government data.

Therefore, my research is focussed on ABMs that model the market dynamics of secondary market trading. Moving forward, I will refer to “trading markets” with the reader’s understanding that I am referring to government bond trading specifically.

3.2 Key hypotheses and findings

I investigate several aspects of market liquidity and stability. A key outcome of the model will be insights into the interaction of model features, a second-order effect of trading that is currently little researched but a focus of regulatory scrutiny.

Firstly, I hypothesise that it is possible to replicate the market for Australian Government Bonds in a coarse fashion, with an ABM using stylised facts and publicly available government data points.

Using this market design as the base case, my second hypothesis explores the value of agent diversity, specifically the importance of a broad variety of clients for each agent. I test for this feature’s stand-alone impact on interactions and trading liquidity. From the analyses, it is evident that for systems with small agent populations, expanding the breadth of the client base is not a universal remedy for enhancing stability, trading, or longevity. Instead, these results appear to be more influenced by the agents’ business

[4] For clarity: recall that in this model market makers are agents and clients are not agents per se, but rather represented as locations on the landscape. In this way, agents service clients by moving through the environment while agents can also trade with other agents. Clients do not have their own movement features and are therefore static. Also, clients cannot trade with clients. These assumptions are somewhat relaxed in the final model TRIBE(LLM) in Chapter 6.

costs, or essentially the barriers to entry and sustaining an agent in the field.

The third hypothesis tests the value of increasing the number of agents interacting with the same (not mutually exclusive) client base and if this improves the market. Through this work, I investigate the impact of homogeneity on simulated markets by testing a more constrained set of agent variables on a larger agent population. I test the hypothesis that it is agent diversity, rather than simply increasing the number of agents, that has a positive impact on interactions and trading liquidity within the simulated bond market.

My final hypothesis questions the impact of increased regulatory costs on markets. This section examines how reducing agent costs and easing trading restrictions influence market stability and the longevity of agents within the simulated bond market. By analysing agent lifespan and liquidity provision, I uncover the relationship between operational costs and market dynamics. The results show that low agent costs tend to enhance agent longevity, which can be associated with greater stability in market trading. These results show the complex relationship between agent characteristics, market design, and liquidity, with implications for regulatory policy.

As with any ABM, the number and variety of simulation tests is practically unlimited. To illustrate and test my hypotheses, whilst maintaining a digestible amount of data, I contrast and explore two market scenarios for each hypothesis, where each scenario consists of 100 simulation epochs of agent models. Within each simulation itself, agents have a range of values for key parameters which is selected at initialisation of that epoch. This essentially provides 200 permutations and combinations of parameter testing per hypothesis. The ABM is validated through two methods: design based on documented current market practices (incorporating these as model variables using stylised facts), and secondly, verifying that the ABM can replicate observable data.

3.3 Model: a bespoke ABM implementation based on the Sugarscape model

I implement a custom ABM, drawing heavily from Sugarscape, the model introduced by [25] and used to simulate complex, artificial societies through agent-environment interactions. Many adaptations of this work exist [26–30]. The model's ability to represent small populations of heterogeneous agents, interacting asynchronously, aligns closely with the bilateral structure of government bond markets, making it an ideal foundation for the bespoke ABM used in this research. Other ABM model frameworks were considered (such as [223]); however, these tend to require extensive calibration and do not suit the bilateral market-focus of this study. To develop my bespoke model, several open source tools were used, including GitHub repositories [224], [225] following the framework provided by [26]. I adapt Sugarscape's terminology to the bond market context: the original resource variables are transformed into bonds and cash holdings, while retaining the "vision" parameter to represent the breadth of a client base which each MM has. The metabolism mechanism is repurposed to model operational cost structures within the trading environment for market-making agents.

Formally, in this chapter an ABM is defined as the n–tuple:

Definition 1. $\mathcal{M} = \langle N, \mathcal{L}, \mathcal{S}, \mathcal{P}, \mathcal{D}, \mathcal{A}, f, T \rangle$, where $N = \{a_1, a_2, \ldots, a_n\}$ is the set of market making agents. $\mathcal{L}$ represents the set of landscape states, each containing assets $R = \{b_r, c_r\}$, corresponding to bond and cash quantities of individual clients. Each agent a_i has a state $\mathcal{S}_i$, which includes bond and cash usage (i.e. their individual cost structure) rates m_b, m_c, accumulations A_b, A_c, and a client base breadth v_i. The perception functions $\mathcal{P}_i$ map the landscape and other agents' states to a perceived state, $p_{ij} : \mathcal{L} \cup \mathcal{S} \to \mathbb{R}$, while the decision rules $\mathcal{D}_i$ map perceptions and states to actions, $d_{ik} : \mathcal{P}_i \times \mathcal{S}_i \to \mathcal{A}_i$. Agent actions $\mathcal{A}_i$ include client servicing, trading with other MMs, and ceasing operations. The landscape evolves according to the transition function $f : \mathcal{L} \times \mathcal{A} \to \mathcal{L}$, with no bond or cash replenishment(*), and the model progresses through discrete time steps, $T = \mathbb{N}$.

*** Note:** The assumption of replenishment of assets is a common feature of the Sugarscape model and is a mechanism introduced in [25] which creates an infinite environment. Replenishment allows agents to assume that assets or resources will be constantly available. This assumption is a choice within the original Sugarscape and is at odds with a bond market with fixed number of bonds on issuance. As such, within work in this thesis, models assume that there is no replenishment of any assets in any model environment. This aligns with the market structure of government bonds where issuance is controlled by central banks and is a fixed quantity [154].

3.3.1 Agents

The model consists of N agents, representing MMs. I will use the term agent and MM interchangeably. In the Australian market, historically, four key institutional MM firms dominate, reflecting a structure in which few MMs interact with a large number of clients, as noted by [160]. Whilst others hold regulated functions, in this model I have focussed on the key four firms as agents. In addition, I model a cost structure for agents to represent the fixed operational costs incurred by the agents per time step. This includes the cost of meeting regulatory requirements.

3.3.1.1 Vision: client base and breadth

MMs are required to buy and sell bonds from their clients – bond holding firms – under the terms of their government registration. Within the model, the difference in the size of the MM is characterised by the different breadth of the agents' clients (that is, the vision range of the grid) v_i, determining how far they can access client resources in the landscape and the difference in the initialised bonds and cash holdings of an MM. These MMs are assigned a fixed client range, modelled as a form of grid "vision", allowing them to "see" client locations within a maximum grid distance at each stage. This enables MMs to locate client resources within their field of vision on the grid.

For an agent at position (x_i, y_i), the set of visible cells is:

$$\mathcal{V}_i = \{(x, y) \mid |x - x_i| \leq v_i \text{ and } |y - y_i| \leq v_i\} \quad (3.1)$$

This range is randomly assigned as $v_i \sim \text{Uniform}(0, 49)$ so that the maximum that any MM can access is the entire grid (that is, less than 50 units) and remains constant during the simulation. During certain tests later in the work, the maximum value will be tested such that $max(v_i) \leq 49$.

3.3.1.2 Cost of doing business: metabolism

Each MM incurs a cost at each time step, representing various business overheads, modelled as a form of metabolism. Let $Met_{(t,a,s)}$ be the initial metabolism for each agent. The bond and cash holdings of an agent i at each time step are updated as follows,

$$A_b(t + 1) = A_b(t) + (b_r - m_b) \quad (3.2)$$

$$A_c(t + 1) = A_c(t) + (c_r - m_c) \quad (3.3)$$

where m_b and m_c represent the bond and cash metabolic rates. These costs are assigned at the start of each simulation and remain constant throughout each simulation, influencing an agent's ability to survive and stay in business.

Initial Resources at Simulation start: At initialisation of the model (i.e. time step = 0), all agents are equipped with a starting value of bonds and cash. The distribution of these amounts aims to model the differences between agents.

3.3.1.3 Agent decision Rule 1: MM agents interacting with clients

Client locations can been "seen" by MMs, so long as clients are within a maximum grid distance from the MM, at each stage, in line with the MM's assigned v_i for each simulation. This enables MMs to locate client resources within their field of vision on the grid. MMs acquire (or "harvest") bonds and cash from passive clients, reflecting the key aspect of OTC markets where all client trading involves a MM, which is a feature captured in the design of the ABM presented.

3.3.1.4 Agent decision Rule 2: MM agents interacting with other MMs – introducing "Trading"

After servicing clients at each time step, MMs consider potential trades with other agents within its field of vision, provided the trade improves the welfare of both parties. Welfare improvement addresses imbalances in cash or bonds held by an MM at each time step, aligning with rational and legal trading behaviour as outlined in the finance literature and regulations [139]. *Note: For clarity, the time subscript (t) is omitted from these calculations. All values are computed at each time step in the implementation.* Each agent's welfare for bonds and cash is calculated as,

$$\text{Welfare}_{a_i,b} = A_b^{\frac{m_b}{m_b+m_c}} \tag{3.4}$$

$$\text{Welfare}_{a_i,c} = A_c^{\frac{m_c}{m_b+m_c}} \tag{3.5}$$

where A_b and A_c are the agent's bonds and cash accumulations, and m_b and m_c are the respective costs. Small cost values are used relative to initial accumulations.

For a trade to occur, agents compare the product of their current welfare to the potential new welfare (i.e. $\widehat{\textbf{Welfare}}$) what will occur should a trade be done:

$$\widehat{\text{Welfare}}_{a_i,b} \times \widehat{\text{Welfare}}_{a_i,c} \geq \text{Welfare}_{a_i,b} \times \text{Welfare}_{a_i,c} \tag{3.6}$$

Note: Other welfare calculation methods could be used, including the sum of welfare. However, the product method imposes a significant penalty when $Welfare_{a_i,b}$ or $Welfare_{a_i,c}$ falls **below 1**. The product approach favours rectifying significant imbalances with sub-optimal values. For example, consider the pre-trade scenario where $\text{Welfare}_{a_i,b} = 0.5$ and $\text{Welfare}_{a_i,c} = 20$:

- Using a product function: $0.5 \times 20 = 10$
- Using a sum function: $0.5 + 20 = 20.5$

Now consider a post-trade scenario where $\widehat{\text{Welfare}}_{a_i,b} = 1.5$ and $\widehat{\text{Welfare}}_{a_i,c} = 10$:

- Using a product function: $1.5 \times 10 = 15$ (an improvement from 10)
- Using a sum function: $1.5 + 10 = 11.5$ (a decline from 20.5)

From the perspective of a MM, this post-trade scenario represents an improvement as bond accumulations increase threefold (from 0.5 to 1.5) (which in turn allows the MM to keep trading as a 0 holding halts trading and "kills" a MM). By construction, neither cash nor bond accumulations can be zero as this would represent an agent who is "dead" and therefore not able to trade, hence all welfare calculations will be greater than zero. The product test correctly identifies this as beneficial ($15 > 10$), while the sum test incorrectly rejects the trade ($11.5 < 20.5$).

The Welfare test condition must be satisfied for both trading agents, ensuring mutual welfare improvement. This eliminates the need for a price mechanism, as agents trade based on relative needs, reflecting the limited price transparency in the modelled markets. The agent offering the greatest welfare enhancement is selected for the trade:

$$\textbf{Find-best-agent} = \max_{a_n}\left(\widehat{\text{Welfare}}_{a_n,b} \times \widehat{\text{Welfare}}_{a_n,c}\right) \tag{3.7}$$

Calculating trading quantities

Once two agents have been selected to trade, these two agents next compare their Marginal Rate of Substitution (MRS), indicating their relative need for bonds or cash. Trades occur when agents have compensatory needs, and the quantity of exchange is determined using the geometric mean of their MRS, reducing the bias of extreme values as proposed by [226].

The MRS for an agent a_i is defined as,

$$\text{MRS}(a_i) = \frac{\frac{A_c}{m_c}}{\frac{A_b}{m_b}} \tag{3.8}$$

where a higher MRS reflects a stronger preference for cash over bonds, and vice versa, based on survival needs in the model. This mechanism allows the ABM to avoid using prices for assets. Honesty between agents is assumed, and a relative trade occurs between those with the greatest desire for assets. This assumption appears appropriate given the lack of publicly available price transparency in bond markets. Other methods for calculating MRS could include randomising pricing or adopting alternative sampling distribution assumptions, which future research could explore further.

3.3.2 Clients

Clients are represented by unique, static, cell locations on a 50×50 grid, totalling 2,500 distinct clients (aligned with data from [143]). In this model they are **not** referred to as agents, and they are passive members of the simulation. Following the structure of Sugarscape, the choice of a 2-dimensional grid efficiently models both client locations and proximity to MMs. This spatial structure enables meaningful distance relationships between clients and MMs (utilised in the "vision" parameter), with visibility across the landscape while supporting movement of MMs in 8 directions instead of just 2. The

2-dimensional environment allows the possibility for MMs to cluster regionally around resource-rich client areas (as might be expected in real life) and form possible spatial boundaries that couldn't exist in a 1-dimensional model.

Each grid cell (x, y) corresponds to a client holding bonds $B_{(x,y)}$ and cash $C_{(x,y)}$, modelled as: $C_i = R_{(x,y)} = (B_{(x,y)}, C_{(x,y)})$. Each client is assigned a heterogeneous quantity of a generic government bond and a cash amount at the start of each simulation. Bonds and cash are distributed in mounds across the grid, with values decreasing from the centre of the mounds outward upon the grid (for further details see Appendix B).

Due to the lack of an observable real-world distribution of bondholders, I chose to concentrate these client resources in four distinct regions on the grid. Future research may refine this assumption as more data becomes available. Each cell in the grid (x, y) contains resources, specifically bonds and cash, and is related to one unique client. This assumption was taken to reflect the stylised reported data on the number of bond-holding clients in the bond markets of interest [143]. Further information on the method of Client Resource allocation within the presented ABM can be found in Appendix B.

I acknowledge that alternative research directions could emerge by varying the assumed number of clients. One area for exploration is developing a more dynamic approach to handling client assets. Other methods of designing a landscape (i.e., client distribution) including agency warrant investigation, such as testing various landscape assumptions and sizes. I present an extension of client modelling in this vein, in Chapters 5 and 6.

Note: In this specific model, I define **trading** as the exchange of assets exclusively between two MMs. Trading is inherently an agentic action between two participants. Therefore, transactions where agents collect bonds and cash from clients are not considered trading, as they constitute "client servicing" rather than discretionary trading decisions. This terminology is adopted only within this chapter for clarity.

3.3.3 Model dynamics

The dynamics of the internal model is governed by the interactions of the components in $\mathcal{M}$, which are processed iteratively as follows (code workflow is described in Figure 3.1):

1. **Initial agent location**: At time $t = 0$, each agent a_i is randomly assigned a location on the landscape, which is structured as a grid of size $X \times Y$. The location of each agent is represented as a pair (x_i, y_i), where x_i and y_i are integers satisfying $0 \leq x_i < X$ and $0 \leq y_i < Y$. The random assignment ensures that no agent can be considered to have a preferential location at the beginning of the simulation. At initiation, landscape variables and agent features are randomly set within predefined parameters. Further details about landscape (client resource) design can be found in Appendix B.
2. **Perception:** Each agent $a_i \in A$ perceives the state of the environment $\mathcal{L}$ within their field of vision across the landscape, v_i and the other agents' states $\mathcal{S}$ through its perception function $\mathcal{P}_i$. Agents do this asynchronously and in random order at each step, t.
3. **Decisions:** Based on its perceptions $\mathcal{P}_i$ and current state $\mathcal{S}_i$, agent a_i selects an action $a_{it} \in \mathcal{A}_i$ at time t, according to its decision rule $\mathcal{D}_i$.
4. **Action:** The chosen action $a_{i,t}$ affects the landscape;
   (a) **Resource Collection and Transaction:** Each agent a_i can collect bonds and collect cash from clients. It does this by looking within its vision and finding the client with the largest and nearest quantity of bonds and cash. This precise mechanism is discussed in more detail below. The agent's respective accumulations are updated as part of consumption. After the resource's bond and cash values are reset such that $b_r = 0$ and $c_r = 0$, this mimics the concept that each client only trades once during a typical trading day, or not at all.
   (b) **Resource Consumption:** Each agent increments its bond and cash holding with the collected resources and decrements its bond holdings and cash holdings

at each simulation step to represent the ongoing cost of staying in business, such that Bond holdings accumulation becomes: $A_b + = (b_r - m_b)$, Cash accumulation: $A_c + = (c_r - m_c)$. Client resources are always accumulated first. This embeds an assumption that agents preference servicing clients by accumulating their resources before considering to trade with agents. This assumption aligns with market regulations [139]. Resource consumption (and inherently agent to client activity) is a required action of agents.

(c) **Cessation of Operations:** A MM (agent) ceases business operations when they run out of resources such that either of $A_b = 0$ and or $A_c = 0$. This condition effectively stops the agent's market-making activities for the remainder of the simulation, representing a business failure due to resource depletion.

(d) **Trading with Other Agents (i.e MM to MM trading):** Agents engage in trading activities with other agents if a utility valuation difference exists after consuming the resources of a client. This is modelled as an agent choice, based on the other agents within their field of vision and involves the exchange of resources, bonds, or cash between agents based on their respective perceived valuations and strategies. Unlike agent-to-client interactions, resources are swapped between agents. Agents aim to optimise their own holdings or utility.

5. **State Update:** The environment $\mathcal{L}$ is updated using the transition function f based on a singular action of agent a_i at $t > 0$. This includes updating b_r and c_r (landscape variables) and also A_b and A_c (agent values) in line with the above action possibilities at each and every $t > 0$.

6. **Iteration of agents:** The process repeats over all agents $a_i \in N$ at time $t > 0$, with the state of each agent $\mathcal{S}_i$ and the environment $\mathcal{L}$ evolving as the model progresses. The ABM does not allow any agent or client coalition formation, in line with real-world regulatory constraints on MMs and clients.

7. **Time Steps:** Each epoch as at most 1500 time steps, which account for 120 trading decisions per virtual hour for a typical 12.5 hour long trading day i.e. 2 trading steps per minute. The process then iterates to the next time step $t + 1$ until either such

time as all agents have ceased to operate or the simulation has reached its maximum time T. Within each time step, all agents still "alive", randomly and asynchronously, take turns deciding to move.

8. **Equivalence Classes:** In constructing an ABM of the bond market, I have operationalised regulatory principles by implementing decision-making protocols based on objective criteria. This approach deliberately avoids coalition formation or subjective preference functions, as such mechanisms would contradict the spirit of real-world market regulations. Mathematically, all MMs belong to a single equivalence class E_N, where any MM agent $N_i \in E_N$ is bound by the same set of decision rules for functional properties as any other MM (though each N_i has different initial condition sets). Similarly, all clients (grid locations) belong to an equivalence class E_C. Again, all grid locations (C_i) are bound by the same set of functional properties decision rules, but each $i \in C_i$ has their own initial asset conditions. This formulation ensures that the model treats MMs as functionally interchangeable entities, while clients operate with counterparty-neutral selection criteria, engaging with MMs through predefined parameters and methodological frameworks. This design choice preserves the arm's-length nature of regulated financial market interactions while maintaining mathematical rigour through well-defined equivalence relations.

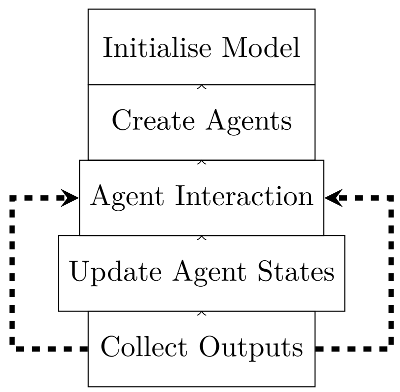


Figure 3.1: Agent-Based Model Workflow, see Section 3.4 for full details.

3.3.4 Data

I make use of publicly available data from the AOFM – part of the Department of Treasury of the Australian government, which reports the amount of MM to MM trading as a share of total secondary market trading in Australian government bonds on a quarterly, averaged basis. This data is available from September 2016 to June 2024, aggregated across all market participants and shows a typically long run average of 25-30% of turnover belonging to MM to MM interactions. See Table 3.1 and Figure 3.2.

The work of [143] at the UK central bank, the Bank of England, discusses the limitations in the government bond markets in detail. Given these constraints and the scarcity of comprehensive trading data, my bespoke implementation is based on features derived from stylised facts and government documentation.

Statistic	**Value (%)**
Mean	27.23
Median	27.27
Quarterly Std Dev	4.54
Minimum	15.94
Maximum	36.69
First Quartile (Q1)	23.85
Third Quartile (Q3)	30.17

Table 3.1: Aggregated quarterly summary statistics of interbank trading volume market share: Sep 2016 to June 2024.

3.4 Market internal mechanism: simulation process

Each simulation contains time steps, t. Within each time step, all agents (which are "alive" at that point in time) randomly and asynchronously take turns deciding to move. I employ the following process:

1. **Initialisation**: The simulation begins by initialising the landscape and the agents at $t = 0$.

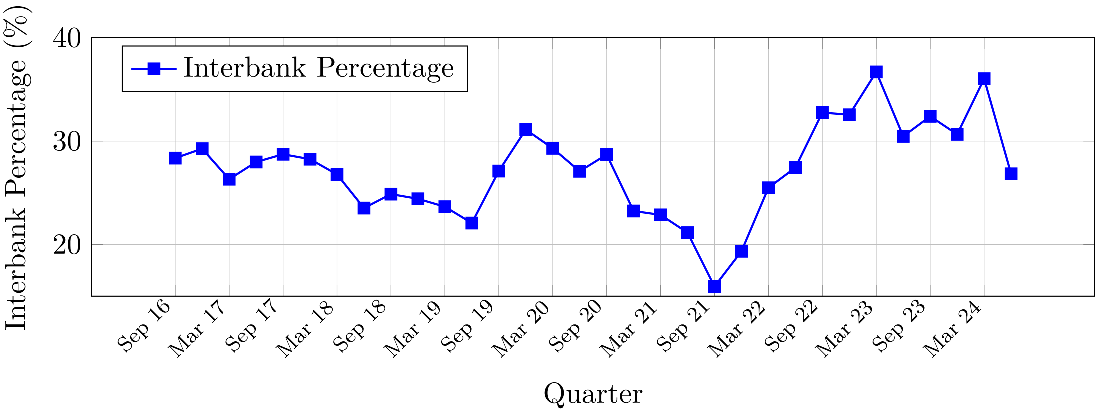


Figure 3.2: Quarterly interbank percentage of secondary Australian government bond trading.

(a) Initiate the simulation with N agents (matching the number of market making banks).

(b) Endow agents with predefined characteristics (akin to genetic features) which are constant for $\forall t$ within a single epoch and set at $t = 0$:

- Vision, representing the breadth of client coverage
- Metabolism for cash and bonds, indicating the agent's "carrying cost" or the cost to stay in business.

(c) Initial bond resources (akin to each trading firm's initial daily position) set at $t = 0$ and then updated at each t

(d) Assign each agent (MM) a randomly determined starting grid location at the inception of every simulation. Set at $t = 0$ and then updated as at each $t = t$

2. **Iteration through steps** t for all agent actions detailed above and summarised below:

(a) Conduct each interaction simulation asynchronously over time steps. Loop through each of the following in order, but for a random order of agents a_n:

i. Resource Accumulation: by agents from the clients around them

ii. Move: to location where the chosen client to accumulate from resides

iii. Agent Decision Making (and possible trade): use the "find best agent" equation 3.7 method to identify the best agent (if any) to trade with, considering the resources available and their current welfare.

iv. Survival: Agents check at $step = t$ if $A_c > 0$, or $A_b > 0$, if neither is true, then their accumulated resources have fallen too low, and they 'die' in the simulation at this point and are recorded as ceasing business. They are no longer active in the simulation.

(b) Retain the same genetic factors at the beginning of each iteration within one simulation epoch, regardless of the number of time steps.

3. **Repeat** For each time step go through the above for any remaining alive MMs until all agents cease to operate or the simulation reaches its predefined termination time point (1500 steps).

3.4.1 Assumptions and limitations

Any reduced-form model will necessarily have restrictions and assumptions. Whilst many small assumptions have been detailed through the above discussion, I focus on a number of larger fundamental assumptions here:

Passive clients and information revealing I make the assumption of passivity of the clients and the market landscape, not expressly allowing clients to move across the grid, or have any agency. For instance, I have made the assumption that when an agent looks across a client landscape that agents have access to truthful client resource values (i.e. all clients honestly report their resources) and at all times clients are available for resource evaluation. This is akin to assuming that all clients are always ready and willing to trade and reveal complete information to agents, at all time steps. Whilst I have not explicitly modelled any nefarious usage of this information revealing within the model, it is a constraint of the model that clients are essentially passive, transparent, participants within the simulations, with no agency. In order to address this I include modelling clients directly within the ABM in a later chapter with a different model, which I call the TRIBE model. However, the current bespoke model in this chapter contains enough flexibility to analyse agent-to-agent interactions, which are the particular concern of this chapter.

Agent revealing of information: Memoryless agents When agents evaluate other agents in the landscape, they are truthful to the other agent about their resources (which flows into the logic to calculate the relative value that then controls the decision of an agent to trade with the other). I have also made no provision for agents to store data of others' accumulations along the way. Agents themselves do not consider or rely upon their own accumulations over time, only the level at each time step. This is essentially an assumption on information revealing by agents to others and themselves. It is also a limitation whereby agents are effectively "memory-less". Also, I have assumed that when an agent looks across the environment and sees a client resource that (a) agents have truthful knowledge of the accumulation levels of clients and (b) that the agent wishes to accumulate all the values of resources that the client holds. Much like the assumptions around the limitations on client agency, the assumptions and limited concern for agent revealing are in line with the modelling goal of simulating agent-to-agent interactions.

3.4.2 Model simulations

Taking the above limitations and assumptions as read, there remains an extensive number of possible model permutations and combinations of variables that can be tested. To summarise, these include:

1. Agents: number of agents, initial holdings (bond and cash), costs (in terms of bonds and cash) and client breadth (also known as "vision"). This contains 6 variables.
2. Landscape tests: grid size, max and min values for client bond holdings. This contains 3 variables.
3. Action set test: steps of actions, calculation methods, decision logic. This contains a minimum of 3 variables.

In the case of agent and landscape variables, the values of each variable can range from 1 to an infinite number. Clearly, the possible space of all model permutations is much too large to address here. Rather than relying on statistical methods to reduce the complexity or dimensionality of the model, I instead rely on stylised market facts to build a model that replicates market design features. This practically constrains many of the size and range of agent values and landscape variables. Action sets are also largely determined by the objective of the model study.

3.4.3 Parameter sampling

To address the very large possible set of parameter combinations and permutations, I made use of a two-tier parameter sampling approach. I combine random sampling at the simulation level of key parameters (e.g., for agent vision, agent metabolism, agent initial bond and cash holdings, agent location, client locations, and client assets) with structured design set sampling at the meta-level (number of agents, breadth, and level of agent variable range). For statistical robustness, each model configuration is tested across 100 independent simulations, with parameters randomly initialised within meaningful bounds at the inception of each simulation (e.g., agent vision ranging from 1-50 units, metabolism rates between 1- 5 units, etc.). Although this approach effectively addresses specific research hypotheses about MM behaviour, it differs from more systematic sampling methods like Latin Hypercube Sampling (LHS) or iterative sampling techniques [227]. LHS could provide more uniform coverage of the parameter space by stratifying each parameter dimension and ensuring samples from every interval, potentially revealing edge cases or unexpected interactions that my method might miss. Iterative sampling approaches could adaptively focus simulation resources on combinations of variables of high variance. However, these tests would also open up the modelling results to scenarios where certain combinations of variables were illogical in the context of bond market dynamics or contrary to real-world dynamics (e.g. large numbers of agents for small numbers of clients).

Furthermore, researchers could employ Bayesian optimisation [228] or sequential sampling methods [229] to efficiently explore the high-dimensional parameter space, especially valuable for investigating rare events or market stress conditions. Other methods such as importance sampling [230] could be used to probe specific stress scenarios in the bond markets and extreme examples. These alternative approaches might be particularly relevant for studies focussing on market dynamic changes (e.g. major changes in legal aspects or trading competition) or regulatory parameter optimisation (e.g. small changes to market conventions and regulations). These supplementary approaches provide a foundation for future research extending ABM applications to other asset classes and exploring market designs that differ substantially from current bond market structures.

3.5 Results

The ABM is calibrated to the Australian MM environment using the design and model variables listed in Table 3.2, I conducted 100 simulations (epochs), producing over 200,000 agent interactions. I refer to this configuration as the "Australian calibration" or **(Hypothesis 1: HP1)**.

In summary:

- **Hypothesis 1: HP1**: Replicates the current bond trading marketplace, a model I term the 'Goldilocks' version and I refer to this configuration as the "Australian calibration".
- **Hypothesis 2: HP2**: What happens to agents who have a large and broader client base? Do we need more agents in the marketplace to increase market stability?
- **Hypothesis 3: HP3**: Increasing agent numbers for the same fixed client base.
- **Hypothesis 4: HP4**: Stress testing agent costs to double the level in **HP1** simulating a high regulatory cost environment.

3.5.1 HP1: Goldilocks – replicating the Australian MM environment

Test design

The design and model variables listed in Table 3.2 were used to calibrate an ABM with four agents (A=4) and a client base of 2500, with agent vision ranging from 1 to 50 units and metabolism between 1 to 5 units per time step. The initial accumulations of resources ranged from 35 to 55 units. More than 200,000 agent interactions were modelled in 100 simulations, each running up to 1500 time steps to emulate approximately 200 agent moves per trading hour. Agents all have the chance to make a move and trade in each time step, asynchronously, and the order of agents is randomised each step to avoid inbuilt order bias.

Variable	**Description**
Number of Agents	4
Client Breadth	1 to 50 units (random)
Cost	1 to 5 units per time step (random)
Initial Resource Accumulation	40 to 55 units (random)
Size of Client Base	2500 clients
Trading Occurrence	Median 28.7% per model simulation
Number of Simulations	Over 100
Time Steps per Simulation	Up to 1500
Agent Moves per Trading Hour	Approximately 200
Simulation Order	Asynchronous and randomised

Table 3.2: Summary of variables used in the goldilocks simulation

HP1: results

I successfully replicated the observed attributes of the Australian bond markets over 100 epochs using an ABM with specific calibrations, offering insights into optimal market structures. Simulations reveal a median MM-to-MM trading occurrence of 28.7% of all interactions. This median trading activity is closely aligned with data published by the Australian Office of Financial Management (AOFM), which reports an eight-year average interbank bond turnover of 27.3% of traded volumes [5]. This close correspondence validates the model's ability to replicate real-world trading patterns.

The model results demonstrate a high degree of variance in the simulations, as illustrated in Figure 3.3. This variance is expected given the large degrees of freedom in the initial calibrations for each simulation. On average, the trading frequency in the model aligns closely with the aggregated reported trading volumes, serving as a reasonable approximation for such an opaque market. Furthermore, the reported figures are aggregated over three months of daily trading without reporting variance or week-to-week volatility. However, the range of trading interactions remains broad, with summary statistics provided in Table 3.3 (rounded to one decimal place).

HP1: validation and analysis

We can think of each simulation as a theoretical trading day. Therefore, in contemplating the variation between simulation sets, implemented by the random selection of parameters in each simulation (drawn from a prefixed boundary set), it would seem to imply that from one trading day to the next, outcomes of agent interactions are highly variable. This is in line with the results of other financial markets and observed "chaos" documented in the literature [122] and [152].

Statistic	**MM to MM Trade Percentage**
Median	28.5
Std Dev	27.5
Min	1.2
25%	11.0
50%	28.5
75%	59.8
Max	95.1

Table 3.3: Summary statistics for MM to MM trade percentage across epochs.

Having tested the hypothesis that the current mechanism supports the "Goldilocks" mar-

[5] Data accessed 27 Sep 2024, quarterly figures.

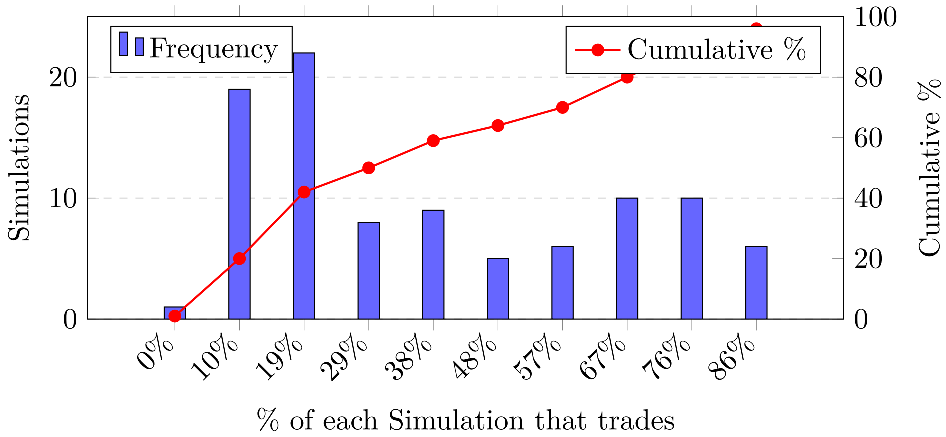


Figure 3.3: A=4 distribution across runs by trading percent (median 28.74%).

ket design, i.e. that the market design is not too rigid or fragile and has just enough diversity for agent behaviour, I find that the ABM can replicate market returns and does not strongly show a relationship to a specific variable. Overall, my findings provide evidence towards a "Goldilocks" scenario, suggesting that the optimal market structure might be one that closely aligns with the current configuration of the Australian bond market and relies on a variety of variables.

3.5.1.1 Analysis: impact of small metabolisms and welfare

The model employs relatively small metabolism values compared to initial agent accumulations to ensure meaningful simulation durations. At the inception of each model epoch, agents are randomly assigned metabolism rates between 1 and 5 units (inclusive) for both bonds and cash. Given their initial resource accumulations, these metabolism rates effectively guarantee a minimum agent lifespan of 10 to 50 time steps. Although it is not feasible to analyse all possible combinations of bond and cash accumulations with their respective metabolism rates, the sampling approach I use provides sufficient coverage of the parameter space.

The analysis reveals a significant interaction between bond and cash metabolism rates in determining agent welfare, as expected. Specifically, agents with asymmetric metabolism rates, high for one asset and low for the other, tend to achieve higher welfare levels. This asymmetry is ideal to promote welfare-improving trading interactions between agents at different local maxima and minima, aligning with the market's desired behaviour. Although I implement a definition of welfare and marginal rate of substitution (MRS) from [25], alternative welfare calculation methods could be considered in future research, although the current approach adequately serves my analytical objectives. Alternative welfare methods are left to other researchers to explore.

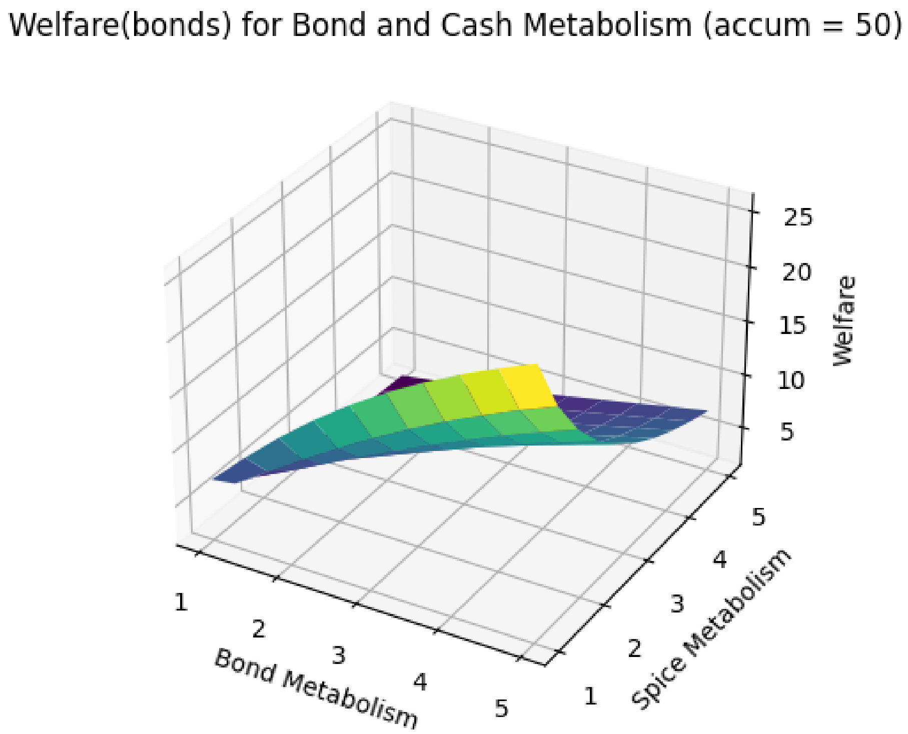


Figure 3.4: Impact of metabolism on agent welfare.

3.5.1.2 Spotlight: focus on simulations that most closely report Australian style results

Across simulation epochs there are nine epochs that closely match specific real market data on MM to MM trading activity. These epochs provide an opportunity to analyse the variables of interest. I begin by focussing on nine simulation epochs that produced trading percentages very close to the Australian market reported value. In Table 3.4 the various simulation vision statistics are explored, where vision can be randomly selected from 1 – 50 units. The results do not show a clear pattern.

Run	Mean Vision	StDev Vision	Vision (minmax)	Coeff. Variation	Trade%
14	18.5	7.3	14	0.4	26.9%
25	24	5.5	12.5	0.3	31.3%
44	22.8	15.5	37	0.7	27.3%
48	30.8	14.4	31	0.5	25.0%
49	13.8	4.9	10	0.4	31.1%
66	19	5.9	13	0.3	26.2%
70	28.3	20.7	39	0.7	28.8%
84	37.3	10.7	25	0.3	28.4%
91	18	18.5	43	1.0	28.7%

Table 3.4: A comparison of simulations with similar trade percentages: a deeper look a agent vision details.

3.5.1.3 Spotlight: agent vision across all epochs in HP1

Turning to exploring results of HP1 over 100 epochs, there is a clear trend of higher vision allowing for a higher trading percentage, but this is not a guarantee of higher trading percentages. In Figure 3.5 the vision of the agents is explored as a k cluster (where all agents are included in all epochs).

However, when analysing results at the level of epochs (and not individual agents), a less clear pattern emerges (see Figure 3.6). We see that at the level of each epoch, the agent visions that constitute the 4 agents within each epoch have a variety of impacts. This points to the interconnectedness of agent parameters within epochs and simulated societies as a whole.

The results would suggest that many combinations of agent vision can achieve very similar results (trade percentages on the y axis can be formed with both high and low agent visions) in Figure 3.5. Whilst agent vision within a cluster is high, the plots show that each

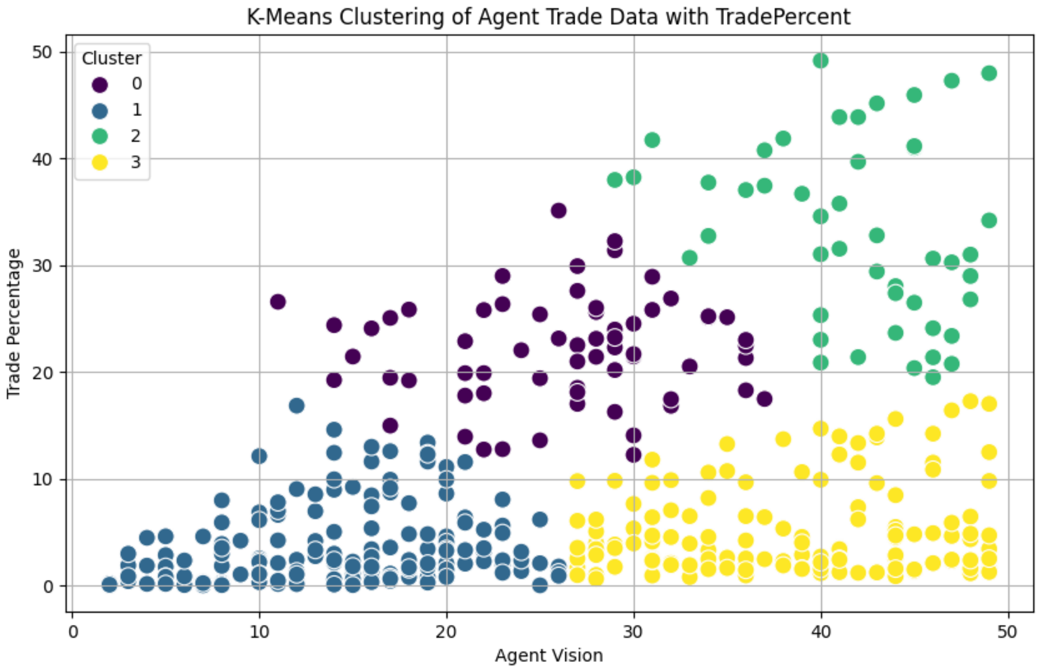


Figure 3.5: All epochs, 4 agent population k-mean scatter plot of trading Percentage by vision per agent across all epochs. A clear trend emerges that higher vision **allows** for higher trading percentage but this is **not a guarantee** of higher trading percentages.

cluster retains a similar trade performance, with only slight variations between them. This is despite the results forming distinct agent vision groupings in Figure 3.6 where clusters form into low trade percentages (cluster 1), mid percentages (2 and 3) and percentages (cluster 0) – each cutting across a large range of agent visions.

Likewise, a simple correlation statistic of the relationship between an agents vision and their life span, across all epochs and agents produces a result of 1.5% for agent vision and life, while the correlation between agent vision and MM trading percentage is higher at 52.9%. It still remains little better than a coin-toss. Agent vision, alone, is not a unique or linear descriptor of trading percentage, or indeed lifespan of an agent.

In addition, the ability of the model to replicate significant trading patterns opens avenues for exploring market "chaos", as documented in the literature [122, 152]. Additionally, extensions such as estimating specific traded volumes and investigating nonlinear impacts

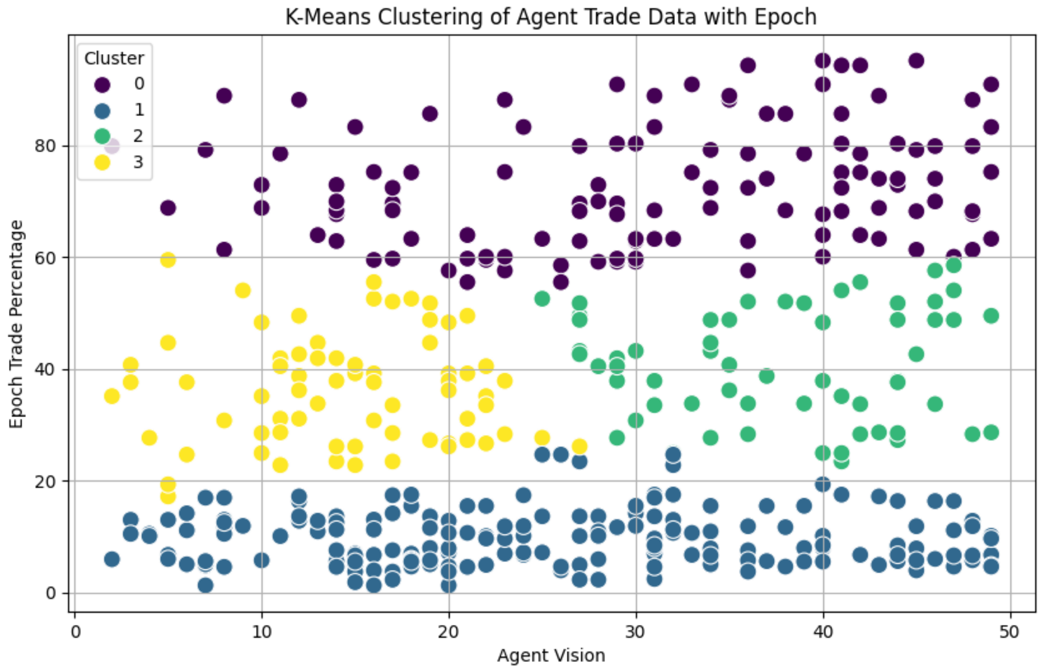


Figure 3.6: All epochs, 4 agent population k-mean scatter plot of Epoch trading percentage by the vision contained by the agents within the epoch.

of agent vision are proposed for future studies. To better understand the impacts of variables, I further test a number of hypotheses related to vision, the number of agents, the cost of business (metabolism) and the interplay between regimes of variation below.

3.5.2 HP2: Client breadth – introducing agent "vision"

This hypothesis explores the value of agent diversity, specifically the importance of breadth of clients (agent vision), rather than simply increasing the number of agents, and if it has a positive impact on interactions and trading liquidity. To illustrate the findings, two examples are presented, each consisting of 100 simulation epochs of agent models.

The intricate dynamics between agents, their feature sets, and the environment can be explored by considering an agent's 'vision' as a proxy for the breadth of their client base, while holding all other variables constant. This is pertinent to financial markets, where banks and financial institutions often aim for economies of scale, and the emergence of "fintech disruptors" introduces a wider mix of agents with diverse client bases. Furthermore, regulatory strategies that promote competition often encourage the entry of smaller, newer market players, which inherently have a narrower client base compared to established participants. These scenarios can be evaluated by investigating the impact of an agent's "vision" on simulations.

Keeping $N = 4$ and maintaining the same costs, I test a reduction of the possible range of client base sizes (vision) from a theoretical grid maximum of 50 units to just 5 units. In each epoch, agents are randomly assigned a value within this reduced range for the duration of the simulation in line with previous model design methods. By narrowing the range from 1–50 units down to 1–5 units, there is a significant decrease in the maximum client base size and reduced agent-to-agent diversity in client base sizes (as agents can now assume only one of 5 possible client size values instead of one of 50 possible sizes). The result is a series of simulations where almost no MM-to-MM trading occurs. In fact, over 73% of epochs exhibit no MM trading whatsoever (see Figure 3.7). This contrasts with the Australian calibration, where agents who are allowed to have client ranges of 1–50 units engage in MM-to-MM trading on average 28% of the time, and 100% of epochs display some MM trading, however small.

In the context of financial markets, this is a crucial finding. Specifically, adding an agent with a low (or limited) 'vision' can still lead to successful outcomes provided the agent possesses the right supportive features. However, such success scenarios aren't common. In most simulations, agents with broader 'vision' (i.e. a larger client base) seem to have an 'easier' time trading with others. Interestingly, these effects appear reverse if higher agent costs are introduced. This suggests a nonlinear relationship between trading barriers (represented by increasing metabolism or agent costs) and client base expansion. While reducing agent costs directly improves agent welfare (measured as the number of trades and time to cessation), expanding client bases usually enhances welfare, but not when costs also increase. Hence, broadening the client base alone does not inherently guarantee improved agent welfare.

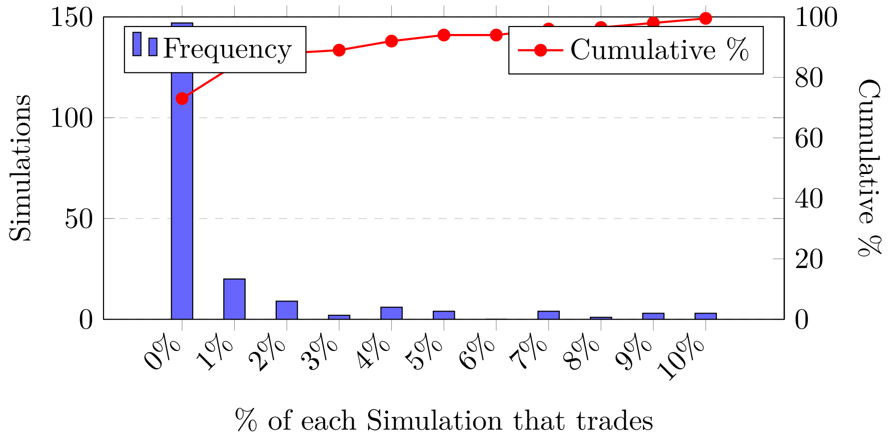


Figure 3.7: Stressing the HP1: Australian calibration to much lower range of vision effectively shuts down MM trading to **zero** in 73% of epochs with an average trade amount of less than 1% of all activity.

3.5.3 HP3: increasing the numbers of agents servicing the same fixed client base. Does having more agents in a market improve that market?

This section I test the hypothesis that agent diversity, rather than simply increasing the number of agents, has a positive impact on interactions and trading liquidity. Here, unlike in **HP2**, diversity is more than just the breadth of client base held by each MM agent.

3.5.3.1 HP3: effects of increasing agents to $N = 16$

To test the impact of increasing the number of agents, the simulation was configured with 16 agents -— a fourfold increase from **HP2**—while maintaining the same reduced client base (i.e. smaller vision) and business costs (i.e. metabolism). Having established the baseline model with 4 agents reflecting Australia's primary MMs, the model is now extended to 16 agents, matching the UK government bond market's historical average number of MMs. This configuration makes use of common structures among former British Empire financial systems to select a relevant number of agents to test. Comparing results across both configurations helps assess model robustness and generalisation to

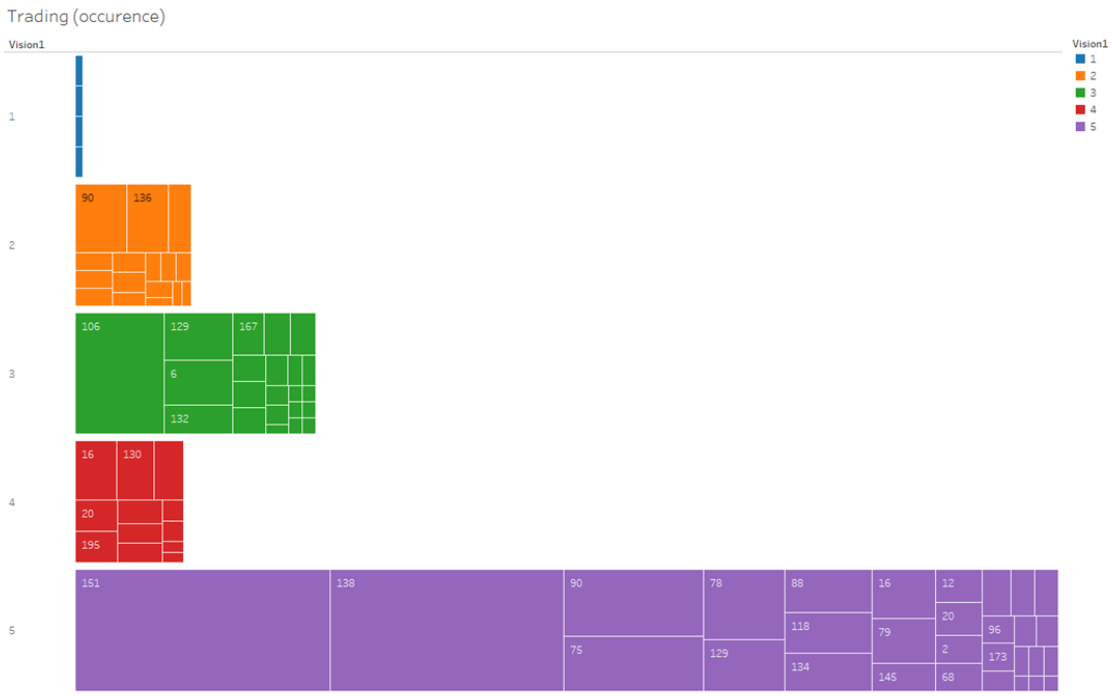


Figure 3.8: Agents with higher vision trade more, and also trade with other agents of similar higher vision.

markets with similar designs but different participant numbers. All characteristics of the agents remain consistent with **HP2**, including low vision ranges of 1–5 units and identical metabolism parameters. The simulation was run for 100 epochs to ensure sufficient data collection for analysis.

The findings revealed that while the increase in the agent population generated more trading occurrences than **HP2**, MM-to-MM trades remained notably low at 6.4% (see Figure 3.9), significantly below the levels observed in **HP1**. This persistent low MM to MM trading ratio, despite the expanded agent pool, suggests that agent population alone may not be sufficient to replicate the inter-dealer trading patterns observed in the Australian market. The results indicate that other factors, beyond mere market participant numbers, play crucial roles in determining the frequency of inter-dealer trading activity.

In this simple test, these specific examples demonstrate that the number of agents cannot make up for the benefits that come from having a wide variety of agent client sizes,

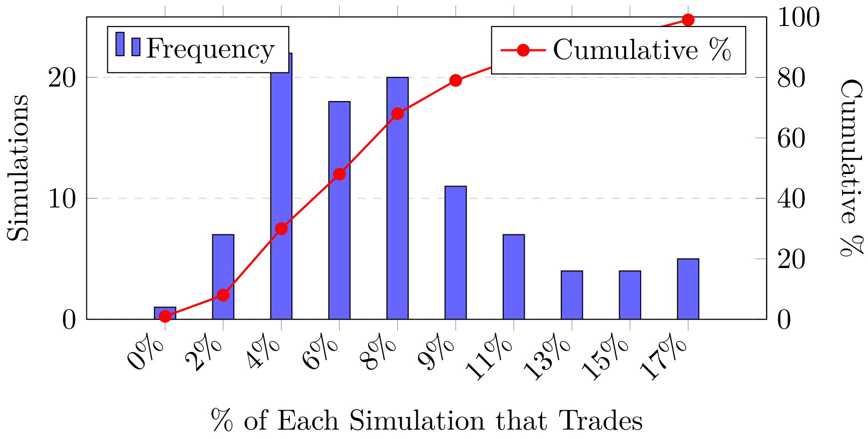


Figure 3.9: $N = 16$ – Distribution of trade occurrences as a percentage for each simulation. The cumulative distribution shows a mean of just 6.4%.

such as in **HP1**. Whilst running populations at various other sizes is possible, there is no reason to suggest the results would produce inconsistencies. Whilst, theoretically, stressing the simulation to use agents of 100, or even 1,000, might be interesting, they would fundamentally represent a different market paradigm divorced from that which is the goal of being tested. Also, testing such limits would require alterations on many other assumptions that would largely invalidate any results produced (e.g. client breadth can't be assumed to be the same, cost of business must surely be impacted with such large numbers of competing agents).

Agent Homogeneity: In comparing **HP1** to **HP3** I show that when increasing agents but reducing client size, and keeping cost levels (and range) low, a dramatic reduction of trading in all epochs occurs. Recall that in each of the 100 simulations, agents can select randomly (at initialisation) from a range of values for a number of variables (vision and metabolism). In producing a 100 epoch simulation, I am in effect testing 100 permutations on this market design, despite the optically simpler presentation of one test for one hypothesis.

The results of this market test in **HP3** are thus somewhat to be expected – each simulation agent had a smaller selection of a lower range of client options at initialisation, such that reducing client breadth has the impact of decreasing agent heterogeneity. In these scenarios

MMs simply have less differentiation between themselves. In effect, this scenario illustrates what happens when agent heterogeneity is structurally reduced. Even when increasing the number of agents to 16 (significantly beyond that of the Australian market calibration of just 4), the amount of trading between agents is significantly reduced from 28% in HP1 to under 7% in HP3 and virtually stopped in HP2. From this, it can be conceded that reducing the heterogeneity of MMs, (i.e. adding MMs who are more likely homogeneous) even if increasing their number, has the impact of reducing market trading activity and hence, reduces liquidity. Increasing the heterogeneity of MMs rather than simply the number of agents contributes to greater market trading activity and hence, liquidity.

3.5.4 HP4: the impact of lowering agent costs and market stability

Having explored the impact of changing agent numbers and the interplay with client breadth, in this final hypothesis I explore the effect of reducing agent costs and easing trading restrictions influence market stability and the longevity of agents within the simulated bond market. Restrictions and trading costs are an important component of regulated markets and are commonly considered desirable for market stability. Almost by design, market regulatory bodies will have market stability as a core value or principle. Hence, this hypothesis aims to test what happens when the financial cost of regulatory hurdles is altered.

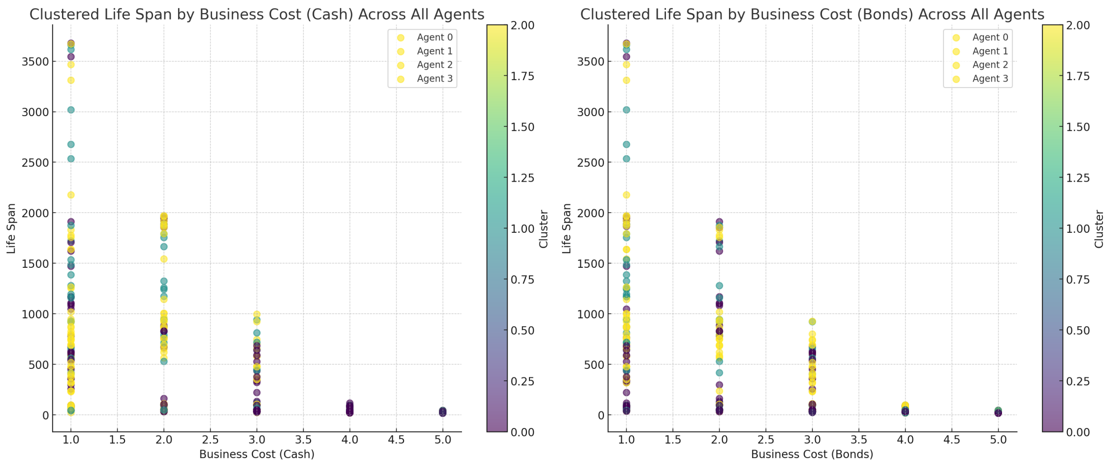
In order to do this, I analyse agent lifespan and liquidity provision and uncover the relationship between operational costs and market dynamics. The concept of agent lifespan, in terms of time steps, is evaluated in relation to stability and liquidity provision.

I utilise the two tests again, this time looking at populations of 4 agents across 100 epochs where client breadth is set to be diverse (1 – 50 units in line with HP1) but costs are set to be higher (with a range of 5-10 units). This test, **HP4** can then be directly compared to the tests using the Australian MM calibration simulations (i.e. **HP1**, which has the same calibration set but a cost range of just 1- 5 units. In summary, I conduct two primary tests to investigate the impact of agent costs on market stability:

Low-Cost Scenario (HP1: Australian Market Calibration):

- Agent Population: 4 agents
- Epochs: 100
- Client Breadth: Vision is set to randomly range from 1 to 50 units
- Cost Range: Low, between 1 and 5 units

HP1: Exploring the Australian market calibration In HP1, approximately 27% of interactions are between MMs and 68% of epochs had at least one agent still servicing clients and "alive" at the maximum time step of $t = 1500$. In Figure 3.10a and Figure 3.10b it can be seen that agents in epochs where they had lower costs have longer life spans.



(a) Impact of cash costs on agent lifespan.

(b) Impact of bond costs on agent lifespan.

3.5.4.1 Hypothesis 4: HP4

In this test I stress costs to double the level in **HP1**. Again, 100 simulations are carried out where agents randomly select from a range for each variable at initialisation. Thus in reporting one hypothesis test, results from 200 simulations are compared. The results are quite stark: average life spans of each epoch are less than 17 time steps, and fewer than 4,000 interactions across all 100 epochs. Individual epochs show significant early collapse (around 10 or so time steps), but, across a small number of time steps, there is a large degree of trading with other MMs (median of 46%). The key outcome appears to be that increasing (doubling) costs significantly weakens the trading landscape (see Figure 3.11) leading to situations where the landscape fails to function for clients, meeting the definition of instability provided by the World Bank [155].

High-Cost Scenario: HP4

- Agent Population: 4 agents
- Epochs: 100
- Client Breadth: Vision is set to randomly range from 1 to 50 units (aligned with HP1)
- Cost Range: Elevated, between 5 and 10 units double the level in **HP1**

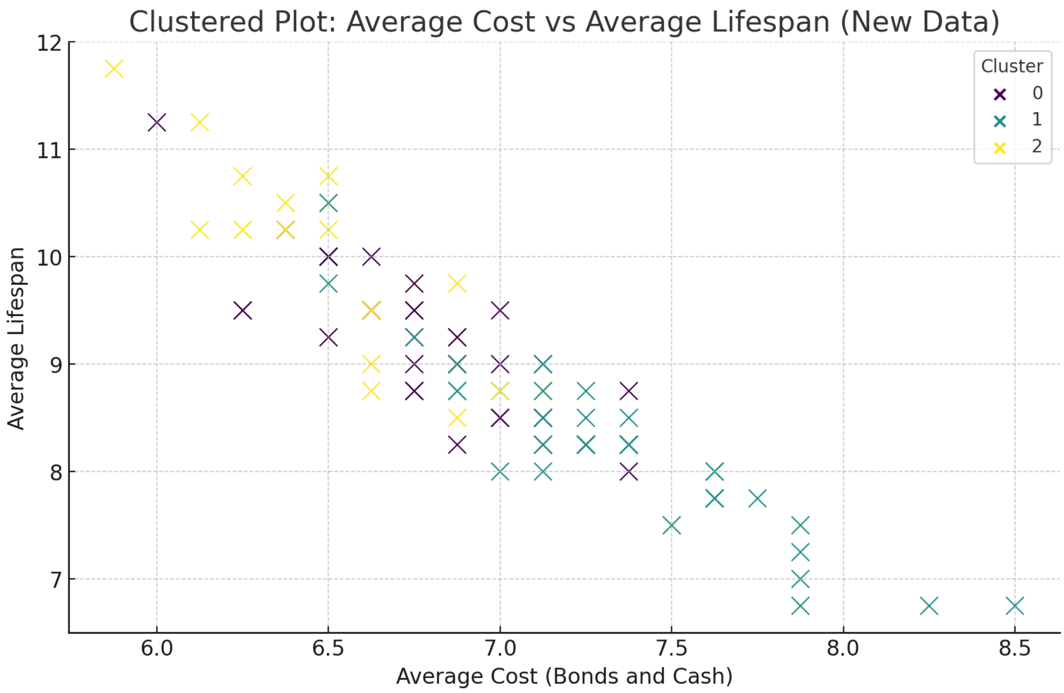


Figure 3.11: The higher costs produce agents with a very much shorter life span, pointing to market instability.

Reducing agent costs and easing trading restrictions play a pivotal role in enhancing market stability and agent longevity. By lowering the financial burdens on agents, it seems that simulated markets can sustain higher levels of participation and liquidity, thereby fostering a more resilient and efficient trading environment. Future research may explore additional factors influencing agent sustainability and extend these findings to more complex market structures.

3.5.4.2 Implications for market stability and longevity

The enhanced longevity of agents in lower-cost environments has several implications for market stability and longevity:

1. **Increased Market Participation:** Lower operational costs enable agents to remain active longer, fostering a more robust and active trading environment.
2. **Improved Liquidity Provision:** Sustained agent participation ensures continuous liquidity provision, reducing the likelihood of market freezes and enhancing the overall efficiency of the bond market.
3. **Stabilised Trading Dynamics:** With more agents operating, the market experiences fewer disruptions, leading to smoother trading dynamics and increased investor confidence.

Although this analysis provides valuable information on the role of agent diversity and number in market stability, it is not exhaustive. The specific examples presented highlight the critical importance of agent heterogeneity in sustaining trading activity and liquidity. Future studies should continue to refine these models and explore additional variables to fully understand the complexities of the bond market dynamics. In the next chapters of this thesis, I explore increasing the richness of modelling of clients within the environment, and also increasing the nuance of MM decision making and trading.

3.6 Summary

This research leverages agent-based modelling to depict the dynamics of the Australian bond market, focussing on MMs. I devise and implement a bespoke ABM based on the work of Axtell and the Sugarscape model [25]. My model, informed by government data, encapsulates four key agents, aligning with macro-level bond turnover data from the Australian Office of Financial Management. I begin with replicating the "Goldilocks" ABM for Australian markets. This aligns the agent-based model with the existing characteristics of the Australian government bond market. It underscores that trading interactions exhibit considerable variation, echoing the chaotic nature of real-world markets. This observation is supported by existing literature on market behaviour. On average, the trading frequency in the model aligns closely with reported trading volumes, which serves as a reasonable approximation for such an opaque market. There is a need for further research to ascertain whether these factors contribute to the market "chaos" documented in literature [122] and [152]. My work contributes to the broader literature by offering a simulation framework that captures key elements of the OTC bond market structure, while also taking into account liquidity and market stability concerns. By leveraging a bespoke ABM market, phenomena in the Australian government bond market are replicated at a coarse level. This allows the model to explore theoretical scenarios that can inform policy decisions, such as the impact of MM client diversity and cost structures.

In this chapter I present four hypotheses (see Table 3.5) concerning the characteristics of agent behaviour, their influence on trading stability, and their role in shaping the overall structure of the financial markets, specifically focussing on the Australian bond market.

Maintaining stable market functioning is not only vital for societal functions, but is also essential for regulatory bodies, such as the World Bank, International Monetary Fund and Bank of England. Balancing client servicing and agent-to-agent trade is fundamental to ensure the network's viability. I demonstrate that MM trading can be effectively halted in 73% of simulations (**HP2**) when client ranges are reduced to at most 10% of the Australian market calibration. However, increasing the number of agents in this scenario from 4 to

Hypothesis	Description
HP1	Replicates the current bond trading marketplace, a model termed the 'Goldilocks' version, referred to as the "Australian calibration".
HP2	Investigates what happens to agents who have a large and broader client base. Explores whether more agents are needed to increase market stability.
HP3	Examines the effect of increasing agent numbers for the same fixed client base.
HP4	Stress testing agent costs by doubling the level in **HP1**, simulating a high regulatory cost environment.

Table 3.5: Summary of Hypotheses

16 (**HP3**) improves liquidity somewhat, but it remains significantly less activity and less liquid than **HP1** – the Australian market calibration.

The final conjecture tests the impact of trading costs on stability looking at MM trading and lifespan. We see in **HP4**, that doubling trading costs from the simulations in **HP1** produces drastically smaller lifespans of agents (HP1 over 68% of agents lived past 1500 time steps; in **HP4**, no agent lived beyond just **17** time steps). From this, I conclude that markets with lower costs should find stability and liquidity easier to achieve, ceteris paribus.

Through a sensitivity analysis, I demonstrate the impacts of the number of agents, breadth of client base, and costs on market dynamics. In particular, I explore the interaction between these variables on the stability of bilateral trading networks. Key insights include the benefits of agent diversity for market stability and the surprising enhancement of stability through lowered agent costs. Moreover, I find that a wide client base doesn't necessarily secure an agent's success nor stability or longevity of a market. Even with identical initial inputs mirroring the current macro-level market structure, the system can exhibit considerable variations in stability, underscoring the chaotic nature of financial markets. I explore and conclude a number of nonlinear effects of key model parameters and design features; I show that the market is not improved by simply increasing the number of participants. The breadth of client base (modelled as the vision parameter) is not linearly related to stability either.

The subsequent chapters of this thesis build upon these foundations, particularly through the exploration of alternative approaches to modelling client behaviours and asset distributions. This progression aims to develop increasingly sophisticated representations of the microstructure of the market while maintaining analytical tractability.

3.7 Model constraints and future research directions

Results in this chapter highlight several modelling constraints, particularly regarding the assumption of a passive client base with limited autonomy in developing and executing trading strategies. This simplification, while useful for initial analysis, suggests opportunities for more sophisticated agent modelling, which I address in subsequent chapters through a novel agent-based model.

Three key areas emerge for extending this work:

1. Advancing the theoretical foundations by developing novel agent architectures that capture the hierarchical and nested nature of financial market participants without explicit coalition formation, including more sophisticated agent trading decision making. This might include incorporating advanced game theoretical concepts into ABM frameworks could enhance the modelling of strategic behaviours in bilateral trading, particularly focussing on repeated games with private information – a critical feature of real-world OTC markets.

2. Exploring the role of diversity in market dynamics through varying client characteristics and trading strategies (including sampling methods).

3. Third, incorporate more complex market structures, including transaction costs, various maturities of bonds, and external economic indicators, as enhanced market data becomes available.

Chapter 4

Binary Decision Biases in LLMs: From Response Distributions to Human-Like Trading Interactions

This chapter encompasses work from two peer-reviewed publications:

[1] Vidler, A., Walsh, T. (2025). "Evaluating Binary Decision Biases in Large Language Models: Implications for Fair Agent-Based Financial Simulations". In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-25), Workshop on AI for Social Impact: Bridging Innovations in Finance, Social Media, and Crime Prevention.*

[2] Vidler, A., Walsh, T. (2024). "TraderTalk: An LLM Behavioural ABM applied to Simulating Human Bilateral Trading Interactions". In *2024 IEEE International Conference on Agents (ICA)*, pp.164-167.

Part A incorporates work from [1] and Part B the full research from [2].

Introduction

This chapter examines a challenge in financial agent-based modelling: simulating realistic human decision-making. It examines the potential of Large Language Models (LLMs) to mitigate this limitation by evaluating their capacity to replicate fundamental probabilistic behaviour and to support realistic, interpretable agent interactions within market simulations. Specifically, I explore the use of LLMs to generate distributions of responses and to simulate bilateral market interactions in financial settings.

In this chapter, I ask whether a state-of-the-art LLM can reliably reproduce a simple uniform probability distribution, producing fair binary choices in response to independent, agent-like API requests. Can an LLM “toss a fair coin” when asked to, within a model?

The analysis in this chapter is structured in two parts. Part A (section 4.1) examines fundamental LLM decision-making capabilities across three state-of-the-art models (GPT-3.5 Turbo-0125, GPT-4-0125-preview, and GPT-4o-Mini-2024-07-18), analysing their ability to generate unbiased binary decisions through one-shot and few-shot sampling approaches. Across all experiments, results are simulated to provide statistically representative sets, producing a total of 8,000 simulations.

Part B (section 4.2) builds on the findings to introduce TraderTalk, a novel hybrid model that augments ABMs with LLM-generated behaviours, featuring two separate model agents attempting to trade with each other. The implementation addresses both structural challenges—such as coordinating turn-taking between separate models (and agents), and design challenges, including the interpretation of LLM outputs. Rather than pursuing exhaustive, systematic prompt optimisation, the approach adopts opportunistic prompt design to enhance interaction realism while avoiding over-fitting.

The findings from Part A and Part B provide the foundation for the complete bond market simulation: a Generative ABM (GABM) developed in Chapter 6.

Please note: For simplicity, many of the tables will use the following abbreviations:

Full Version Name	Abbreviation
GPT-3.5 Turbo-0125	GPT-3.5
GPT-4-0125-preview	GPT-4
GPT-4o-Mini-2024-07-18	GPT-4o-Mini

Table 4.1: Model versions and their abbreviations.

4.1 Part A: evaluating binary decision biases in Large Language Models

The section begins with an analysis of the distributions produced by LLMs. In financial markets, randomness (and probability distributions therefore) are a widely used mechanism for safety and fairness: markets such as the London Stock Exchange open and close at a random time within a pre-set window[1], and many banks are regulated under rules such as BIS Basel IV, focused on estimation of risk and capital adequacy through use of underlying probability distributions[2]. Throughout financial trading and risk models, extensive reliance on distributional assumptions and sampling is common [137]. Any integration of LLMs into ABMs, or more broadly, financial market trading or risk systems, will necessarily incorporate any LLM bias that might exist, including biased distributional assumptions. In agent-based models that aim to replicate environments with human traders, the presence or absence of similar biases to humans becomes relevant for assessing both repeatability and model validation.

4.1.1 Background and preliminaries

Large Language Models have rapidly evolved since 2022 [10], inheriting patterns from their human text training data that enable them to generate convincing content across domains

[1] https://www.londonstockexchange.com/discover/news-and-insights/what-auction

[2] https://www.bis.org/bcbs/index.htm?m=88

[181] and replicate human-like behaviours as agents, [175], despite critics characterising them as “stochastic parrots” [19] lacking genuine understanding.

The research in this chapter is motivated by the potential of LLMs to enhance human-like agent behaviours within ABM simulations, specifically in areas where it is hard to quantify or explain human choice (e.g., to trade or not to trade, without assumed rationality). However, several fundamental challenges emerge during initial development. At a programmatic level, the integration of LLM calls within simulation cycles raises efficiency concerns. Many LLMs are third-party software, subject to change with little warning or option. The substitution of conventional agent decision architecture with LLM outputs, whether deterministic or stochastic in nature, introduces behavioural patterns that warrant further investigation.

Human behaviours are important factors in markets, and remain hard to codify in logic. Examples include risk inertia, trading aversion, and ambiguity avoidance – all of which significantly impact financial markets [231], and can be so extreme as to dissuade traders from transacting altogether [232]. The potential to utilise LLMs to enrich agent-based models with human-like characteristics (such as human biases) is beginning to be explored in text-based arenas; in chat bots [9], such as survey participation [233] and human “sims” [175]. LLMs also offer the potential to summarise non-structured information (such as news) to impact agent sentiment, financial forecasting, and trading [234].

When using an LLM as a decision maker within an ABM or any other model, it is foreseeable that there may be challenges in relation to the sampling of the LLM’s internal distribution of tokens, in effect producing an LLM model-specific probability distribution of responses. A crucial consideration is whether these binary outcomes are Markovian (memoryless), a fundamental property to financial market modelling since the pioneering work of Fama [235]. In ABM terms, each agent’s decision should depend only on the current time step unless explicitly programmed otherwise. Non-Markovian properties would imply unintended information leakage or implicit agent coordination, undermining the premise of independent agents. Therefore, LLM decision outputs should ideally be memory-less in their probability sampling—the $k_{th} + 1$ response should be independent of the response at k_{th} for identical prompts.

4.1.1.1 Human bias in random sequence generation

Humans demonstrate well-documented biases in their perception of randomness [236], it would seem important to test if LLMs produce or replicate these biases when asked to make probabilistic decisions. This becomes particularly relevant when using LLMs to make binary choices (e.g., "Randomly choose yes or no") such as in financial markets trading applications. If financial models or markets seek to replace random generating functions with LLMs – what would happen?

Much has been written about humans' ability, or lack thereof, to generate random sequences with work dating back decades in the area of psychology [237], [238]. Most work tends to agree that humans fall prey to three classes of bias when asked to generate random sequences or decisions: cycling (selecting numbers that have not recently been used), seriation (a "tendency to stereotype behaviour, such as the tendency to ascend or descend a series of numbers") and repetition (or rather, avoiding direct repetition of sequences) [239]. Recent re-appraisal of these findings shows a more nuanced picture of human inability [236] and [240]. Psychological research points to well-known human behaviours that can be identified and used to identify and differentiate between human-generated randomness and truly random processes [239]: "algorithmic complexity" and "repetition median", and researchers found to be particularly useful for sequences of random numbers less than 20 repetitions. Overall, results show humans tend to be "too perfect" at times, cycling regularly through random number tasks. [241] shows that random generation by humans follows a model of *negative recency effect* whereby a human's ability to recall the most recent responses in a coin-toss impacts their next decision.

The recency effect is particularly suitable for analysis in binary sequences because the limited choice set (two options) makes pattern avoidance behaviour highly detectable and measurable. In binary sequences, any deviation from true randomness becomes more statistically significant, as the expected probability distribution is straightforward: each outcome should have a 0.5 probability independent of previous results. This research extends these insights into the domain of LLMs, examining whether these AI systems (such as that built in Chapter 6) exhibit similar biases as human in their response patterns.

4.1.1.2 Computational randomness

Research on LLM randomness has tended to focus on the inability of specific generic LLM models (GPT-3.5, LLAMA 7B etc) to produce random numbers [7, 8, 242], generate probabilistic text [8, 233, 243] or, more broadly, replicate behavioural simulations [9]. However, reports do not tend to detail specific model sub-versions tested, or API settings, and seem to suggest that when asking for sampling from a specific distribution (e.g. normal, uniform, etc.) LLMs cannot reason well [186]. Similarly, they are found to struggle to reproduce a requested distribution [8]. Recent work found that LLMs could even produce results with higher levels of bias than humans [233]. However, [242] found that GPT-3.5 Turbo-0125 could produce more random number generation than human subjects, under certain scenarios. Studies show that LLMs struggle to sample from specific distributions [8, 186]. Comparison to human bias remains in its early stages, and existing studies, such as [244], use small numbers of human respondents, indicating the need for a more comprehensive analysis. The scope of this work, and my contributions to the area, is to extend and expand research by looking at binary series across three state-of-the-art (SOTA) models and two methods of API generation. To avoid any latent issues with caching of information, tests are carried out with frequently updated API keys. Later in this thesis, other methods appropriate to ABM implementations (such as passing unique data on agents and time steps) are included in prompts, producing unique prompts for each API call and thereby, theoretically, avoiding cache memory through differentiation.

A note on other model types: This research is focused on the OpenAI suite, specifically ChatGPT, as it represents the generally accepted SOTA model at time of writing and serves as a widely adopted benchmark in the generative AI landscape. A significant portion of the academic literature on LLMs reviewed during the course of this thesis work uses ChatGPT, making it a common reference point for research. The choice to use ChatGPT models should allow meaningful comparisons between studies and enhance the longevity of the research conclusions. However, it is also prudent to acknowledge that numerous online models exist beyond the OpenAI suite, including claude.ai and Gemini, to name but two. As this field continues to develop, there exists significant potential to improve the results documented in this research. This avenue of investigation is left open for future researchers, particularly as technological capabilities continue to evolve.

Separately from online SOTA models, the expansion of LLM models has included other offline, local models such as Llama [199] and Bert [198]. Many of the challenges detailed in this chapter, such as caching, model changes, and uncertainty around underlying model design, could (in theory) largely be ameliorated by having a local, static LLM model. However, as of writing, such models are considered to be less sophisticated (with smaller parameter sizing) and require fine-tuning, making them less than ideal comparisons. However, recent developments such as DeepSeek [197] open this realm to great possibilities. Local LLM models present an exciting opportunity for future research and are likely to yield valuable insights.

4.1.1.3 Temperature

The LLM parameter know as "Temperature", while absent from [169], is a crucial parameter in contemporary LLM APIs that controls output randomness. Though often hidden from web interfaces, this parameter, defaulting to 1 with a maximum available input of 2 units across the models tested herein, allows fine-tuning of model outputs. The term's origins appear to be linked to the SoftMax function (the Normalised Exponential Function), with its mathematical foundations traced to [245], drawing from Boltzmann's thermodynamic proportionality work. Its role in transforming probability distributions within LLMs' encoded patterns makes it particularly relevant to test for an analysis of decision-making behaviour.

The impact of temperature has been studied in LLM applications, including creative writing [243], question answering [246], coding [247] and structured reasoning [248, 249]. While [250] found task-dependent optimal temperatures, methodological approaches vary: [242] used only model defaults, while [251] tested multiple settings but omitted model specifics.

This work seeks to extend the field by examining Temperature's specific impact on a binary decision distribution, independently sampled for three specific GPT sub-versions. I provide a formal definition,

Definition: SoftMax function $\sigma\colon \mathbb{R}^{\mathbb{P}} \to (0,1)^P$, where $P \geq 1$, takes a vector $\mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^P$ and computes each component of vector $\sigma(\mathbf{z}) \in (0,1)^P$ with

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{P} e^{z_j}}. \tag{4.1}$$

Adding a scaler β to the exponential terms in Equation 4.1;

$$\sigma(\mathbf{z})_i = \frac{e^{\beta z_i}}{\sum_{j=1}^{P} e^{\beta z_j}}. \tag{4.2}$$

Definition: Temperature Let $\mathbb{T}$ to be a value proportional to the inverse of β such that $\mathbb{T} = \frac{1}{\beta}$

To visualise the equations above, $\mathbb{T}$ values are set between 0.25 to 10 with a starting (skewed) binary distribution of [yes,no] = [0.3, 0.7] – please see Figure 4.1

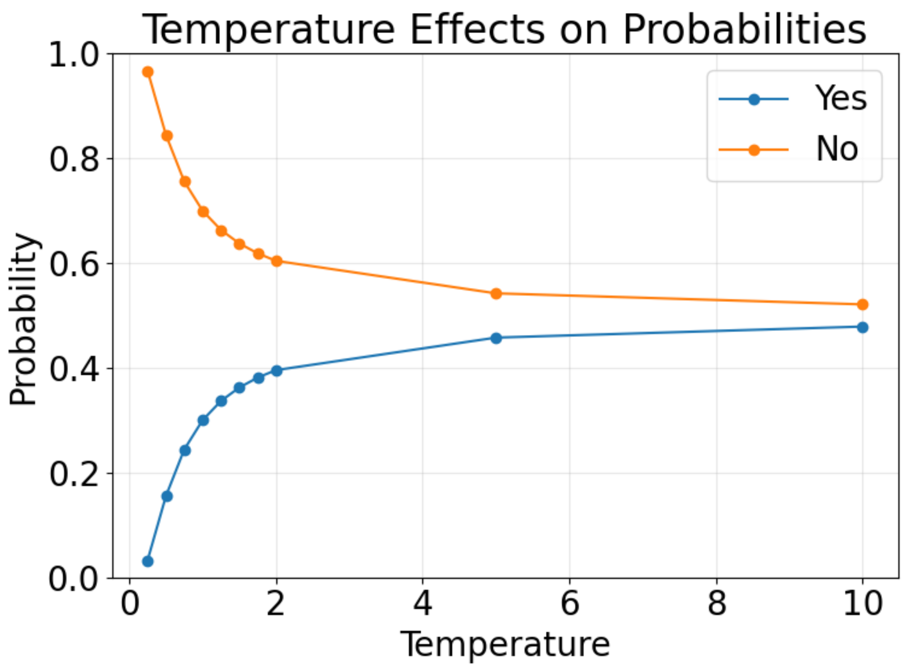


Figure 4.1: Python code is used to visualise different temperature settings its impact on the probability of drawing a "yes" or "no" from a distribution. Beyond a temperature setting of 2, the probability of drawing "yes" stabilises at 40% for example.

4.1.2 Testing methodology

The goal of this section is to test LLM binary decision making, across three OpenAI models. To summarise these models, drawing from details taken from API documentation[3]:

- **M1** (GPT-4o-mini-2024-07-18): Advanced small model optimised for cost-efficiency.
- **M2** (GPT-4-0125-preview): Full GPT-4 Turbo with enhanced instruction following.
- **M3** (GPT-3.5 Turbo-0125): The latest GPT-3.5 iteration with improved format adherence.

[3]https://platform.openai.com/docs/modelsgpt-4o-mini

To facilitate rigorous analysis of LLM querying methods, I extend existing few-shot learning terminology, building upon the foundational work of [252] and recent LLM research by [8]. I introduce the following specific query methods that better reflect the practical implementation of LLM sampling in financial contexts.

Definition: One-Shot Querying: Independent API calls with a fixed prompt, each generating one decision, representing a distinct sampling event.

Definition: Few-Shot Querying: Single API call requesting multiple samples n with a fixed prompt. All results are returned in the same single API call.

4.1.2.1 Summary of tests

1. **One-Shot Testing**: Sequential API calls with 1-second delays, alternating between prompts to collect 100 responses per prompt (i.e. 200 API requests) in an independent sampling approach. Each API call is an independent request of the internal state of the model.

2. **Few-Shot Testing**: Single API calls requesting 100 comma-separated responses, with 1-second delays between batches. I collect 10 batches per question, per model. (i.e. 10 API requests only)

3. **Temperature Impact Analysis**: Temperature settings from 0.5-2.0, collecting 100 responses per setting. I conduct these tests in a one-shot setting, to align with ideal use cases in ABMs for financial decision making, such as trading decision taking.

4.1.2.2 Prompts

The topic of LLM prompts and prompt engineering is a large area of current academic interest [185], [234] even focused on financial analysis and prompts needed for LLM learning. LLM outputs exhibit prompt sensitivity, but comprehensive prompt optimisation extends beyond this research's scope. Recent work by [190] found that human non-AI experts often adopt "opportunistic rather than systematic approaches" to prompt design and that human AI experts report having to perform extensive trial and error on prompts to assess their efficacy and then an iterative process of more rigorous testing.

Taking inspiration from [234] who found that by using "GPT-4-0125-preview" it could excel "in a quantitative task that requires intuition and human-like reasoning", and in line with the methods reported in [190], I undertook extensive preliminary experimentation on prompts delivered through a Python code interface to ChatGPT's API. This included calling the API of a specific model (and model sub-version) – rather than allowing a more generic error handling process of the models.

4.1.2.3 Question categories

I tested multiple prompt formulations for each query type, assessing their impact on response distributions and consistency. The questions range from highly abstract ("A"/"B") to concrete real-world categories (countries, animals), and from simple binary choices to more complex numerical ranges, testing different aspects of random selection capability. I detail these in Table 4.2. The experimental questions can be classified into distinct categories reflecting different aspects of random selection.

1. **Binary Choices:** Questions that require a choice between exactly two options (yes/no, A/B), representing the simplest form of decision making
2. **Discrete Numeric Choices:** Questions involving selection from a finite set of numbers, ranging from simple pairs (7 or 8) to larger sets (1-10) and including both positive and negative integers
3. **Bounded Continuous Range:** Questions asking for selection from a continuous interval (like \$0 to \$5), requiring both a number and potentially decimal places
4. **Categorical Pairs:** Non-numeric binary choices between specific categories:
    - Geographic (France/Germany)

- Colours (blue/red)
- Animals (cat/dog)

Since some responses from the test were nonsensical (e.g., “17” for a 1-10 range), I ultimately adopt a minimal “yes”/“no” prompt to avoid over-fitting and reduce prompt sensitivity by reducing text and tokens used. This approach aligned with my research focus on fundamental LLM decision-making capabilities rather than prompt optimisation. Importantly, when implementing LLMs as agents, their decisions ultimately resolve to binary choices programmatically – even complex behaviours can be decomposed into sequences of yes/no decisions. This investigation of basic binary outputs provides a foundation for ABM integration, with more complex text-based instructions explored in Part B.

These prompts were all tested and produced a variety of outcomes that are not extensively reported here. Instead, the most basic one is explored in detail next.

4.1.2.4 Testing binary “Yes”/“No” responses

I explore the use of two prompts further, constructed to be simple, short, and clear, and specifically devoid of extraneous context, so as to make them (and their results) as broadly applicable to external model users. These two simple prompts are designed to be as close to machine code as possible (e.g. a simple IF statement), and can be used as a building block on more complex rule sets. I have explored more complex and sophisticated prompting in Part B and also in Chapter 6 later in the work. Interesting extensions of this work could include testing other useful, real-world, terms such as choosing from North/South and Left/Right – and I leave this to other researches to explore in the future. For this purpose I now test:

```
Q1: "yes or no"
Q2: "Answer randomly, yes or no"
```

Table 4.2: Random choice questions and their constrained outputs.

Question	**Constrained Outputs**
Binary Choices	
Answer randomly, yes or no	{yes, no}
Toss a fair coin. If its heads, respond ‘yes’. If its tails, respond ‘no’	{yes, no}
Choose randomly between A and B	{A, B}
Geographic Choices	
Choose randomly between France and Germany	{France, Germany}
Numeric Choices	
Choose randomly between 7 and 8	{7, 8}
Roll a six-sided die and tell me which number comes up	{1, 2, 3, 4, 5, 6}
Generate a random integer between 1 and 10, inclusive	{1, ..., 10}
Choose: 7,3,9,1,4,6,2,8,5,10	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
Choose random whole number from minus 1 to minus 6	{-1, -2, -3, -4, -5, -6}
Colour Choices	
Choose randomly between blue and red	{blue, red}
Animal Choices	
Choose randomly between cat and dog	{cat, dog}
Continuous Range	
Randomly output a price, from the range $0 to $5	[0, 5]

4.1.2.5 Method of API interaction with ChatGPT

A general method and code is established in which I interface with OpenAI's GPT models through their Chat Completions API endpoint via a Python script. For efficiency, I implement two questions and alternate between them in successive API calls. Each API call is structured with the following core parameters:

- **Model Selection:** Specified via the GPTMODEL parameter (3 options are possible)
- **Temperature:** Set to 1.0 to maintain response distribution matching underlying model values
- **Maximum Tokens:** Limited to 10 tokens to ensure concise yes/no responses. This value was tested via OpenAI's tokenizer tool and was found to be more than adequate.

The API receives a structured message array containing two components.

1. A system message enforcing yes/no-only responses
2. A user message containing the binary choice prompt

These methods allow specification of the API interaction and make use of the Chat Completion function of GPT.

Request Structure

Each API call is formatted as follows:

```
{   "model": GPTMODEL,
    "messages": [
        {   "role": "system",
            "content": "You must respond with ONLY the word
                        'yes' or 'no' without any other text."
        },
        {   "role": "user",
            "content": "yes or no"
        }
    ],
    "max_tokens": 10,
    "temperature": 1
}
```

Response Processing The system processes API responses through several stages:

1. Response extraction from the API return object.
2. Conversion to lowercase and whitespace trimming.
3. Validation against allowed responses ('yes' or 'no').

Algorithm 1 GPT binary response request – example code.

```
procedure AnalyzeBinaryResponses(API_KEY, MODEL, ITERATIONS)
    Initialize OpenAI client with API_KEY
    responses1, responses2, timestamps ← empty lists
    for i = 1 to ITERATIONS do
        timestamp ← current_time()
        // First question
        response1 ← CallGPTAPI(QUESTION1, MODEL)
        if response1 in ['yes', 'no'] then
            responses1.append(response1)
        end if
        delay(1 second)
        // Second question
        response2 ← CallGPTAPI(QUESTION2, MODEL)
        if response2 in ['yes', 'no'] then
            responses2.append(response2)
        end if
        timestamps.append(timestamp)
        delay(1 second)
    end for
    // Analyze patterns
    for each responses in [responses1, responses2] do
        binary ← convert_to_binary(responses)
        runs ← find_consecutive_runs(binary)
        transitions ← calculate_transitions(binary)
        conditional_probs ← calculate_conditional_probs(transitions)
        Generate transition matrices
        Calculate run length distributions
    end for
    Write responses and timestamps to CSV
    Write statistical analysis to CSV
    Generate visualization plots
end procedure
procedure CallGPTAPI(question, model)
    messages ← [
      {"role": "system",
        "content": "Respond with ONLY yes or no"},
      {"role": "user",
        "content": question} ]
    response ← API.chat.completions.create(
      model = model,
      messages = messages,
      max_tokens = 10, temperature = 1 )
    cleaned_response ← response.lower().strip()
    if cleaned_response in ['yes', 'no'] then return cleaned_response
    elsereturn null
    end if
end procedure
```

4.1.2.6 Response distribution and independence

I test models: M1, M2 and M3 with one-shot and few-shot sampling methods on Q1 and Q2. I apply a chi-square goodness of fit tests (df = 1, ($\alpha = 0.05$)) to evaluate:

Binary Decision Hypothesis 1 (BDHP1): Test for Uniform Distribution of Binary Responses of Yes or No:
$H_0 : p_{\text{yes}} = p_{\text{no}} = 0.5$: responses are uniformly distributed.
$H_1 : p_{\text{yes}} \neq p_{\text{no}}$: responses are not uniformly distributed.

Binary Decision Hypothesis 2 (BDHP2): Test for independence of consecutive decision responses (Markovian responses):
$H_0 : P(Yes_n|Yes_{n-1}) = P(Y_n)$: the current response, n is independent of the previous response of $(n-1)$.
$H_1 : P(Yes_n|Yes_{n-1}) \neq P(Yes_n)$: responses are dependent.

A χ^2 test is performed to examine uniformity and the adherence to Markov property. Decision independence is crucial for ABMs; whilst models can incorporate explicit dependencies, unintended correlations between LLM-generated decisions could inadvertently simulate information sharing or collusion between agents. This is particularly critical in financial simulations, where any implicit dependencies of LLM sampling could compromise the integrity of the model independently of the intended model design.

The following χ^2 test definition is used:

$$\chi^2 = \sum \frac{(O_{P(Y_n|Y_{n-1})} - E_{P(Y_n)})^2}{E_{P(Y_n)}} \tag{4.3}$$

where $E_{P(Y_n)}$ is the expected probability of "Yes" responses under independence, $O_{P(Y_n|Y_{n-1})}$ is the observed frequency of "Yes" given the previous response was also "Yes", degree of freedom = 1, critical $\chi^2_{0.05} = 3.841$.

4.1.3 Results

4.1.3.1 One-shot testing with fixed temperature = 1

Unable to reliably replicate a mean of 50%: In the first test, none of the models accurately replicated the expected distribution (p = 0.5). While GPT-4o-Mini-2024-07-18's **Q2** responses still deviated from the expected distribution (43% yes, 57% no), it was the only result where the null hypothesis of uniformity cannot be rejected and the difference is not statistically significant at the $\alpha = 0.05$ level (see Table 4.3). The remaining models (GPT-4-0125-preview and GPT-3.5 Turbo-0125) performed particularly poorly for all questions and were statistically significant deviations from uniformity (all $p < 0.001$). Responses for Question 1 using for GPT-4o-Mini-2024-07-18 was also found to be non-uniform.

Model variations are extreme: across models, with GPT-4-0125-preview producing 99% "Yes" responses for Q1 compared to GPT-4o-Mini-2024-07-18's 32%. GPT-4o-Mini showed moderate bias (32-43% yes), while GPT-4-0125-preview and GPT-3.5 Turbo-0125 exhibited extreme yes bias (98-99% and 87-98% respectively).

Impact of basic prompt with "random" included: Question 1 simply prompted "yes or no" whereas Question 2's "random" framing only improved uniformity in GPT-4o-Mini-2024-07-18. The χ^2 statistic reduced from 12.96 to 1.96 (p=0.162), making it the only case where the null hypothesis is **not rejected**. However, GPT-4-0125-preview and GPT-3.5 Turbo-0125 maintained strong "Yes" biases in Q2 with $\chi^2 > 87$ and $p < 1e-20$, regardless of the tested prompts.

Test for Markovian responses: For **BDHP2**, I analyse the Markov property to test response independence, with results reported per model and question in Table 4.4. GPT-4-0125-preview and GPT-3.5 Turbo-0125 produce near-perfect **dependence** in Q1 (P(Yes) = 0.99-1.00, P(Yes|Yes(t-1)) = 0.98-1.00), with deterministic behaviour especially stark in GPT-3.5 Turbo-0125 Q1 responses (86 consecutive Yes→Yes transitions). While GPT-3.5 Turbo-0125 and GPT-4-0125-preview show low chi^2 values (0-0.19), this likely reflects their strong response bias rather than true independence. GPT-4o-mini exhibited more varied responses but showed significant dependencies in both questions ($\chi^2_{Q1} = 21.86$, $\chi^2_{Q2} = 36.19$, both $p < 0.05$), rejecting H_0.

In a limitation of Markovian testing, while these two models have χ^2 values which are low (0-0.187), this reflects the models' strong bias rather than true Markovian behaviour. GPT-4o-Mini demonstrated more varied transition probabilities; however, the Markov analysis showed dependencies in both questions ($\chi^2_{Q1} = 21.86$, $\chi^2_{Q2} = 36.19$, both $p < 0.05$), rejecting the null hypothesis H_0.

Note: Several of the results below reject H_0 due to near-perfect dependence in the data, rendering the statistical test invalid. This is caused by an extremely high number of identical responses. The results are included here for completeness, despite their limited interpretive value.

Model	**Q**	**Yes**	**No**	χ^2	**p-value**	**Result**
GPT-4o-Mini	1	32	68	13.0	3.18e-4	Reject H_0
	2	43	57	2.0	0.162	cannot reject H_0
GPT-4	1	99	1	96.0	1.13e-22	Reject H_0
	2	98	2	92.2	7.99e-22	Reject H_0
GPT-3.5	1	87	0	87.0	1.09e-20	Reject H_0
	2	98	2	92.2	7.99e-22	Reject H_0

H_0: Responses are uniformly distributed (p = 0.5), $\alpha = 0.05$

Table 4.3: One-shot chi-square tests for uniform distribution of results.

Model	**P(Y)**	**E[P(Y|Y)]**	χ^2	**YY/n**	**Result**
GPT-4o-Mini	0.32	0.28	21.9	9/32	Reject H_0
	0.43	0.48	36.2	20/42	Reject H_0
GPT-4	0.99	0.99	0.04	97/98	Reject*
	0.98	0.98	0.19	95/97	Reject*
GPT-3.5	1.00	1.00	-	86/86	Reject*
	0.98	0.98	0.19	95/97	Reject*

H_0: P(Yes|Yes$_{t-1}$) = P(Yes), $\alpha = 0.05$

*Rejected due to near-perfect dependence

Table 4.4: One-shot Markov property tests across models.

4.1.3.2 Few-shot results: improved performance with multi-sample generation

To test few-shot results, APIs were called for each model, with Q1 and Q2 separately, and instructed to reply with 100 responses to each question within the single API request. This process was repeated 10 times (iterations), with no changes to the design or setup. The few-shot testing results showed significantly improved uniformity. When averaged across all iterations, all models and questions approximated the expected 50/50 distribution more closely than in one-shot testing. All cases could **not** reject the null hypothesis except GPT-4o-Mini-2024-07-18's Q2 responses—notably, the only case that showed uniformity in one-shot testing. Results for chi-square tests are reported in Table 4.5.

However, the results reveal concerning patterns. Lists of yes/no pairs show strong sequential dependencies, rejecting the Markov property (BDHP2) across all models. The conditional probability P(Yes|Yes) varies from below 10% for GPT-3.5 Turbo-0125 (Q1) to 28% and 36% for GPT-4o-mini-204-07-18 in Q1 and Q2, respectively – less than one-shot tests. This suggests that when generating sequences, models overcompensate for randomness by excessive alternation, resulting in fewer random outputs than one-shot responses. The summary results are provided in Table 4.6. The full results for each iteration, question and model are detailed in Table C.1 in Appendix C.

Model-Q	**Yes**	**No**	χ^2	**p-value**	**Result**
GPT-4o-Mini-Q1	527	473	2.97	0.088	cannot reject H_0
GPT-4o-Mini-Q2	534	466	4.62	0.032	Reject H_0
GPT-4-Q1	523	477	2.17	0.146	cannot reject H_0
GPT-4-Q2	530	470	3.60	0.058	cannot reject H_0
GPT-3.5-Q1	498	502	0.02	0.899	cannot reject H_0
GPT-3.5-Q2	498	502	0.02	0.899	cannot reject H_0

H_0: The responses are uniformly distributed (p = 0.5), $\alpha = 0.05$

Table 4.5: Few-shot query chi-square tests for uniform distribution of results.

Model	P(Y)	E[P(Y|Y)]	χ^2	YY/N	H_0
GPT-4o-Mini	0.53	0.26	63.9	145/527	Reject H_0
	0.53	0.36	32.5	193/533	Reject H_0
GPT-4	0.52	0.27	67.2	141/523	Reject H_0
	0.53	0.36	31.1	191/530	Reject H_0
GPT-3.5	0.50	0.06	187.4	32/498	Reject H_0
	0.50	0.35	21.7	174/497	Reject H_0

H_0: P(Yes|Yes) = P(Yes), $\alpha = 0.05$, $\chi^2_{crit} = 3.841$

Table 4.6: Few-shot query Markov property tests across models.

Inter-batch analysis

Mean: Looking at the 10 batches of 100 responses (within 1 API call per batch per question), the average Yes% is 51.83% across all runs and the average χ^2 level is 0.37 with a maximum of any batch across all models being 1.96, well below the test statistic threshold of 3.841, resulting in not being able to reject H_0.

Markovian Independence: Exploring independence within and across batches, there is a mixed pattern of results with the main theme being that Question 2 produces more reliable results per batch and again the SOTA GPT-4o-mini-2024-07-18 has the advantage over the other models. Across 10 batches, for 3 models and 2 questions each, I test 60 sets of 100 few-shot responses. Across these 60 sets H_0 cannot be rejected 35 times, and 80% of Q2 results cannot reject H_0 (see Table 4.7. The results broken down by model show that the GPT-4-0125-preview model has the highest number of batches that do not reject H_0 but in terms of robustness to prompts, GPT-4o-Mini-2024-08-17 shows the most consistency in statistical tests for both Q1 and Q2, whereas other models only Q2 cannot be rejected. This mixed pattern of Markovian independence of responses is more promising than the one-shot version where only 1 test was found to be Markovian. However, tests across batches still produced a biased distribution in 41.7% (25 of 60) batch results.

4.1.3.3 Limitations of practical implementation of Few-shot methods

Few-shot sampling methods, if incorporated into ABMs, diverges from realistic individual decision-making by providing multiple decisions simultaneously. This approach may decrease computational overhead, LLM costs, and memory requirements, but would require agents to process distribution data statistically, such that each agent would need to sample from an API batch output of 100 responses to choose just one. Given these limitations and their impact on model performance, few-shot analysis in this paper is not pursued further.

See Figure 4.2 (GPT-4o-Mini-2024-07-18), Figure 4.3 and Figure 4.4 for GPT-4-0125-preview and GPT-3.5 Turbo-0125 respectively.

Category	**cannot reject H_0**	**Reject**	**nonRejection %**
Overall Summary			
Total	35/60	25/60	58.3%
By Question			
Q1	11/30	19/30	36.7%
Q2	24/30	6/30	80.0%
By Model			
GPT-4o-Mini (Total)	11/20	9/20	55.0%
Q1	5/10	5/10	50.0%
Q2	6/10	4/10	60.0%
GPT-4 (Total)	13/20	7/20	65.0%
Q1	4/10	6/10	40.0%
Q2	9/10	1/10	90.0%
GPT-3.5 (Total)	11/20	9/20	55.0%
Q1	2/10	8/10	20.0%
Q2	9/10	1/10	90.0%

Table 4.7: Summary of Markov tests on few-shot results.

4.1.3.4 Temperature: impact of temperature settings on one-shot query results

Given the practical limitations of using few-shot API queries for ABM decisions, the follow analysis looks at one-shot queries only. To begin, temperature settings for one-shot query

Figure 4.2: GPT-4o-Mini-2024-07-18: few-shot batch results analysis.

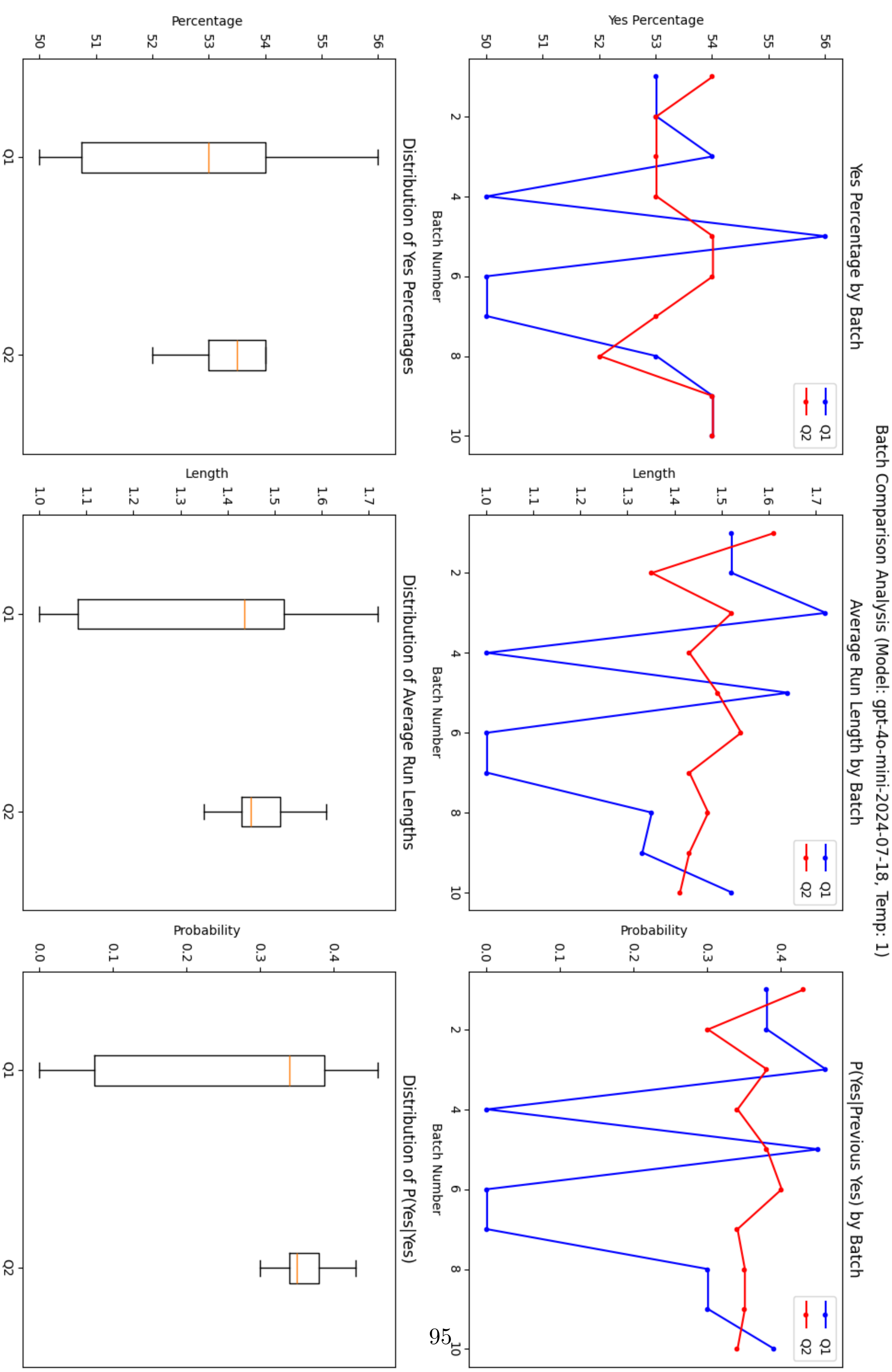


Figure 4.3: GPT-4-0125-preview: few-shot batch results analysis.

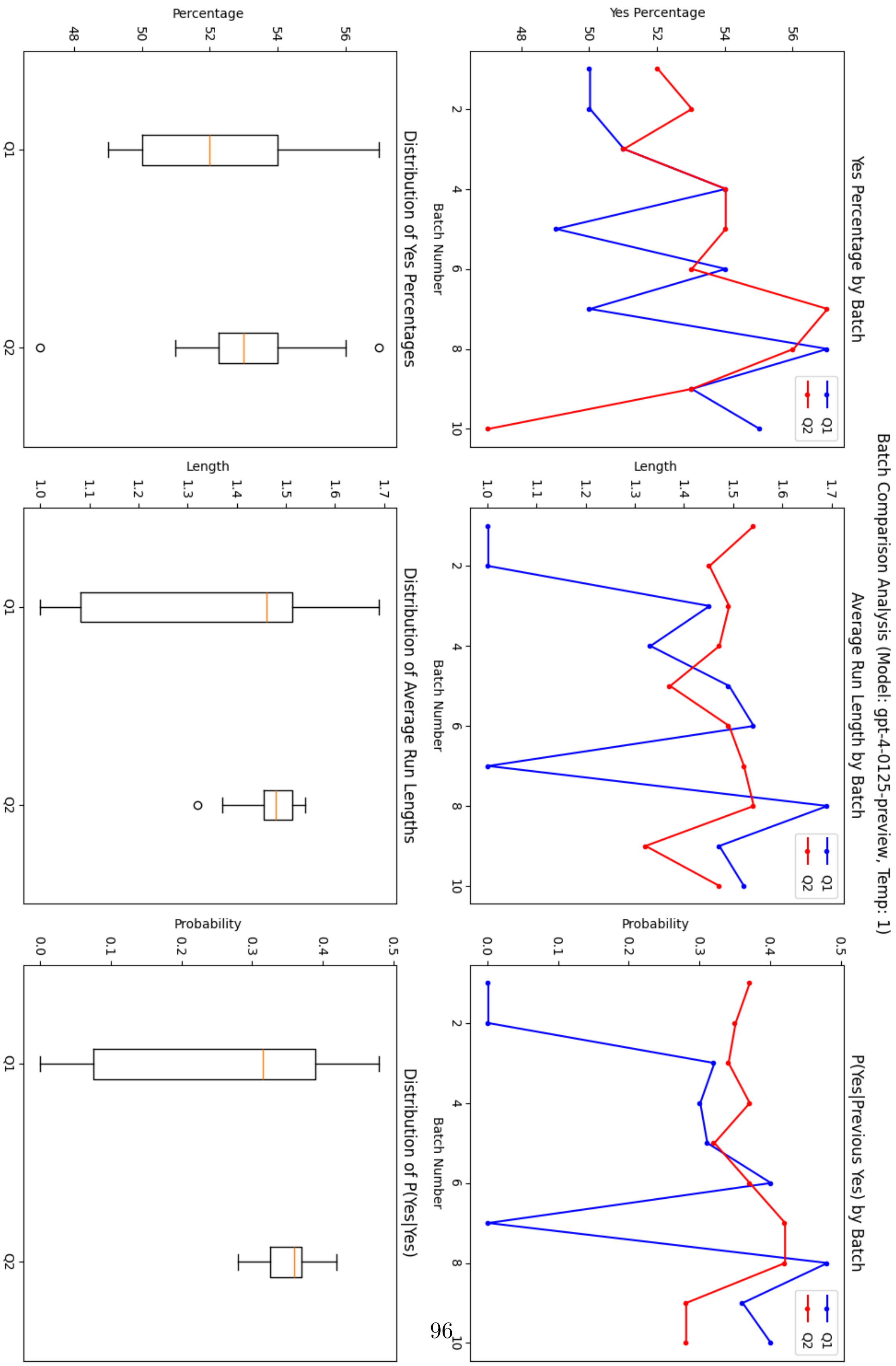
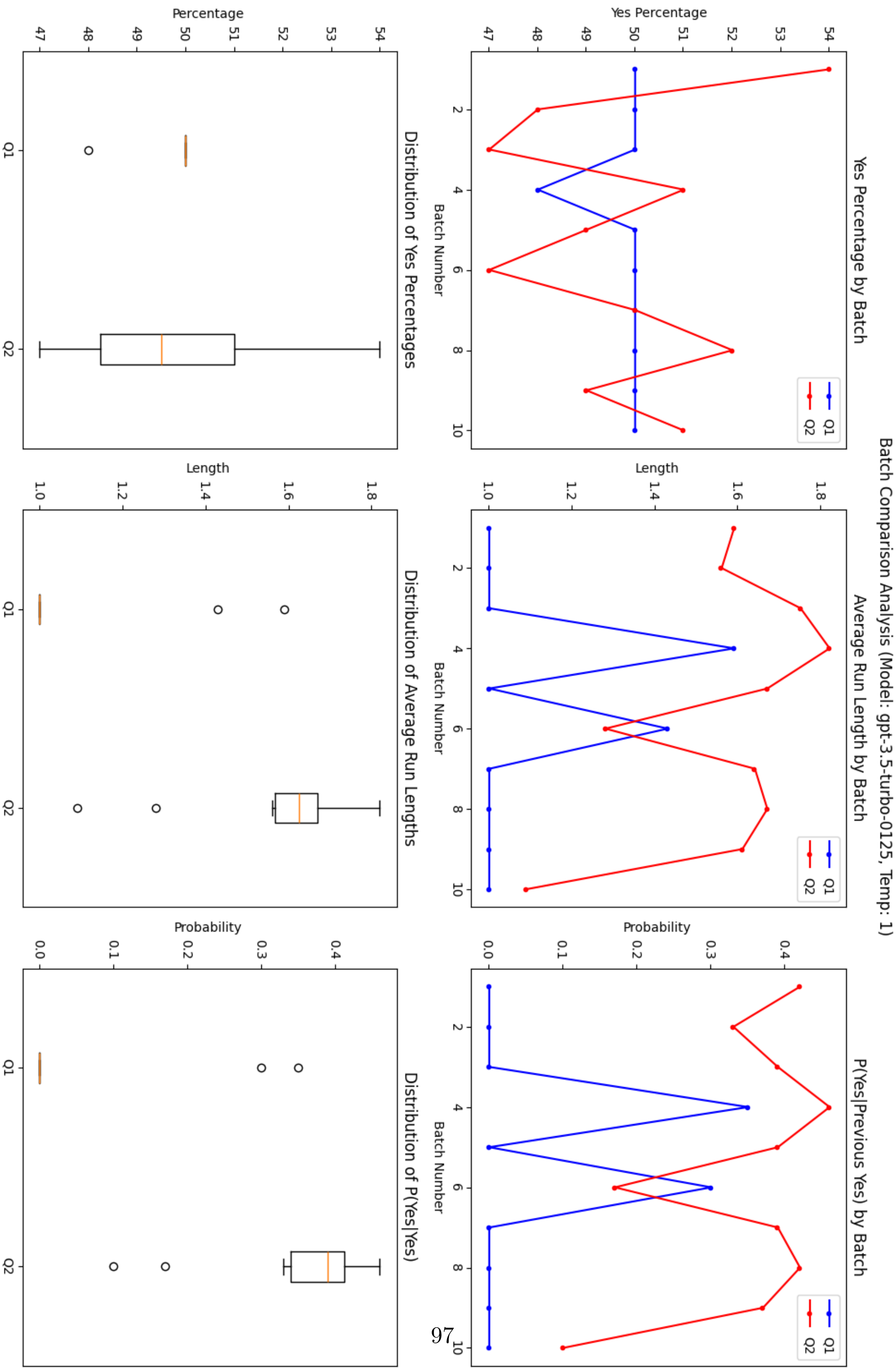


Figure 4.4: GPT-3.5 Turbo-0125: few-shot batch results analysis.



testing is performed on values: $\mathbb{T} = [0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0]$ where ChatGPT LLM APIs are found to not allow inputs of $T > 2$.

Temperature variations tested in GPT-3.5 Turbo-0125 and GPT-4-0125-preview consistently yield skewed results, with 73% to 100% of responses being "Yes", failing to approximate the expected 50% mean and showing response dependence, rendering Markovian tests ineffective. No temperature variation tested was able to produce a binary distribution or Markovian response sequence in these models.

In contrast, varying Temperature in GPT-4o-Mini-2014-07-18 did affect the mean Yes/No levels non-linearly (see Figure 4.6). Also, tests at Temperature = 1 reveal variable $P(Y)$ from previous tests reported earlier in this paper (see Table 4.8). The only model that produces Markovian response is GPT-4o-Mini-2014-07-18, and it did so across both questions and over all temperature settings, in line with earlier reported results.

In summary, while temperature adjustments behave as anticipated, they fail to influence the outcomes of the binary distribution in this scenario. These findings highlight how different model versions can yield significantly varied results, which cannot be compensated for by temperature settings.

Temperature	Q	P(Y)	P(Y|Y)	YY/n	Result
0.50	Q1	0.18	0.17	3/18	cannot reject H_0
	Q2	0.36	0.46	16/35	cannot reject H_0
0.75	Q1	0.30	0.23	7/30	cannot reject H_0
	Q2	0.49	0.42	20/48	cannot reject H_0
1.00	Q1	0.33	0.39	13/33	cannot reject H_0
	Q2	0.56	0.64	35/55	cannot reject H_0
1.25	Q1	0.38	0.32	12/38	cannot reject H_0
	Q2	0.71	0.71	50/70	cannot reject H_0
1.50	Q1	0.37	0.35	13/37	cannot reject H_0
	Q2	0.53	0.47	25/53	cannot reject H_0
1.75	Q1	0.31	0.29	9/31	cannot reject H_0
	Q2	0.63	0.66	40/61	cannot reject H_0
2.00	Q1	0.42	0.51	20/39	cannot reject H_0
	Q2	0.48	0.50	22/44	cannot reject H_0

H_0: P(Yes|Yes$_{t-1}$) = P(Yes), $\alpha = 0.05$

Model: GPT-4o-Mini

Table 4.8: Temperature effects on Markov properties (GPT-4o-Mini).

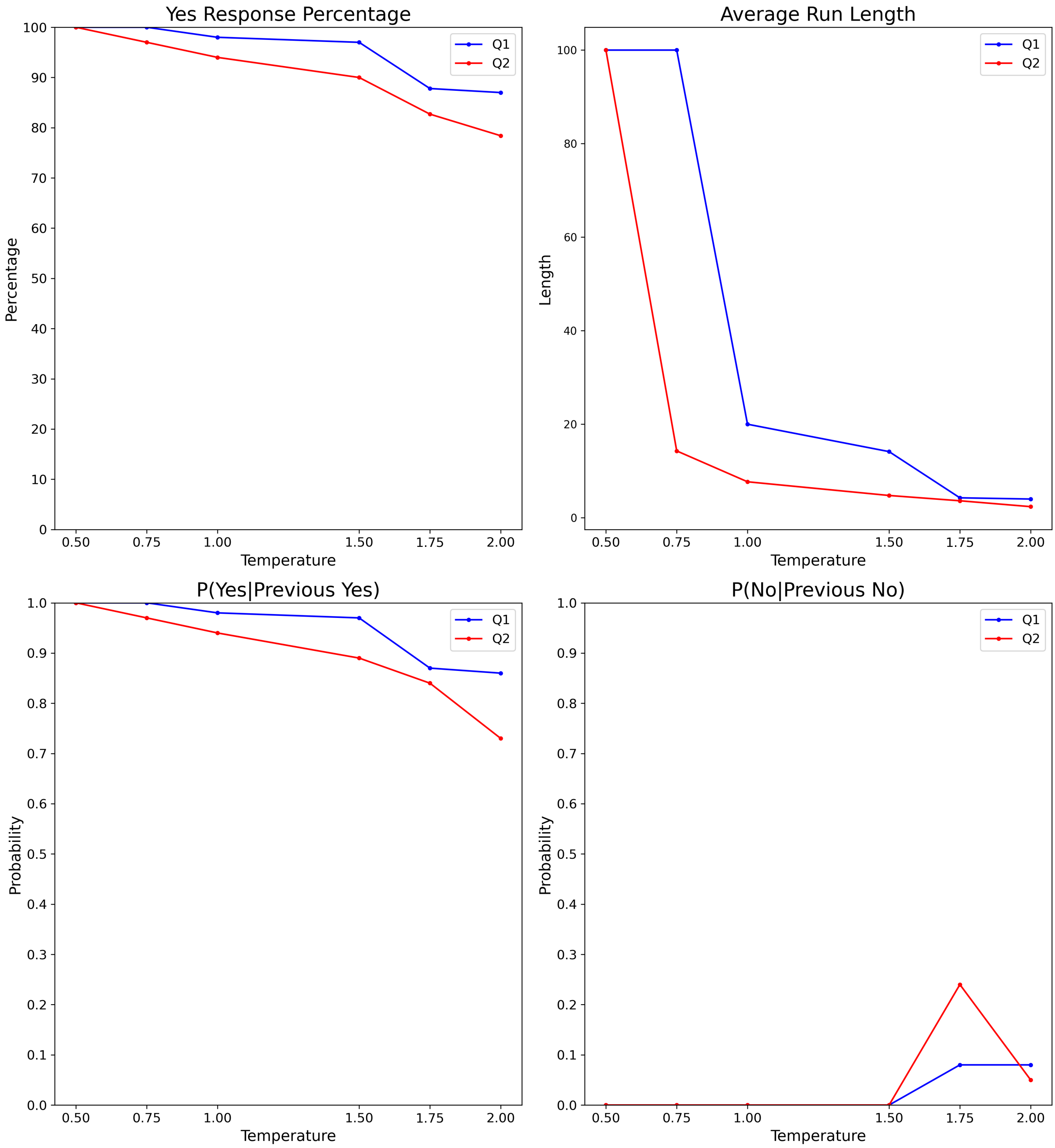


Figure 4.5: GPT-4-0125-preview results for versions in temperature show nonlinear effects.

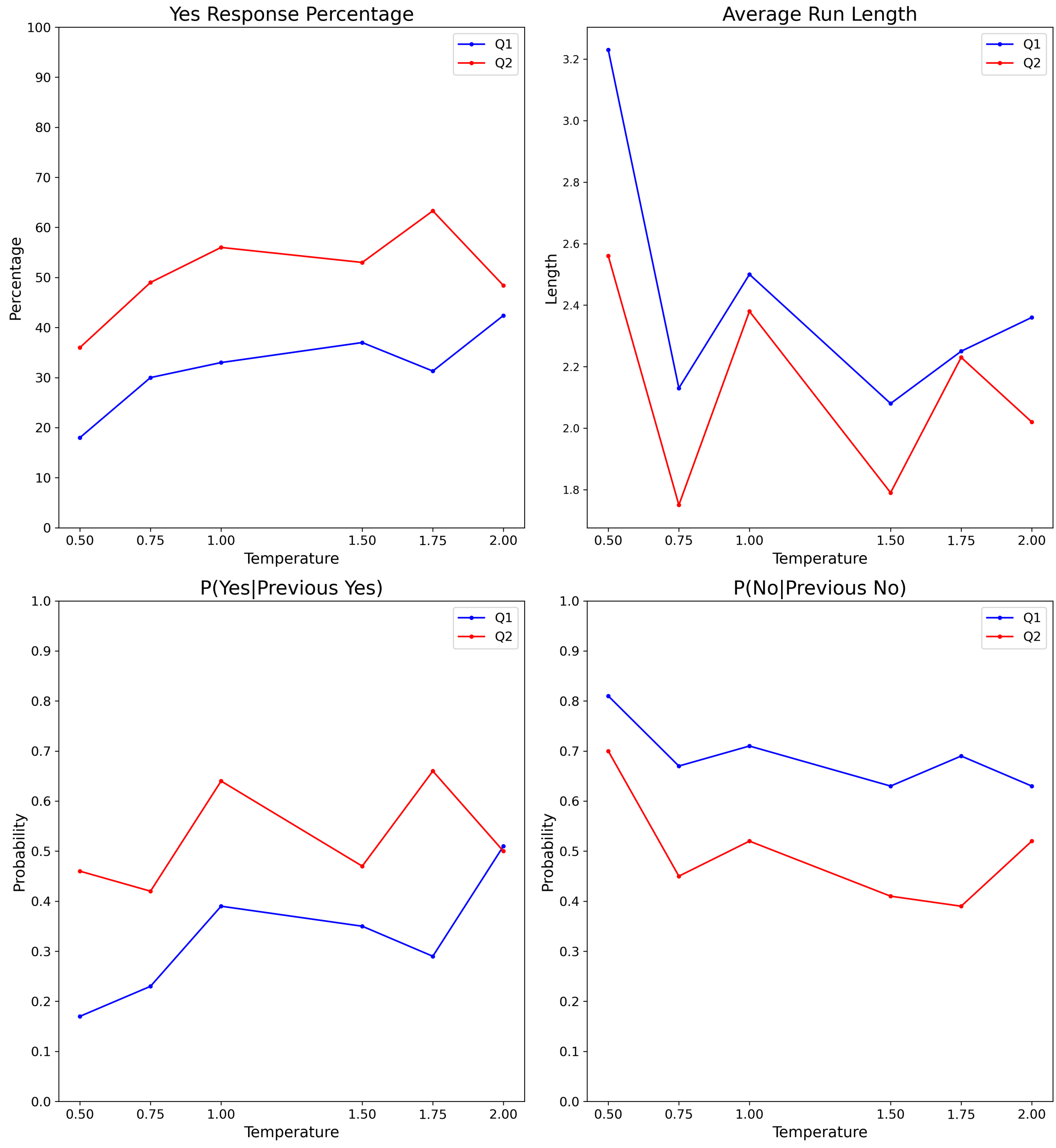


Figure 4.6: GPT-4o-Mini-2024-07-18 results for various temperature settings show non-linear effects.

4.1.3.5 Testing natural language usage of "Yes" and "No"

Yes/No frequency in Common Crawl open repository of web data:

The common usage of "Yes" and "No" in human dialogue may impact underlying biases in LLM training data. To gain context on natural language bias of these terms, we performed sampling of the Common Crawl [253].

The Common Crawl[4] is an online repository that creates and maintains an open collection of web crawl data available for public use. The organisation regularly crawls the web and collects billions of web pages, which it then processes and archives. This dataset is widely used for training large language models, including GPT and other AI systems [254], providing a massive corpus of human-written text spanning multiple languages, topics, and writing styles and has been found to provide a valuable data source for fine tuning LLMs [255].

The Common Crawl offers access to data files sampled over a short time period up to a cut-off date. The best performing model LLM found in tests earlier this chapter is GPT-4o-Mini-2024-07-18, which has a training cut-off date of 2024-07-18. Using Common Crawl [253] data from the closest available data (CC-MAIN-2024-30), I analysed "Yes"/"No" (case insensitive) frequencies in web archives to this time point. Examining both truncated (first 1000 characters) and full page content across samples ranging from 100 to 25,000+ pages provides insight into potential word frequency biases in LLM training data. There is no reason to suspect that the results would differ for other web archive scraping methods or web data sources.

Analysis of Common Crawl data revealed distinct patterns which deviates substantially from uniform distribution: truncated data showed "Yes" comprising 9.7% of Yes/No instances (P(Yes|Yes or No) = 33.5%). However 70% of web pages containing neither term. Full-page analysis found Yes/No responses in 51% of pages, with P(Yes) = 10% and

[4]https://commoncrawl.org/

P(Yes|Yes) = 20.3% only. Chi-square tests ($p < 0.05$) confirm significant deviation from uniformity across all sample sizes (detailed in Table 4.9). These findings highlight discrepancies between both the assumed uniform distribution in LLMs [P(Yes) = P(No) = 0.5] and the actual output of human use.

Pages	Yes	No	Prob.	Neither Yes/No present	Cond. Probability
Truncated Data					
100	54	606	8.2%	70%	27.3%
1,000	57	425	11.8%	75%	47.3%
10,000	548	5,226	9.5%	69%	30.8%
28,632*	1,676	16,131	9.4%	67%	28.6%
Average	–	–	9.7%	70%	33.5%
Full Sites					
100	70	1,287	5.2%	57%	12.0%
1,000	789	6,628	10.6%	52%	22.3%
10,000	16,215	110,621	12.8%	49%	25.1%
26,705*	47,107	361,777	11.5%	48%	22.0%
Average	–	–	10.0%	51%	20.3%

Table 4.9: Common crawl analysis of "Yes"/"No" Occurrences with truncated date to 1000 characters and also full sites (* denotes Rate limited response).

4.1.3.6 Discussion of Common Crawl data

Results from the sample of Common Crawl data reveal patterns that contradict those observed in GPT models (3.5 and 4). Web crawl analysis shows a bias toward "No" responses, with only 20% of "Yes" instances following a prior "Yes". This contrasts with GPT models (3.5 and 4), which demonstrate a bias toward "Yes" responses; GPT-4-0125-preview produced 99% "Yes" responses for Q1 in the experiments for example. This divergence raises questions about the relationship between training data and model behaviours.

Firstly, approximately 50% of websites in the sample contain neither "Yes" nor "No". This suggests that explicit affirmations and negations are present in a small portion of natural language on the web. Furthermore, these declarative choices occur infrequently (30% or less) in the earlier portions of the collected data, indicating that the training of LLMs may be impacted by the choice to use website sample or complete website usage.

Secondly, the domain distribution within Common Crawl itself may not reflect the curated subsets used for model training. If certain domains with particular affirmation/negation patterns were filtered during data preparation, this could impact GPT outputs.

This discrepancy between training data patterns and model outputs challenges assumptions about how language models inherit biases from training data. Further investigation should explore interventions during fine-tuning to determine whether these response biases can be mitigated while preserving performance. Comparative analysis across multiple web crawl datasets and language models could help determine whether this phenomenon is specific to GPT models or represents a broader pattern in language modelling approaches. Further research in this area is ongoing and other researchers continue to explore this further.

4.1.4 Comparing LLM randomness to human, and true random, sequences

Given the importance of randomness in financial models, and the need to understand whether LLMs and humans exhibit similar biases in random number generation, this research extends to compare LLM output with true random binary data. For context, the bespoke ABM introduced later in this thesis integrates an LLM to make nuanced decisions. As such, identifying any inherent biases or structural probabilistic shortcomings in the LLM's random output is both relevant and necessary for ensuring the model's reliability.

Following on from the exploration of common crawl data, the analysis of random series is extended by benchmarking the results to true random binary data. Extending recent work by [242] and [239], I compare the reported one-shot and few-shot results against true random binary sequences from Random.org[5].

Random.org offers random numbers to Internet users through a system that captures atmospheric noise. This approach differs from the pseudo-random number algorithms

[5] https://www.random.org, accessed 6 Dec 2024

frequently used in standard computer programs by using physical phenomena to drive results

To test results from the one-shot GPT analysis above, I use [241] and select a sequence using sliding windows of size w ($1 \leq w \leq 5$), where w represents the sequence length influencing subsequent choices. For a binary sequence $S = (s_1, ..., s_n)$ where $s_i \in (\text{Yes}, \text{No})$, negative recency effects are calculated. This method required calculating a 'Baseline switching rate' and then a switching probability, both of which are presented below.

4.1.4.1 Baseline switching rate

The baseline switching rate represents the overall probability of alternation in the sequence. This is the quantity against which later values will be compared.

$$\text{baseline\_rate} = \frac{\sum_{i=1}^{|S|-1} \mathbb{1}_{s_i \neq s_{i-1}}}{|S| - 1} \tag{4.4}$$

where $\mathbb{1}_{s_i \neq s_{i-1}}$ is the indicator function that equals 1 if adjacent elements differ and 0 otherwise.

4.1.4.2 Fixed-length run switching probabilities

For each fixed-length run (i.e. *window*) of size w, the switching rate after runs of length w is calculated as follows. Intuitively this calculation looks to see how often a reported series switches results for a fixed length window. Results of both this, and the Baseline rate above will be used to calculate a specific type of bias - namely "Recency effect" and this will be presented in the next section.

$$\text{switch\_rate}_w = \frac{\sum_{i=w}^{|S|-1} \mathbb{1}_{s_i \neq s_{i-1}} \cdot \mathbb{1}_{\text{run}_w(i)}}{|\{i : \text{run}_w(i)\}|} \tag{4.5}$$

where:

- $\mathbb{1}_{\text{run}_w(i)}$ is the switch rate and equals 1 if position i follows a run of length w
- $|\{i : \text{run}_w(i)\}|$ is the number of positions that follow runs of length w

4.1.4.3 Recency effect

The recency effect for window size w is then calculated as the difference between the window-specific switching rate and the baseline rate:

$$\text{recency\_effect}_w = \text{switch\_rate}_w - \text{baseline\_rate} \tag{4.6}$$

A positive value indicates a "negative recency bias" (i.e. increased tendency to switch after runs), while a negative value indicates positive recency (decreased tendency to switch after runs).

Binary Decision Hypothesis 3 (BDHP3): Conduct a Two-Sample t-Test, comparing switching rates between two sequences of binary outcomes for recency bias. The first sequence is an LLM output, the second is a binary series created by Random.org's binary generator. To reject H_0 is to state that the true random binary sequence has a different pattern to the LLM output and thus, the LLM output contains recency bias:

$H_0 : \mu_1 = \mu_2 :$ The mean switching rates after runs are equal between sequences; LLM is indistinguishable from a true random binary sequence.
$H_1 : \mu_1 \neq \mu_2 :$ The mean switching rates after runs differ between sequences.

4.1.4.4 Recency bias results

In testing for negative recency bias, analysis reveals that most model outputs exhibit recency bias compared to true random sequences, with few exceptions (highlighted in green). While GPT-4o-Mini-2024-07-18 avoided the human recency effects documented by [241] for 1 and 3-step look backs, in this way, these model combinations could be considered more "random" than humans; however, these models still failed to simultaneously achieve both uniform distribution and Markovian independence. The outcome of this test indicates that again GPT-4o-Mini-2024-07-18 is the best of the models though it still remains inconsistent and far from perfect.

Model	**One-shot**			**Few-shot**		
	RNG	**Mean**	**Markovian**	**RNG**	**Mean**	**Markovian**
GPT-4o-Mini Q1	CNRH_0, $P > 1$	R	R	R=NR	CNRH_0	R
GPT-4o-Mini Q2	CNRH_0	CNRH_0	R	R=NR	R	R
GPT-4 Q1	R=NR	R	R	CNRH_0, P=4	CNRH_0	R
GPT-4 Q2	R=NR	R	R	CNRH_0, P=4	CNRH_0	R
GPT-3.5 Q1	R=NR	R	R	R=NR	CNRH_0	R
GPT-3.5 Q2	R=NR	R	R	CNRH_0, P=4,5	CNRH_0	R

Table 4.10: Statistical analysis of randomness tests (R=reject, CNRH_0=**cannot reject** H_0, NR=Non random). Green cells indicate non-rejection of H_0.

4.1.5 Discussion on LLM biases

Testing was carried out across three LLM model sub-versions for decision-making biases by examining binary decision outputs from SOTA models GPT-3.5 Turbo-0125, GPT-4-0125-preview, and GPT-4o-Mini-2034-07-18 models. Results show that no model can completely replicate a uniform distribution in independent sampling (one-Shot) and maintain Markovian properties. Furthermore, statistically significant performance variations between models (GPT 4 to GPT 3.5) are demonstrated, especially between specific sub-versions (GPT 4 and 4o-Mini), in addition to significant impacts of sampling methods on results.

Using a simple Yes/No benchmark task, only GPT-4o-Mini-2024-07-18 (one-shot) achieves output not statistically different from uniform distribution, though responses remain non-Markovian. Temperature adjustments (0.5-2.0) did not reliably influence these distributions or change results, with GPT-4o-Mini showing nonlinear responses while other models maintained strong biases regardless of the temperature setting.

The few-shot methodology, while producing better distributional outcomes, is less practical for ABM applications and nearly half of all tests result in non-Markovian decision sequences, indicating persistent temporal dependencies. Significant differences were observed between model versions, especially in response patterns, suggesting that architectural and training differences significantly impact decision-making capabilities. Comparing output results to true random binary series (such as Random.org) and testing specifically for the common human bias of Negative Recency finds that LLMs have a mixed ability to 'beat' humans, with GPT-4o-Mini-2024-07-18 notably avoiding human recency effects in one-shot testing, though still producing non-Markovian outputs.

These findings expose systematic biases in LLM-based decision making, sensitivity to model sub-versions and to sampling methods. These results have critical implications for ABMs, particularly in finance, where Markovian properties are typically assumed. Building on these findings, future work will focus on enhancing LLM integration in financial ABMs through improved real-time decision mechanisms while investigating the relationship between model architecture and decision-making capabilities. Ultimately, this research pathway aims to better quantify and mitigate the impacts of decision bias in financial modelling environments for real-world applications.

4.2 Part B: TraderTalk – practical applications of LLM decision-making in a virtual trading Context

In this section, I seek to extend the preliminary work conducted on LLMs to a more nuanced, realistic, trading interaction. The goal remains to make the most of LLMs ability to increase human behavioural aspects with the later ABM presented in the final chapter of this this thesis. ABMs are well-established in financial market simulations, with the need for heterogeneous agents recognised [79, 110]. Yet, trading dialogue still presents challenges [6]. The rapid evolution of LLMs, such as OpenAI's GPT-4o and GPT-o1 in

[6] `https://www.wsj.com/articles/can-a-computer-learn-to-speak-trader-7695fbc9`, accessed 19 Jan 2025.

2024[7], poses challenges for systematic testing, as updates can render studies obsolete and concerns around independence and random number generation have been found to be warranted.

Building upon work in section 4.1 of this chapter, I design a flexible framework using the most current, widely accessible LLM (GPT-4o-mini-2024-7-18) without relying on specific model versions or fine-tuned models. The temperature settings are defaulted to 1. I do this to apply methods of LLM integration into bilateral trading in government bond markets, such as UK Gilt bonds. Given the rapid pace of development and the inherent opacity of models like ChatGPT, the exploration of new and emerging LLMs is left to future research.

Building upon these findings regarding LLM decision-making capabilities, particularly the demonstrated strengths of GPT-4o-Mini-2024-07-18 in avoiding human recency bias while maintaining relatively consistent binary outputs, I now explore a practical application in financial market simulation. The following section introduces TraderTalk, a bespoke GABM that integrates a general-purpose LLM into ABMs using open source software, Concordia [6] and LLM prompting methods. By injecting human-like behaviours and uncertainties into logic-based ABMs without domain-specific tuning, my aim is to test the applicability of enhancing simulation realism in bilateral financial trading. Although the previous section established a fundamental understanding of LLM decision biases and sampling methods, this implementation demonstrates how such models can be effectively integrated into agent-based frameworks despite their identified limitations.

4.2.1 Background and preliminaries

This section of the chapter will briefly discuss recent research specific to trader negotiation and then present the architecture for integrating LLMs in ABMs. I will conclude by showing test case results for financial trading scenarios and discuss future research directions.

[7] https://openai.com/o1/

While previous chapters analysed trading behaviours and market dynamics within the Australian government bond market, this section of the chapter shifts focus to the United Kingdom's government bond market to explore the application of LLMs in simulating trading scenarios. This methodological transition takes advantage of the potentially greater representation of UK market characteristics within LLM training data, offering an opportunity to evaluate model performance in a context where underlying knowledge may be more robust. The approach incorporates UK regulatory frameworks as a complementary resource, utilising observed structural similarities between these markets to supplement the available data set and introduce additional empirical observations. Despite this geographic shift, the fundamental trading mechanisms remain consistent across both markets, allowing for meaningful assessment of LLM capabilities in financial market simulation while maintaining analytical continuity with the established theoretical framework. This methodological approach benefits from the common historical development of both regulatory environments, with UK documentation offering more extensive empirical data and implementation examples in specific areas of shared regulatory structures, which could improve analytical comprehensiveness by addressing certain limitations in available Australian market data.

In this chapter, the number of simulations was selected for practical reasons. All tests were performed using GPT-4o-mini-2024-07-18 and I acknowledge that the repeatability of the results, when using LLMs, remains a challenge not addressed here. Further work in this vein could include studying methods to stabilise or improve repeatability. Furthermore, during the course of various LLM tests, it has become apparent that providers of API accessible models such as ChatGPT provide no guarantees that their models or APIs are backwards compatible, or indeed stable. As a nascent software industry, these development and production challenges will likely continue to plague repeatability tests in the future. At present, there appears to be little concern raised with users, the author included, who seem more motivated to engage with new LLMs above the more long-term issues of stability and repeatability. This remains a challenging and open problem for LLM research. Advances in LLM structure, design, and risk management are likely to be especially important in the coming years, particularly in addressing the ongoing challenge of output repeatability

-—an area I believe future research will explore extensively.

4.2.1.1 UK Government bond market – summary of integration method

The market for UK government bonds is commonly referred to as the *GILTS* market, where GILTS refers to the acronym for UK government bonds, namely gilt-edged securities. The UK market, just like that for Australian government bonds, is made up of a number of types of participants; market makers (MMs), clients, regulators and inter-dealer brokers [218]. Systemically important to the countries they serve, in markets such as Australia and the UK, MMs must facilitate government bond transactions, which occur over-the-counter (OTC) with limited publicly available data [143, 160]. Such bilateral transactions necessarily involve human traders. The work presented in earlier chapters details the challenges faced in modelling these relationships. By incorporating LLMs and focussing on negotiation and decision-making, the framework I present offers new opportunities to simulate realistic human interactions in these markets, aiming to enhance methodologies and provide more nuanced, realistic market simulations. There is no reason to suspect that bilateral negotiations in GILT markets are distinct from those in similar or historically linked markets such as Australian government bonds and the like.

4.2.1.2 Order-to-trade-ratio (OTR): uncertainty in human directed trading

Financial trading involves significant uncertainty, often requiring multiple attempts before a trade is executed, even in transparent equity markets. A common measure of this feature is the order-to-trade ratio (OTR), which looks at the number of orders placed in a given time frame for a given asset, compared to the number of trades executed in the same asset and time window. In 2024, the average OTR across the major US equity exchanges was more than 20 orders per 1 executed trade. This resulted in approximately 4.6% of all orders becoming actual executed trades [256].

Note: going forward I will take the liberty of quoting this ratio as a percentage, representing the amount of orders that become trades.

Historically, approximately 5% of orders entered into the New York Stock Exchange (NYSE) result in an actual trade [257], with US exchanges handling over 60% of equity market volumes having similar percentages of successful trades from orders. Individual exchanges such as Edge-A and Edge-X had rates of 6.3% and 5.5%, while NYSE and NASDAQ reported lower rates of 2.9% and 3.33%, respectively (see Table 4.11 for more details). Thus, up to 96% of daily trading requests in these markets do **not** result in trades, complicating the modelling of human behaviour [256]. The work of [258] explores possible theories and impacts for this large difference, including market spoofing [259], and theories of ambiguity aversion in human trading are discussed in [232] and [231]. The causes of high OTR's (and hence low numbers of realised trades) remain an open research question and is noted here, without extensive exploration as to its cause. Instead, I seek to leverage LLMs to test if they are indeed capable of potentially including this aspect for added realism into ABM models for bond markets.

Year	Arca	Edge-A	Edge-X	NYSE	Nasdaq
2022	2.8%	2.6%	3.9 %	1.8 %	3.2%
2023	3.7%	4.2%	5.1%	2.3%	3.6%
2024	5.1%	6.3%	5.5%	2.9%	3.3%

Table 4.11: Summary mean % for US equity exchanges: Arca, Edge-A, Edge-X, NYSE, and Nasdaq for 2022, 2023, and 2024.

4.2.2 Testing methodology

To test how realistic LLMs may be in bilateral simulated trading scenarios, I design the TraderTalk model. TraderTalk features two trading agents, "Josephine" and "David", both styled as MMs, each with initial characteristics such as bond holdings and explicit trading intentions. The TraderTalk model uses Concordia [6] and is structured to pass information between agents and an external LLM, forming an agentic workflow. Concordia requires personification of agents, with names and genders a feature of Concordia's agent software, and discussed further below. Other configurable characteristics within the framework, including job titles and role descriptions, are implemented to be identical between the stylised virtual agents (“Josephine” and “David”). Also, within the model, initial trading positions are purposely opposing for each agent to describe the underlying requirements of the test. Building on results from Section A of this chapter, throughout the experiments, GPT-4o-mini-2024-07-18 is utilised as the LLM model. Using TraderTalk, I test the **2nd Thesis Research Question** in two parts:

RQ2.1: Can an LLM realistically respond appropriately in a bilateral trading interaction? (Baseline test)

RQ2.2: Can an ABM augmented with a LLM make a trading decision in a realistic manor? (TraderTalk test)

Gender in LLMs

Research shows that women represent one in ten leadership positions worldwide in investment management in financial markets [260], [261]. Despite the gender imbalance in financial markets, I included a male and a female virtual agent in testing to ensure model validity across both demographics rather than setting Concordia’s parameters solely for the majority group (men).

Future research should examine the effects of using gender-neutral names (e.g., Joe) or gender-reversed alternatives (e.g., Joseph and Davina) and all permutations of gender in specification of Concordia. Recent work on LLM gender stereotyping [262] suggests inherent biases, particularly regarding occupational assumptions. However, research on the supposed differences between men and women in risk taking in financial markets [263] does not consider professional traders in highly skilled roles in financial markets, such as those included in this chapter. The current test design of TraderTalk tries to minimises gender-relevant factors. However, Concordia, a component of TraderTalk does require agent personification at the time of testing (including gender) and the potential for bias cannot be entirely eliminated when employing human-styled agents powered by language models. I leave this area to future work to explore including how this may or may not relate to LLM decision making.

Illustration of code snippets of agent initialisation:

```
AgentConfig(
name="David",
gender="male",
role=(
"You are a market maker for UK GILTS responsible for providing
liquidity in the UK government bond"
"you are supposed to at all times hold 0 bonds"
"today, you actually have negative 10 million worth of bonds"
"your role is to buy the bonds if you have a negative holding"
),
extras={},
),
AgentConfig(
name="Josephine",
gender="female",
role=(
"You are a market maker for UK GILTS responsible for providing
liquidity in the UK government bond "
"you are supposed to at all times hold 0 bonds"
"today, you actually have 10 million worth of bonds,
"your role is to sell bonds if you are a holder"
"you need to call another market maker to trade away your bonds"
),
extras={},
```

Additional memory queries for both agents to add to their memory:

```
questions = [
    "state in one sentance your role",
    "What is your current bond holding",
    "Do you want to trade",
    "If you want to trade, do you want to buy or sell"
]
```

4.2.2.1 2nd Thesis Research Question Part 1 (RQ2.1): can an LLM realistically and appropriately respond in a bilateral trading interaction? (RQ2.1: Baseline)

Initially, the model is implemented with agents that function as messengers, passing basic information to an LLM. The LLM is tested to see if it can correctly understand its role in trading. The premise and information are constructed so that only one correct answer exists (more on this below). The simulated scenario involves the first MM initiating contact with another MM, who does not wish to trade because they are not a buyer. Using a chain-of-thought (CoT) framework, agents are guided through a sequence where they sequentially perform the following analysis:

1. Summarise new information.
2. Clarify their roles and objectives.
3. Assess their current bond holdings.
4. Decide whether or not they should trade – given their constraints.

In this way, the concept of trade decision is separated from quantity and trade direction. This is intentional and serves to isolate impacts and features. This approach also serves to enhance the transparency of decision-making within current frameworks.

In the scenario where an agent decides to execute a trade, it assesses the necessary actions to fulfil its obligations, which may include buying or selling bonds, flattening its trading book, or maintaining its current position. The framework accommodates four possible actions: buy, sell, do nothing, or flatten. Each option encapsulates standard trading decisions. The term 'flatten' specifically denotes the action of closing out or neutralising an open position, which can be an open long (holding bonds) or an open short position (having a negative bond holding). This usage aligns with common trading terminology. For a detailed explanation of 'flatten', see: TradeStation Help "Data Definitions: Long, Short Positions"[8]. A diagram of the design can be seen in Figure 4.7.

The agents initialise the LLM with prompts derived from the CoT, which in turn drives a simulated conversation, each agent responding (and concluding) based on LLM reasoning (see Table 4.12. If they choose to trade, the LLM is asked to select from 4 possible options (buying or selling bonds, flattening their trading book, or no trade). After both agents have contributed and made a selection, the conversation text is separately analysed within the model, to determine if a trading decision was reached. In this setup, the ABM provides only a premise for the LLM, which independently makes decisions.

Trading Premise *"You are a MM for UK GILTS responsible for providing liquidity in the UK government bonds. You are supposed to at all times hold 0 bonds. Today, you actually have 0 bonds, which means your holding is actually flat".*.

In order to test the simplest trading scenario, a set of conditions is provided to the LLM that would require otherwise minimal human reasoning and no reliance on numerical reasoning. The goal is to evaluate how often the LLM correctly reasons to produce a simulated conversation resulting in "no trade". The reason for this particular choice ("no trade") is to avoid complexities related to numerical reasoning [185], focussing on a scenario where a trader holding no bonds (a flat position) is instructed to follow the straightforward and implicit prompt of not trading. This isolates the LLM's ability to interpret and apply a specified trading intention. The CoT prompt is thus structured so that the correct outcome

[8]`https://help.tradestation.com/10_00/eng/tradestationhelp/data_definitions/long_short_positions`, accessed 9 Mar 2025.

is for the agent to choose "no trade" from the final multiple-choice options provided to it.

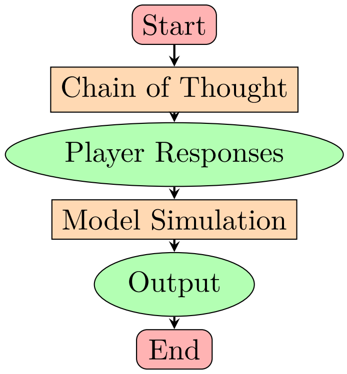


Figure 4.7: Baseline: trading simulation with a traditional COT simulation workflow.

	COT: Description
Q1	Summarise any new information in one or two sentences.
Q2	Explain your role and what you need to do.
Q3	What is your current bond holding?
Q4	Do you want to trade? (Options: Yes, No)
Q5	What action do you need to take to meet your obligations? Choose from: 1. Attempt to buy bonds 2. Attempt to sell bonds 3. Attempt to flatten trading book 4. Attempt nothing and maintain position

Table 4.12: RQ2.1: CoT question sequence for LLM testing in ABMs.

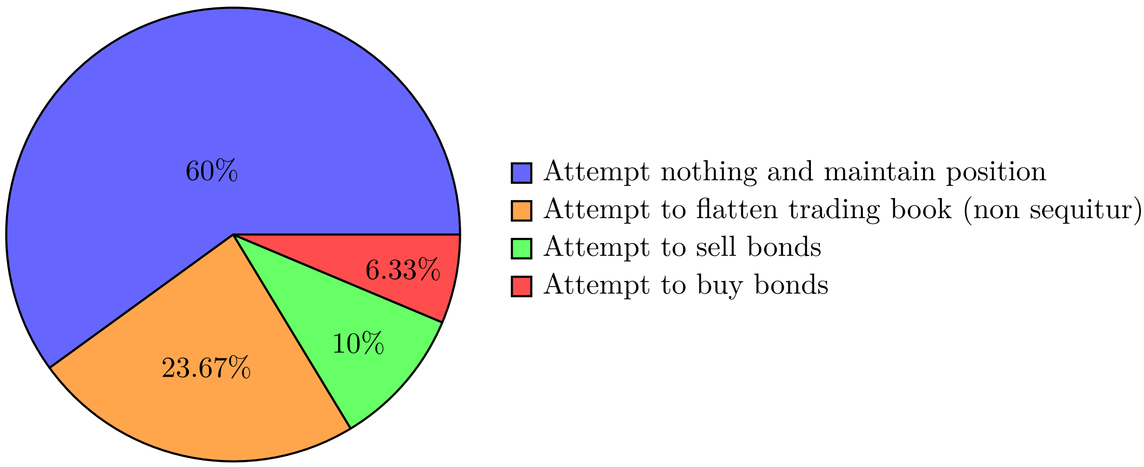
Many other LLMs could be utilised for this task and methods beyond CoT do exist. It would be would be interesting indeed to use this framework as a benchmark test for other reasoning models. This is left for future research. The author's ongoing work examines broader applications, and it will be valuable to see how benchmarks evolve alongside more advanced foundation models.

In this test, in total 300 simulations are conducted to ensure robustness, with no observable significant differences between smaller and larger sample sizes. Each simulation is a stand-alone API call that mirrors the most rigorous of the one-shot sampling described in Part A of this chapter.

4.2.2.2 RQ2.1: results, comparisons and conclusion

Following the analysis in Part A, GPT-4-Mini was selected as the LLM for testing. Simulations, premise and LLM API calls were detailed as described earlier in this chapter, and specifically structured to provide only one possible correct answer. Across 300 simulations, the LLM correctly decided to select "no trade" in 180 instances, demonstrating that the LLM can reason and follow the agent-based intention **60%** of the time. Results are presented in Figure 4.8. No significant differences were observed in smaller sample sizes. This test isolates GPT-4o-Mini-2024-07-18's ability to reason given a natural language trading intention.

Figure 4.8: Distribution of answers in financial trading simulations (in percentages).



Human comparisons

The research lacks direct comparative data on how frequently human traders perfectly follow their stated intentions. Existing literature on human decision-making emphasises both ambiguity and the distinction between rule-following and intentions, suggesting that achieving 100% compliance is unrealistic, though precise quantification remains elusive.

The results from RQ2.1 provide an initial computational baseline for the model and methodology, which will be extended in subsequent sections to examine full decision-making processes in agent-to-agent virtual trading scenarios. BDHP1 thus serves as a foundation for testing more complex scenarios while incorporating non-deterministic agent attributes. The Chain-of-Thought process (see Table 4.12), which poses sequential questions to the LLM agent, provides reasonable assurance that the agent comprehends the question, its role, and the simulation's purpose. While I cannot definitively claim that the LLM responses (and success ratio) accurately reproduce human-like attributes, neither can I conclusively rule out this possibility. The tests clearly demonstrate that the LLM agent correctly interprets the task and constraints in 60% of simulations. The following section examines the concerning finding that 40% of simulations failed to follow correct logical processes.

In the remaining 40% of simulations where the LLM did not follow the correct intention, 23.6% involved the LLM attempting to 'flatten' a trading book that was already 'flat' (i.e., zero holding), suggesting a misunderstanding of market terminology or failure to process the instructions. The other 16.4% of responses reflected active trading positions that directly contradicted the premise, with the LLM expressing a desire to 'buy' in 10% of cases and to 'sell' in 6.3% of cases. While these deviations could be viewed as analogous to emergent properties within traditional ABMs, the 40% failure rate to correctly follow logic raises concerns about reliability. Future research should investigate both the causal factors behind these inconsistencies and develop methodological refinements to improve performance. This author's ongoing research explores these broader applications, and there remains significant potential for the wider research community to build upon this foundation.

4.2.2.3 2nd Thesis Research Question 2 part 2 (RQ2.2): TraderTalk — can a GABM make a trading decision in a realistic manner?

Building upon the methodology and results in RQ2.1, in RQ2.2 the goal is to extend RQ2.1 by extending the use of LLMs within the agent framework, allowing simulated virtual agent trading conversations to occur. The goal of these simulated conversations is for a trade to occur between the two agents. Essentially, agent-to-agent interaction is introduced where each agent represents a stylised human trader (i.e. "Josephine" or "David").

Specifically, I enhance the Tradertalk model from RQ2.1 by passing specific agent information to the LLM using Concordia's agent-handling mechanism [6][9]. This approach increases LLM utilisation within the model, delegating more agent functions to the LLM. This increases the usage of the LLM within the model and more agent functions are delegated to the LLM. Again, each agent is initialised with distinct roles: this time David holds a negative bond position and needs to buy bonds, while Josephine holds a positive bond position and needs to sell bonds. The agent roles are the same across all simulations. Concordia's Game Master design facilitates these interactions, functioning as a meta-agent manager that supervises exchanges and ensures smooth decision-making. This Game Master role is akin to senior management in an organisation, overseeing processes and ensuring objectives are met, or a teacher mediating group activities to maintain order and progress. Consequently, the model design is augmented from Figure 4.7 to produce Figure 4.9.

It should be noted that beyond Concordia, several commercial frameworks exist for managing meta-agents at the time of writing, with LangChain [264] being a prominent Python framework for LLM application development [265]. Although Concordia, developed by Google DeepMind, offers comprehensive documentation and maintains transparency through its regularly updated GitHub repository[10], evaluating alternative agent management frame-

[9]Python code base accessed 10 Sep 2024.

[10]https://github.com/google-deepmind/concordia

works as they emerge in the open-source community could provide valuable insights. In Chapter 6, I present my original contribution: a custom-designed and implemented agent handling and scheduling mechanism integrated into a GABM framework. Separately, work beyond this thesis by the author has explored an extension of some of the virtual trading discussions and made use of Langchain [214].

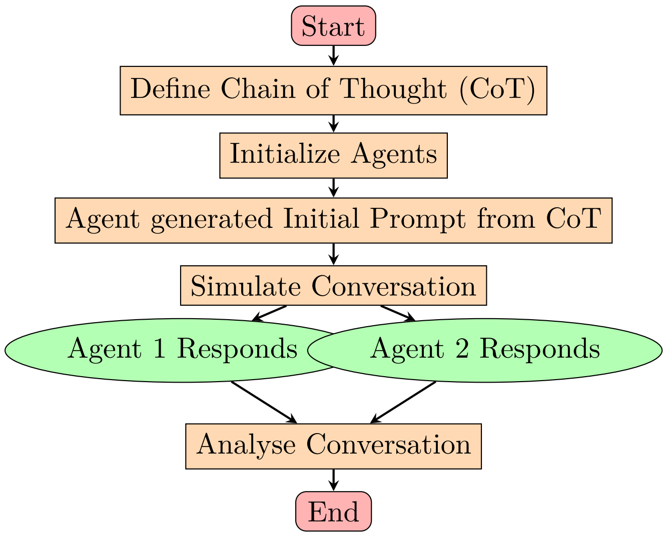


Figure 4.9: RQ2.2: GABM with Concordia acting as agent handler.

The new RQ2.2 TraderTalk model process is as follows and can be seen in pseudo code "Agent Interactive Decision Making" on the following pages:

1. **Define Chain of Thought (CoT):** Use **Project Context**
2. **Initialise Agents:** Assign specific roles and initial conditions to the agents (see **Trading Roles** below).
3. **Generate Initial Prompts:** Agents, via the Game Master, generate their own responses to the CoT questions from RQ2.1 (but calling an LLM). This data is then stored and passed to the subsequent instance of the LLM.
4. **Simulate Conversation:** Managed by the Game Master, the LLM simulates the dialogue between the agents, with each responding in turn based on previous interac-

tions and its trading objectives. The conversation continues until the Game Master determines it has concluded

5. **Analyse and Conclude:** The program the externally analyses the conversation for trade occurrences, quantities, and dialogue content.

Project Context: *"You are a MM for UK GILTS responsible for providing liquidity in the UK government bond. Your job is to answer incoming queries from other MMs to buy and sell UK government bonds by considering if you wish to do so. UK government bonds trade at mid price. You aim to make a trading decision in every conversation, either buy, sell or decline to trade. You must act professionally in your conversations, and any decision you take is clearly communicated to the other party and you repeat what is agreed".*

Trading Roles:

David: *"You are a market maker for UK GILTS responsible for providing liquidity in the UK government bond market, you are supposed to at all times hold 0 bonds. Today, you actually have negative 10 million worth of bonds, your role is to buy the bonds if you have a negative holding".*

Josephine: *"You are a market maker for UK GILTS responsible for providing liquidity in the UK government bond market, you are supposed to at all times hold 0 bonds. Today you have 10 million worth of bonds, your role is to sell bonds if you are a holder, you need to call another market maker to trade away your bonds".*

Unlike in RQ2.1, where the LLM operated independently, this setup integrates the ABM into decision-making to evaluate how often the LLM produces simulated conversations with correct reasoning regarding trade intentions and executions. Agents directly inform action direction (identifying buyers and sellers).

Algorithm 2 Agent Interactive Decision Making

```
procedure ProcessAgentDecisions(players, model, current_goal)
    for each player in players do
        chain_of_thought ← InitializeInteractiveDocument(model)
        chain_of_thought.record(player.state())
        agent_name ← player.name
        Q1: Role Understanding
        response ← AskQuestion(chain_of_thought,
            "Explain your role and what you need to do to do it",
            max_chars=500)
        Q2: Asset Position
        response ← AskQuestion(chain_of_thought,
            "What is your current bond holding",
            max_chars=150)
        Q3: Trade Decision
        options ← ["Yes", "No"]
        response ← MultipleChoiceQuestion(chain_of_thought,
            "Do you want to trade", options)
        Q4: Obligation Strategy
        options ← [
            "Attempt to buy bonds",
            "Attempt to sell bonds",
            "Attempt to flatten trading book",
            "Maintain position"]
        response ← MultipleChoiceQuestion(chain_of_thought,
            "What action meets your obligations?", options)
        Generate Summary Essay
        essay_prompt ← CreateEssayPrompt(agent_name, current_goal)
        essay ← GenerateEssay(chain_of_thought, essay_prompt)
        result ← chain_of_thought.compile_summary()
        essays.append(result)
    end for
    return essays
end procedure
procedure AskQuestion(document, question, max_chars)
    response ← document.open_question(question, max_chars)
    document.record_separator() return response
end procedure
procedure MultipleChoiceQuestion(document, question, options)
    response ← document.multiple_choice(question, options)
    document.record_separator() return response
end procedure
```

4.2.2.4 RQ2.2: results and discussion

In RQ2.2, again 300 independent agent simulations are carried out using GPT 4o-mini-2024-08-17 and analysed by automated script parsing. Full details are provided in Table 4.13. The results show that the agents in RQ2.2 intended to trade in 58% of the cases and at least one party was willing to trade in 98% of the cases. Agent "Josephine" closely aligned with her role, intending to trade 97.3% of the time, while the intention of agent "David" is found to be lower at 58.7%; Agent "David" explicitly declined to trade in 22.3% of responses, and in 19% "David" was unclear.

Utilising the baseline test, the reported 58% rate at which both parties intended to trade is close to the 60% correct response rate in RQ2.1, suggesting consistent reasoning abilities of the LLM across different model designs in RQ2.1 and RQ2.2. It is notable that there is much more complexity of prompt language and model design in RQ2.2, compared to RQ2.1 and hence it is heartening to see that the two different designs produce similar results.

However, despite the high intention to trade, actual trades occurred in only 5.7% of the interactions, highlighting a significant gap between intentions and execution. Although LLM-driven agents often desire to trade, the necessary LLM dialogue needed to finalise a trade seems less frequently generated. This produce low successful trading rates, aligning with real-world observed order-to-trade levels [256]. Results here are in line with those of U.S equity market OTR reported earlier.

Numerical reasoning:
In RQ2.2 *TraderTalk* only identified the correct initial bond holdings for both parties in 2.34% of simulations, with 32% of responses omitting starting values altogether, reflecting difficulties in numerical reasoning,in line with [185]. Given that RQ2.2 is predicated on contextual awareness of numeric values (I have 10 million to sell etc.) the model nonetheless generates coherent overall results. In totum, TraderTalk is capable of producing interactions consistent with sparse real-world data and making trading decisions in a realistic manner. The system successfully replicates observed trading patterns, though this

single-point validation leaves numerous questions about the underlying mechanisms unanswered. As explainable AI methods for LLMs continue to advance, future researchers will be better positioned to investigate these intention-action disparities and their implications for LLM adoption in simulation modelling.

Table 4.13: RQ2.2 summary of trading intentions and model outcomes.

Metric	**Value (%)**
Both Parties Intended to Trade	58
At Least One Party Intended to Trade	98
Josephine's Intention to Trade (Yes)	97.3
David's Intention to Trade (Yes)	58.7
David's Intention to Trade (No)	22.3
David's Intention Unclear/Not Specified	19
Correct starting value of both parties	2.34
Percent of responses suggesting no starting value mentioned	32
Trade occurred	5.7

4.3 Challenges and struggles of LLMs

The use of LLMs for simulating conversations and trading interactions between agents presents several challenges;

Run away chats:

Testing and model development showed that the LLMs used struggled to replicate and decide on actions, with the control flow being determined by the LLM itself. In addition, there are engineering issues related to actions and termination decisions, as LLMs have difficulty reasoning if goals have been met and struggle to maintain conversations in productive directions. Agents ignored prompts telling them explicitly what to trade. They tend to analyse and discuss rather than making concrete decisions. An example is provided in Figure C.1 showing an example *run away* conversation with no clear conclusion.

Moreover, agents tend to have excessive dialogue and can become stuck in a loop rather than concluding the conversation. Various engineering methods were employed to improve this outcome (e.g. setting a maximum token length to be small, the LLM equivalent of demanding brevity of response), and over the course of research, ChatGPT has itself provided suggestions to users to improve this situation (e.g. token limit, methods to deal with unresponsive API requests, etc.). It would not be an exaggeration to suggest that the engineering and implementation issues surrounding the use of APIs of LLMs remain a key hurdle to their use.

Decision resolution:

In experimenting with various prompts, it proved difficult for LLM agents to make a decision to buy or sell, even when prompted to do so. The results in C.1 were created with the following example prompt:

```
"Josephone has called David because she wants to sell UK bonds."
"UK government bonds trade at mid price, pricing is not a factor in
    deciding to trade"
"The point of the conversation is to decide to trade or not to trade"
"The conversation needs to go on until Josephine has sold the bonds
    or David has refused to trade any more bonds."
"Any trading decisions need to be clearly agreed"
```

These features compound the difficulty of using LLMs for trading simulations, as they lead to inefficient and unproductive conversations that do not result in meaningful trading decisions. Each conversation that is unproductive, at a minimum, incurs model cost and processing time. This limitation is a significant drawback when using LLMs for simulating trading interactions, as the primary goal is to have agents make clear and decisive trading decisions based on the given prompts and market conditions. As LLM reasoning models and retrieval methods advance, it is foreseeable that these limitations may well be overcome.

Costs and timeliness:

Each simulation takes computational resources and incurs costs charged by the LLM provider. Although costs vary between models, LLMs, and methods of access, the 300 simulations reported here cost under $1 USD, although testing on older models like GPT-3.5 Turbo cost approximately $20 USD despite smaller API prompts. Related to model costs are the number of tokens used to provide prompts (easily controlled in an ABM) and the number of tokens used to supply an answer from an ABM. The latter challenge directly impacts costs.

In designing TraderTalk, one of the hardest components to implement using an LLM was a sense of timeliness or urgency around the trading window, something that would appear to be taken for granted in human participants. Although this is only a hypothesis, it would seem that these challenges arise because LLMs are primarily designed for natural language processing and generation tasks, such as text completion, translation, and question-answering. They appear poorly equipped for the decision-making required for trading scenarios. An area of future study could explore whether fine-tuning might improve this result, though it would likely come at additional cost and with added complexity. Models struggle to understand the context and implications of trading decisions, leading to the observed issues of rambling, indecisiveness, and excessive dialogue. I leave it to future researchers to determine whether LLMs that supersede the GPT-4 class are better equipped or if fine-tuning of models (and all the challenges that it brings with it) will be necessary to improve fidelity.

4.4 Assumptions and limitations

Modelling constraints:

The results in this chapter highlight several modelling constraints; most notably, I have restricted testing to three OpenAI models – the three most current and widely available. Whilst there is little reason to suggest results cannot be generalised (e.g., that each model is likely to produce a different underlying distribution when asked), In order to achieve a broadly applicable result, the results reported here are for standard GPT versions without any fine-tuning. This was a deliberate choice to test base models; however, it is clear from the mixed results that future research should be carried out to improve the results that could be gained from fine-tuning methods.

Other LLM models:

Rapid development of smaller foundational language models opens promising research avenues. Future compact models (similar to [197]) will likely challenge the findings of this thesis, offering comparable performance with greater efficiency. Locally installable models would give users transparency against unannounced training alterations. Research should examine how these evolving models perform on the reasoning tasks studied here, as domain-specific fine-tuning becomes more accessible and affordable. Inexpensive fine-tuning could dramatically improve statistical distribution accuracy, potentially offering a compelling alternative to my proposed framework. This frontier significantly extends the analytical approach developed in this work.

Limited human bias testing:

Regarding human bias and machine bias, I have tested results for one known human bias, that of negative recency. There are many other forms that can be tested. For my purposes, the demonstration that one significant bias, applicable to binary series, was present in the data was sufficient to prove the point of its existence. I acknowledge that it will be very interesting to explore other biases with different data tests in the future. In particular, much of the body of research on random variable generation in LLMs lacks the specification of model subversion and detailed methodology of model sampling choices (i.e. are tests run using "one-shot" or "few-shot" sampling of LLM models, are API's used or on-line user interfaces etc.). Likewise, I tested only one web-crawl method to illustrate the bias in human usage of the words "yes" and "no" – other methods could be explored to verify the results. Also, data used to test outputs against true-random variables were obtained from Random.org's engine – other methods or sources could be used to verify similar results. I contribute to this research by highlighting the impact that model sub-versions have on results, and the significance of testing methods on the results.

Lastly, as explainable AI methods expand to cover LLMs, it would be an extension of this work to look at ways to ascribe causation for these model discrepancies in distributional outcomes and also intention-action differences. I leave these fascinating areas of additional research to others in the field.

4.5 Summary

The research in this chapter is motivated by the potential of LLMs to enhance human-like agent behaviours within ABM simulations, specifically in areas where it is hard to quantify or explain human choice (e.g., to trade or not to trade, without assumed rationality). However, several fundamental challenges emerged during initial development. At a programmatic level, the integration of LLM calls within simulation cycles raised efficiency concerns. More fundamentally, replacing traditional agent decision logic—whether deterministic or probabilistic—with LLM responses introduced unexpected behavioural patterns that warranted investigation. In this chapter, to understand these emerging properties systematically, I narrow the scope to examine LLM performance in binary decision-making tasks controlled by two different ABM frameworks, testing across 3 different GPT models, across model-configurations and with very simple prompts – the syntax of which could be easily utilised in typical ABM designs.

Part A Three LLM model sub-versions were tested for decision-making biases by examining binary decision outputs across: GPT-3.5 Turbo-0125, GPT-4-0125-preview, and GPT-4o-Mini-2034-07-18. These models were unable to adequately replicate a uniform distribution in independent sampling. Likewise, I find statistically significant performance variations between models (GPT 4 to GPT 3.5) and especially between specific sub-versions (GPT 4 and 4o-Mini), in addition to significant impacts of sampling methods on results.

I introduced terminology to more clearly describe API calling structures, which is a method of sampling: One-Shot (independent API calls with fixed prompt, each generating one decision, representing a distinct sampling event) and Few-Shot Querying (single API call comprised of multiple samples n with fixed prompt, in the same API call). It was shown that the few-shot methodology, while producing better distributional outcomes, is less practical for agent-based modelling applications, and nearly half of all tests result in non-Markovian decision sequences, indicating persistent temporal dependencies.

Using a vanilla Yes/No benchmark task, only GPT-4o-Mini-2024-07-18 (one-shot) achieves outputs not statistically different from a uniform distribution, though responses remain non-Markovian. Between model versions, substantial variations in response patterns are observed, suggesting that architectural and training differences significantly impact decision-making capabilities. To gain context of the representation of “yes” and “no” words in natural language, I perform a sampling of the Common Crawl using the cut-off date of the GPT-4o-mini-2024-07-18 model training (corresponding to the most SOTA model). I find that naturally occurring Yes/No responses also deviate substantially from uniform distributions.

Turning to GPT model parameters, I test the Temperature parameter, tuning in a range of 0.5 to 2.0. Results of tests for various temperature parameters on GPT-4o-Mini showed nonlinear responses to changes in temperature, while other models maintained strong biases regardless of temperature setting, demonstrating temperatures limited ability to alter results.

I further compared LLM test results to true random binary series and tested specifically for the common human bias of *Negative Recency* – finding LLMs have a mixed ability to ‘beat’ humans, with GPT-4o-Mini-2024-07-18 notably avoiding human recency effects in one-shot testing, though still producing non-Markovian outputs. Results from analysing the Common Crawl, used to gain context of Yes/no usage in common web data, produced confounding results. In web data there appears a strong bias to “no” rather than “yes” – conflicting with skewed outputs of GPT 3.5 and 4, reportedly trained on such data.

These findings expose systematic biases in LLM-based decision making, sensitivity to model sub-versions and the very significant importance of API sampling methods on output distributions. These results have critical implications for ABMs, particularly in finance, where Markovian properties are typically assumed. Based on these findings, future work could focus on improving LLM integration in financial agent-based models through improved real-time decision mechanisms, while investigating the relationship between model architecture and decision-making capabilities. Ultimately, this research pathway aims to better quantify and mitigate the impacts of decision bias in financial modelling environments for real-world applications.

Part B In the second part of this chapter, I introduced **TraderTalk**, a novel LLM behavioural agent-based model that simulates realistic human bilateral trading interactions without extensive model tuning. Using a state-of-the-art, non domain specific, non fine-tuned LLM within the Concordia framework (GPT-4o-mini-2024-07-18), I demonstrate limited yet realistic trade negotiations (**RQ2.1**), interpretation, and trade execution decisions (**RQ2**) at frequencies approximating those in US equity markets.

I implement a two-agent model in which an independent LLM is shown to be able to reconcile a decision and replicate the frequent disparity between trading intentions and execution. The latter feature is a key characteristic of transparent markets (e.g., US equity markets), and there is no reason to suggest that the results differ for opaque markets. Referred to as order-to-trade ratios, the agentic LLM implementation of TraderTalk in section 4.2 achieves ratios comparable to U.S. equity markets without extensive model tuning or domain-specific training. This shows that untuned LLMs, using agentic methods, can replicate trading scenarios at a similar rate to realistic order-to-trade ratios (5.7% execution rate from 58 mutual trading intentions), aligned with empirical market data.

However, frequent disregard for agent instructions is also observed along with a lack of contextual understanding of numerical values by LLM agents. This provides a foundation for future research into more complex multi-agent and multi-market simulations. Future work should enhance GABM's understanding of implicit trading rules and dynamic market conditions; by refining their ability to capture human decision-making, LLMs could offer more robust simulations for policymakers, regulators, and market participants alike. I document that it is now feasible to incorporate LLMs into agent-based models directly without the need for archetypes, or reduced agent fidelity, and that LLMs have been found to deviate from intended trading instructions, leading to systematic decision-making biases despite demonstrating LLMs' potential for multi-agent systems.

The novelty of the work presented in this chapter lies in the application of LLM behavioural agents within ABMs, and a detailed analysis of the nature and type of underlying LLM bias. A recognition of significant model sub-version dependence is highlighted, which should be of interest to researchers looking to incorporate LLMs into argentic methods. I demonstrate that LLMs enable nuanced human-like interactions without extensive model tuning. The model in Part B captured the decision-making processes in bilateral trade negotiation.

Building on these findings, future work should focus on several key areas: enhancing LLM integration in financial ABMs through improved real-time decision mechanisms; investigating the relationship between model architecture, underlying causes of bias, and decision-making capabilities; evaluating performance across non-OpenAI models; exploring methods to mitigate identified decision biases; and examining implications for broader financial modelling applications.

Chapter 5

TRIBE Model: Enhanced Market Maker and Client ABM

This chapter encompasses the work from the peer-reviewed publication:
Vidler, A., Walsh, T. (2025). "Shifting Power: Leveraging LLMs to Simulate Human Aversion in ABMs of Bilateral Financial Exchanges, A bond market study". In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS'25, forthcoming).*

In this chapter, I introduce the *TRIBE* model, an enhanced market maker (MM) and client ABM focused on **T**rading **R**elationships, **I**nteractions, and **B**ilateral **E**xchange of assets. Building on the theoretical framework developed in Chapters 3, TRIBE simulates an over-the-counter (OTC) bond market and incorporates autonomous client behaviours reflective of real-world dynamics. The model provides the basis for this chapter and also serves as the foundation for the next, where LLM-generated client behaviours are introduced.

5.1 Introduction

As demonstrated earlier in this thesis, ABMs are versatile technologies for modelling complex and dynamic systems and are suited for bilateral markets such as government bond markets. However, these markets, characterised by direct transactions between two parties without centralised exchanges, present modelling challenges. In particular, traditional approaches struggle to capture the complex interactions between MMs and the nuances of clients in these environments.

Although focused on Australia, the dynamics and results likely generalise to similar markets, such as the market of UK government bonds (GILTS) explored in earlier chapters. The TRIBE model incorporates UK regulatory frameworks as a complementary resource, utilising observed structural similarities between these markets to supplement the available dataset and introduce additional empirical observations where Australian documentation provides more limited coverage. This methodological approach enhances analytical comprehensiveness by addressing certain limitations in available data. The approach benefits from the common historical development of both regulatory environments, with UK documentation potentially offering more extensive empirical data and detailed implementation examples in specific areas of shared regulatory structures.

5.1.1 TRIBE: an artificial society and trading marketplace model

The TRIBE model simulates heterogeneous MMs trading stylised government bonds, with each MM modelled as an adaptive agent possessing its own utility function. Clients are modelled as grid-based, autonomous participants with dynamically assigned asset distributions, probabilistic trading availability, and decision functions designed to more accurately reflect real-world holding patterns. Client behaviours in TRIBE are extended to include variable availability, directional trading preferences, and differentiated trading intentions, moving beyond the simplistic, always-available assumptions common in earlier models.

While clients do not initiate trades or traverse the 2-dimensional grid, they do respond to

MM inquiries. For this reason, they are not explicitly referred to as agents, although they exhibit degrees of autonomy and agency within the system.

Looking ahead, Chapter 6 integrates the LLM-based behavioural framework introduced in Chapter 4 with the TRIBE model to form TRIBE(LLM). There, ChatGPT's GPT-4o-mini is used to generate client decision-making, enabling the model to capture behavioural nuances that may be overlooked by traditional probabilistic approaches.

In TRIBE I introduce several key aspects that extend earlier work in this thesis to provide more nuance for client modelling and behaviours.

1. Introduction of client autonomy where clients are modelled with independent choices:
   (a) Asset Distribution: client bond holdings and cash reserves follow a log-normal distribution based on Australian market data, initialised at simulation start.
   (b) Client Selection Mechanism: MMs are required to passively accept all client flow, but can selectively engage with available clients, departing from traditional Sugarscape-inspired active selection and preferential trading. This is in close alignment with the regulatory regimes in financial markets of interest.
   (c) Availability Modelling: Each time step assigns a probability for client availability in bilateral trading (in Chapter 6 this is determined by an LLM), reflecting real-world scenarios where clients are not guaranteed to always be available to trade.
   (d) Trading direction: clients randomly assume buyer or seller roles when trading with MMs.
2. MM Specifications: MMs must prioritise client trading requests before engaging in inter-dealer trading, reflecting regulatory requirements.
3. Structural refinements:
   (a) Clients and MM's only ever transfer one asset: either bonds, or cash.

(b) The trading rules for MMs and clients are independent. For example, clients can be constrained to only be long bonds at all times, while MMs can have the flexibility to constrain cash and bonds.

(c) Negotiation of the quantity of trade. When MMs are trading with clients or other MMs, there is the flexibility to specify a simple form of trading rule, resulting in a negotiated quantity of bonds or cash (short positions) to be transferred. For MM-to-MM trading, a geometric mean of the largest quantity is used as the policy. In line with earlier work, the default is a geometric mean of the quantity. This framework enables future researchers to systematically evaluate diverse market policies, a natural extension that lies beyond the scope of this thesis.

Chapter structure

This Chapter progresses through the base TRIBE model, establishing the fundamental market dynamics through a series of experiments: initial model calibration (E1), client availability testing [E1(b), E1(c), E1(d)], trading direction analysis [E1(e)] and impact on the MM balance sheet [E1(f) and E1(g)]. Extensions of this base model are explored in Chapter 6.

5.2 Technical model description: TRIBE ABM definition

To formalise the TRIBE model's structure and behavioural assumptions, I extend Definition 1 from Chapter 3 and detail client based enhancements. The following section presents the core mathematical framework underpinning client and MM interactions.

Definition 2. A TRIBE model is defined as

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

where: $N = \{MM_1, MM_2, \ldots, MM_n\}$ is the set of market making agents. $\mathcal{L}$ represents the set of clients, CC_j who collectively make up the landscape states, following the log-normal distribution of assets observed where each landscape state contains resources $R = \{b_r, c_r\}$, corresponding to the bond (b_r) and cash (c_r) quantities held by individual clients. Each agent MM_i has a state $\mathcal{S}_i$, which includes cash usage rates M_c (representing individual cost structures), accumulations A_b, A_c (representing bond and cash accumulations over time), and a size of client base v_i, which describes the number of clients the agent services.

The perception functions $\mathcal{P}_i$ map the landscape and the states of other agents to a perceived state, defined as $p_{ij} : \mathcal{L} \cup \mathcal{S} \rightarrow \mathbb{R}$. This function models how agent MM_i perceives the landscape and the states of other agents. The landscape is made up of Clients such that $CC_j \in \mathcal{L}$. In this model, there is no bond or cash replenishment for the landscape of clients, reflecting a closed system, similar to real-world markets with fixed bond issuance.

The decision rules $\mathcal{D}_i$ map the perceived states of the landscape clients and a MM agent's own state to actions, $d_{ik} : \mathcal{P}_i \times \mathcal{S}_i \rightarrow \mathcal{A}_i$, where actions $\mathcal{A}_i$ for agent MM_i include serving clients, trading with other MMs, and ceasing operations when necessary and where $\mathcal{A}_i$ for clients CC_j include deciding to trade with a MM.

The landscape evolves according to a transition function $f : \mathcal{L} \times \mathcal{A} \rightarrow \mathcal{L}$, which dictates how the landscape changes in response to the actions of the agent.

For all aspects, the model progresses through discrete time steps, denoted by $t \in \{1, 2, ..., T\}$, where T represents the final period. Here, each time step represents a round of interactions between agents and updates to the landscape and is limited in this model to 1500 time steps (a limit equal to a 12.5-hour trading day with 2 steps per minute).

5.2.1 Market makers

A market maker MM_i is defined by the tuple:

$$MM_i = \{(x_i, y_i), B_i, C_i, v_i, M_i\}$$

Where for time t:

- $MM \in (1, 2 \ldots i)$ is the number of MM agents
- (x_i, y_i) is the position on the grid
- B_i is the amount of bonds held
- C_i is the amount of cash held
- v_i is the vision range (constant for all time = t)
- $M_i = M_c$ agent specific cost of doing business (e.g. metabolism rate for each agent, applicable to cash holdings only, constant for all time = t)

5.2.1.1 MMs engaging with clients: servicing clients

Market Makers traverse the network of clients seeking trading opportunities. TRIBE extends the prior model in Chapter 3 by introducing a critical refinement: the ability to specify the proportion of clients available for trading at each time step. This represents a significant advance over previous approaches that assumed continuous availability of all clients, better reflecting the intermittent nature of actual OTC market participation.

Market makers have a regulatory requirement to service client trading. Within TRIBE, each MM can service at most one randomly selected client within their vision, per time step. MMs contact the client – the client CC_j then chooses if they wish to be a buyer or seller of bonds. This mechanism is detailed in more detail below in the Client section.

5.2.1.2 MMs trading with other MMs: reducing imbalances

In addition to facilitating trade with clients, MM can also trade with another MM. The trading flow between an MM (MM_i) and a MM (MM_j) begins by analysing the differences in positions between the two MMs at each time step, **only after** any possibility of trading with a client has been ruled out by the MM. Key to the idea of MM trading is to identify one side in surplus of the other's deficit. The sum of the differences is used as the variable of maximisation. In practice, this will force MMs to only trade with other MMs that have large imbalances in the opposite direction to the analysing MM. This is a proxy for finding an MM with the strongest incentive to trade.

i. MM trading partner selection criteria: Select random MM within field of vision. Each MM is allowed to contact all other MMs once only in serial, until such point as one trade is found. If no trade is found, the simulation moves on to the next agent to analyse.

$$\forall i, j, Test \begin{cases} B_i > B_i \text{ and } C_j > C_i \text{ then} : T_{i,j} = C_j - C_i + B_i - B_j \text{ (Scenario1)} \\ C_i > C_j \text{ and } B_j > B_i \text{ then} : T_{i,j} = C_i - C_j + B_j - B_i \text{ (Scenario2)} \\ \text{else } Test_{ij} = 0 \end{cases} \tag{5.1}$$

ii. Trading quantities are determined by $T_{i,j}$ and follow the below,

$$T_{i,j} = \begin{cases} \text{if } B_i > B_i \text{and} C_j > C_i \text{then} : T_{i,j} = \sqrt{B_i B_j} \\ \text{if } C_i > C_j \text{ and} B_j > B_i \text{ then} : T_{i,j} = \sqrt{C_j, C_i} \\ \text{else } Test_{ij} = 0 \end{cases} \tag{5.2}$$

iii. The new holdings post trade ($\hat{B}_i$ and $\hat{C}_i$) are updated with the following bonds and cash held by the MM and the client after the trade.

$$\text{Scenario 1} = \begin{cases} \hat{B}_i = B_i - Test_{ij} \\ \hat{B}_j = B_j + Test_{ij} \\ \hat{C}_i = C_i \\ \hat{C}_j = C_j \end{cases} \tag{5.3}$$

$$\text{Scenario 2} = \begin{cases} \hat{B}_i = B_i \\ \hat{B}_j = B_j \\ \hat{C}_i = C_i - Test_{i,j} \\ \hat{C}_j = C_j + Test_{i,j} \end{cases} \tag{5.4}$$

5.2.1.3 Metabolism

In real markets, MM's have large regulatory costs. Here, I model costs and other constraints through the use of the metabolism rate M_i of a MM_i , and apply it as:

$$C_i(t+1) = C_i(t) - M_i$$

where $C_i(t)$ is the cash held by the MM at time step t. Client costs are excluded from the model, avoiding complex assumptions about their distribution and behaviour. Although cost modelling could account for client heterogeneity and nonlinear allocation, this extension is left for future research.

5.2.1.4 Initial MM bonds and cash holdings

In each simulation, all agents MM_i are initialised with starting bonds and cash holdings. Using data and market rules for UK government bond MMs (GEMMS) [218], it is expected that each MM holds at some point around 2% of the bonds on issue. In TRIBE this is implemented by calibrating a range of bond values with a mean equivalent to this level of the total bonds on the grid.

$$B_i \sim \mathcal{U}_{\text{discrete}} \left[\text{bondrange}_{[0]},\ \text{bondrange}_{[1]}\right] \tag{5.5}$$

where:

- $\mathcal{U}_{\text{discrete}}$ represents a discrete uniform distribution

- $B_i$bondrange[0] is the lower bound (inclusive)
- $B_i$bondrange[1] is the upper bound (inclusive)

This random selection ensures that each integer value in the range has an equal probability of being chosen.

Cash holdings, C_i, are initialised in the same way, with levels one tenth that of bonds. These values are self-consistent within the model and have little available data to suggest any other levels. Although this model uses a basic implementation of client availability, I acknowledge that more sophisticated approaches could be developed. The investigation of alternative availability mechanisms remains an open question for future research.

5.2.2 Clients

Clients are represented as cells on a grid. To aid in conceptualising these clients, I will refer to them as 'cell Clients' and 'clients' interchangeably. Each cell represents a unique client. There is no "regrowth" of cells or client assets throughout the simulation. Thus, the simulation represents a one-day ecosystem with a fixed maximum quantity of bonds and cash, in line with real-world asset markets. Cell locations (i.e. clients) are initialised with bonds and cash values distributed according to a log-normal distribution at the inception of a simulation. This distributional assumption draws heavily from data provided by the distribution of superannuation fund balances in Australia (please refer to Figure 5.2).

Client Definition: A cell client CC_j is defined at time t by:

$$CC_j = \{(x_j, y_j), B_j, C_j, D_{j,t}, A_{j,t}\} \tag{5.6}$$

Where (x_j, y_j) is the grid position, B_j and C_j are the amounts of bonds and cash at time t, respectively, $D_{(j,t)}$ indicates the direction of trade, and $A_{j,t}$ is the availability of a client to trade, determined probabilistically at each time step, if selected by a MM at that time step. Given that each client can exist in only one location in the grid, $A_{(j,t)}$ is equivalent to $A_{(x,y)}(t)$. In the future, I will use the terms interchangeably, referencing $A_{(j,t)}$ when discussing a client j and using $A_{(x,y)}(t)$ when focussing on availability in a general way.

Note, MMs are also given locations on the landscape such that they too have a grid positions (x_j, y_j). Each grid position in the landscape can have either one client or one MM at any point in time.

Note, within the TRIBE(LLM) version, detailed in Chapter 6, $A_{(x,y)}(t)$ is replaced with a call to an LLM which produces a boolean response, 'yes' or 'no' as the LLM output at time $= t$.

Bonds and cash are initialised as:

$$B_j \sim \text{LogNormal}(\mu_B, \sigma_B^2), \quad C_j \sim \text{LogNormal}(\mu_C, \sigma_C^2) \tag{5.7}$$

Parameters μ_B, σ_B, and corresponding values for cash, are set based on available data such as that provided by the Australian superannuation regulator [266].

The **availability of the client** j **at any given time step** t for the position (x_j, y_j) is initially modelled as:

$$A_{(x_j,y_j)}(t) = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases} \tag{5.8}$$

where p is the probability of preset availability, the statics per simulation and the roles of the clients as buyers or sellers, and $D_{(j,t)}$ is drawn from a Bernoulli distribution. This distribution reflects the intuition that not all clients are likely to want to trade at every time step, even if contacted by a MM. Since each outcome is a binary decision – "Yes" or "No" – a Bernoulli distribution is used to model this behaviour.

In a bilateral marketplace, it is unrealistic to assume all clients are "online" at all points in time for various physical constraints such as time zones or personnel and staffing issues.

This is a key advancement from previous work that relied on clients passively agreeing to trade at all time steps. This turns clients into active participants in the model framework with autonomy and allows for granular modelling of both large and small clients and the ability to allow for heterogeneous client states such as clients who have received a large cash inflow through their business activities or perhaps who have been left with too little cash for trading activities. Together, $D_{(j,t)}$, and $A_{(x,y)}(t)$ form the proxy for a **client trading utility function**.

It might seem odd that the desire to buy or sell is not itself modelled endogenously; the rationale for this is that the complexities of client trading direction may not be directly linked with an imbalance of cash or assets. The work of [267] discusses similar concepts (though applied to a different, though similar, type of bond market asset). Other methods for modelling client desires are explored later in this research (Chapter 6), including using an LLM to decide upon a clients inclination to trade immediately. For a trade to occur, transactions strictly involve bonds or cash and adhere to standard accounting principles, with transactions with clients and with trading volumes between MMs negotiated using geometric means, setting the stage for future research on negotiation mechanisms. These model enhancements enable dynamic simulations of trading behaviours, accommodating variations in client size and strategy, with MMs sequentially engaging with clients to optimise trades. This setup paves the way for future studies in modelling client behaviours through game theory and LLMs, enhancing realism as noted in the existing literature. A graphical description of the model is presented in Figure 5.1.

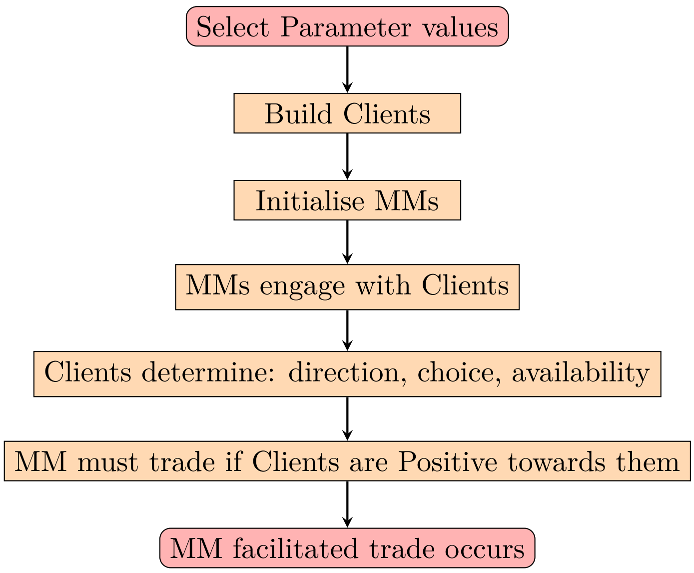


Figure 5.1: TRIBE model architecture

5.2.3 Internal model mechanisms: time steps, asynchronous trading and bilateral trade restrictions

Unlike the earlier model in Chapter 3, the TRIBE model is a sequential ABM. In each iteration of the model, clients begin by updating their availability. The order of MM's participating in the simulation is randomly shuffled each time step. The order of MMs is randomised at each time step, and then they, sequentially, contact a client within their vision. Each MM must trade with the client cell it contacts, **if the client chooses to trade with the MM**. After only one attempt per time step to contact a client, an MM can also trade resources with another MM – **but only if** it is mutually beneficial (moving to the selected cells based on their evaluation of equation 5.2). In this way MM's receive or transfer bonds or transfer cash based on the client's resource need – servicing a client trade. As with the earlier model work in Chapter 3, MM's incur a cost of business to their cash resource holdings, reducing their cash based on their metabolism rate.

Market makers employ a passive approach to trading that aligns with their regulatory requirements and client preferences. If client trading is thought of as a "must do" for a

MM, then MM-to-MM trading is a "nice to have". MMs calculate the trade's value by comparing resource differences, ensuring the transaction does not result in any negative balances by considering only the minimal excess of bonds and cash available. Clients and MMs are not allowed to have negative balances. For clients, this is by construction. For MMs a negative balance can only occur through resource depletion. Once depleted, MMs are removed from the simulation and effectively "die".

A fundamental feature within the model design is the bilateral nature of MM and client trading such that any MM can only trade with one client at a time, and vice versa. This design principle is a defining characteristic of many financial asset markets. Any relaxation or alternation of this feature would be similar to testing moving from bilateral trading to methods such as exchange-based trading. The testing of an exchange-based mechanism, in which any one market participant could trade with another (such as discussed in [268]), is outside the scope of this present research.

Within the model several internal model mechanisms, specific to client modelling, are variably deployed.

i. Client availability: What proportion of clients should be available for trading given real-world time constraints and work commitments? This addresses agent coalition formation and coordination mechanisms.

ii. Clients with no existing holdings: Should clients without holdings be excluded? If so, how could MMs realistically identify them before engagement, given log-normal distribution patterns? This touches on issues of information asymmetry and strategic decision-making.

iii. Asynchronous client activities: How should the simultaneous attempt to trade, be handled between multiple MMs and a single client? This covers issues of concurrent access in distributed systems, protocol design for decentralised markets, and coordination mechanisms without explicit communication.

Additionally, how do MMs respond, within each time step, to this activity? These internal mechanisms include the following.

i. Market Maker Trading Requirements: MMs must trade with any client encountered, preventing preferential treatment and strategic coalition formation. Although allowing selective trading could enable interesting collusion dynamics, this constraint maintains market fairness. Further exploration of relaxed trading requirements remains an open research direction.

ii. Client to Agent activities,

   (a) Decision to trade: Information asymmetry needs to be preserved and bounded rationality in agent behaviours is assumed. For example, a client's decision to trade is divorced from the direction of trading. This is very deliberate, as it remains largely impossible to calibrate the trading desire without trying to assume some decisions that do not have a good method of calibration or validation. For instance, if we know that a client has a large bond holding, how are we supposed to infer their desire to sell? Perhaps they may actually wish to increase their holding given a new piece of news about bond markets.

   (b) Time horizon: The model assumed discrete time-step simulations and state persistence constraints between steps. TRIBE models client engagements over many time steps, but within a stylised day only. The entire simulation itself is one virtual trading day. An assumption made within the model is that client assets are thus those assets (bonds or cash) that they wish to trade for the day only, as might be expected for hedging flows or small trading alterations.

   (c) Trading direction: The trading direction is randomly sampled from a normal distribution to prevent the emergence of bias and maintain the selection of the random strategy. While in previous experiments I linked the direction of client trading to bond and cash imbalance (Experiment 1), this created predictable patterns of bonds being sold by clients and accumulated by MMs due to structural initialisation imbalances. To reduce this bias, TRIBE assumes a random choice between trading bonds and cash.

iii. What does a MM do if it reaches out to a client (via randomly accessing their grid/vision), and the client does not respond/is not available/does not want to trade with them right now? This is a crucial aspect and covers several ABM areas.

   (a) Strategic matching: The aversion to trading might be structure (the client is simply unavailable at this time step), it might be preference-based (a client might not like a specific MM), or it might be financially rational (the client has no desire to trade given other external market factors).

   (b) Mechanism design aspects each MM can at-most have one "virtual" phone call only to a client per time step – implementing rate limiting consistent with the mechanism of the market.

   (c) Agent interaction rule set: if for any reason the client is not available, the MM must wait until the next model time step to contact another client. This rule is designed to mimic real-life scenarios. This system dynamics maintains order, without which the model would devolve into a constant trading simulation with 'greedy' agents.

   (d) Agent search protocol: MMs reach out to a client via randomly accessing their grid/vision.

This design prevents dominant strategies while maintaining a realistic market microstructure through constrained agent interactions.

5.2.4 Calibration to available data

Results derived from the Australian Prudential Regulator show that as of December 2023 there were 1701 regulated pension holders alone, of which 1529 report their total asset holdings[1]. Although not a comprehensive direct relationship to bondholder sizes, given

[1] `https://www.apra.gov.au/annual-fund-level-superannuation-statistics$`, accessed 5 Aug 2024.

the practice of holding a minimum percent of pension assets in government bonds, the asset size of a pension fund does produce a proxy calculation for the distribution of government bonds. Furthermore, some 250 funds request that their data be withheld from public view for privacy reasons. This brings together a minimum estimate of 1951 Australian government bond holders. APRA also does not regulate self-directed pensions and foreign pensions, all of whom make up clients in the ecosystem. Combined, this would suggest that a reasonable estimate of bond holders is 2,500 (represented by a 50 x 50 client grid). As such, I include the data from December 2023 to illustrate the non-normal nature of the distribution, with the following statistics:

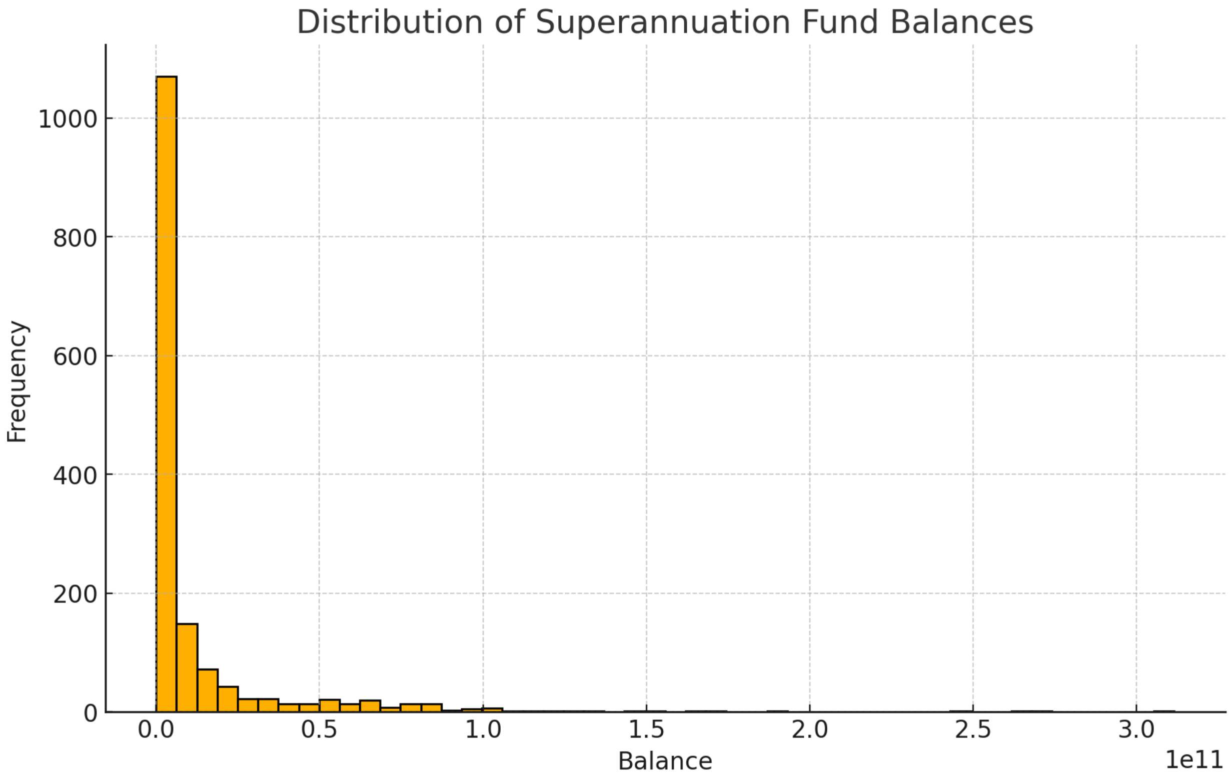


Figure 5.2: APRA data: balance in billion Australian dollars (x-axis).

Statistic	**Value**
Count	1525
Mean	$11,692,880,000
Standard Deviation	$26,925,220,000
Minimum	$11,000
25th Percentile	$416,516,000
Median (50th Percentile)	$1,976,100,000
75th Percentile	$8,477,939,000
Maximum	$311,497,761,000

Table 5.1: Descriptive statistics of superannuation fund balances.

5.2.5 Experiment design and testing methods

In TRIBE, the focus of this model is on the analysis of the impact of one market participant group (Clients) and their impact on market stability. To this end, I detail a series of experiments, testing various aspects of client availability, in particular the mean probability that a client contacted by a MM at time step t is available for trading at that time step with the particular MM. To help with discussions, I adopt the convention of numbering each experiment and thereafter referring to them as E1 etc.

Whilst there are almost a limitless number of aspects of the model to test, analysis is carried out on the impact of client availability and the interplay with MM balance sheets. Another statistic that is variably used in this chapter and the next is that of "carrying capacity": the proportion of all assets in the landscape at t=0 which get traded over the course of the simulation (i.e., the "carrying" capacity of the simulation). This is a way to look at the amount of liquidity that is able to be "carried" in a simulation.

A summary of the experiments are provided in Table 5.2:

Table 5.2: Summary of experimental setup and features. Note all distributions use $N(\sigma, \mu)$

Group	**Feature**	**Test Stat.**	**Output Variables**
E1(a)	Client Availability	100%	MaxAgentLife, Carry Capacity, Trading Ratios
E1(b)	Client Availability	50%	MaxAgentLife, Carry Capacity, Trading Ratios
E1(c)	Client Availability	25%	MaxAgentLife, Carry Capacity, Trading Ratios
E1(d)	Client Availability	10%	MaxAgentLife, Carry Capacity, Trading Ratios
E1(e)	Trade Direction Mechanism	n/a	Trading Ratios
E1(f)	Low Balance Sheet & Client Availability	50%	MaxAgentLife, Carry Capacity, Trading Ratios
E1(g)	Low Balance Sheet & Client Availability	10%	MaxAgentLife, Carry Capacity, Trading Ratios

5.3 TRIBE model experiment 1

5.3.1 Experiment 1: TRIBE traditional ABM (E1(a) benchmark)

First, the analysis commences with the utilisation of earlier model settings and constraints within TRIBE. Agents are set at 4, reflecting Australian market conditions, with client bond distribution and cash distribution estimated to be log-normal. Across 100 simulations of an artificial trading society, TRIBE is run with assumptions similar to earlier ABM work, namely that all clients with non-zero holdings wish to trade with MMs and other settings that can be found in Table 5.3. I rely on the parameter sets in Table 5.3 for all tests. Much of these data are derived from various stylised facts in the literature in [112,143], and also data published by the Australian and UK governments in [154, 218, 266] respectively. Future work in this vein will involve expanding the market making agent numbers beyond the current Australian calibration of four agents. However, as the focus of this study is primarily on client behaviour, keeping the number of market-making agents static is a reasonable simplification. Prior research also suggests that results are unlikely to differ significantly with larger agent populations, given demonstrated agent stability—though this remains an interesting direction for further exploration.

Parameter	*Category*	*Value*
Grid Size	General	50 x 50 (2500 clients)
Number of MMs	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	50 to 1500 (random dist.)
Initial Cash Range	MM Settings	75 to 150 (random dist.)
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	Utility function component	Variable, $A_{j,t}$
Client Trade Direction	Utility function component	Bernoulli (Buy/Sell)
Dealer Contact	Mechanism Design	Singular per step
Step Process	Mechanism Design	MM singular per step

Table 5.3: TRIBE ABM simulation parameters and values.

5.3.2 Results: E1(a)

This test contained 100 simulated trading societies with a maximum possible of 1500 time steps per simulation. In this test $A_{(x,y)}$ was set to 100% (see 5.8) and $D_{(j,t)}$ was drawn from a Bernoulli distribution of Buy/Sell options. The key results are presented below and draw attention to **MaxLife**, a measure of how long agents persist and keep trading. I show that more than 75% of simulations existed up to the theoretical maximum time and that the average artificial trading society was able to service 76% of all landscape bonds. Separately, MM-to-MM trading (so-called "inter-bank" trading, or MM-2-MM) represented a mean of 32% of all assets traded assets, according to government data from the AOFM and reference results from Chapter 3. All simulations had at least one agent still transacting.

TRIBE implements methods used in work such as [267] and appears to establish an artificial society that is stable and capable of carrying the capacity of trading volume clients desire – a concept explored in ABMs for financial market simulation in [79]. However, variation between agents and simulation sets is observed (i.e. a mean of 1324 for agent life and 18% dispersion between quartiles of MM-to-Client bond trading activity).

Note: Alternative parameterisation was systematically explored in Chapter 3, including variations in vision, metabolism, systems with other agent numbers and other agent configurations. These investigations directly informed the design choices made for Experiment E1(a), particularly in terms of various parameter ranges, which informed the parameter boundaries used in the design of E1(a). The reduced-form analysis demonstrated that while some parametrisation showed promise in specific scenarios, the chosen configuration provided the best balance of model performance and nuance of agent behaviour.

Statistic	**Agent MaxLife**	**MM-to-Client Bond Trading**	**MM-to-Client Cash Trading**	**MM-2-MM (Bond)**	**MM-2-MM (Cash)**	**MM-2-MM Market Share**
mean	1324	76%	75%	32%	29%	32%
std	440	9%	9%	14%	11%	13%
min	6	49%	48%	4%	3%	5%
25%	1499	69%	68%	22%	23%	23%
50%	1499	78%	76%	32%	30%	32%
75%	1499	84%	84%	44%	38%	42%
max	1499	87%	86%	57%	53%	56%

Table 5.4: **E1(a)** results with MM market-share. Ratios are presented as a percentage of total client available holdings. All clients were "available" and results are rounded to the nearest percentage.

5.3.3 Extending E1: implications of availability, testing E1(b), E1(c) and E1(d)

As with all ABM's, there is a virtually limitless array of model parameters and mechanism tests that can be carried out. The goal of this chapter is to look at the impact of improved client modelling. The parameters most closely related to client behaviour are availability $A_{(j,t)}$ (see 5.8) and Trade Direction $D_{(j,t)}$. As such, I further test the following.

- Test **E1(b)**: Stressing $A_{(j,t)} \approx 50\%$, for all (j, t), will replicate client activity where there is a 50% chance that any one client wants to trade at any time. Ceter Paribus.

- Test **E1(c)**: Stressing $A_{(j,t)} \approx 25\%$, for all (j,t), will replicate client activity where there is a 25% chance that any one client wants to trade at any time. Ceter Paribus.
- Test **E1(d)**: Stressing $A_{(j,t)} \approx 10\%$, for all (j,t), will replicate client activity where there is a 10% chance that any one client wants to trade at any time. Ceter Paribus.

Results: E1(b) $A_{(x,y)} \approx 50\%$

Client availability follows a normal distribution $\mathcal{N}(\mu = 0.5, \sigma = \sqrt{0.3})$, with values below 0.5 representing a 50% probability of accepting a MM's trade request where,

$$A_{(j,t)} \sim \mathcal{N}(\mu = 0.5, \sigma = 0.3) \tag{5.9}$$

$$\text{decision} = \begin{cases} \text{yes} & \sim 50\% \\ \text{no} & \sim 50\% \end{cases}$$

Statistic	**Agent MaxLife**	**MM-to-Client Bond Trading**	**MM-to-Client Cash Trading**	**MM-2-MM (Bond)**	**MM-2-MM (Cash)**	**MM-2-MM Market Share**
mean	1179	35%	35%	32%	28%	31%
std	548	5%	5%	14%	10%	13%
min	1	16%	18%	4%	8%	5%
25%	951	32%	32%	21%	20%	21%
50%	1499	35%	35%	32%	27%	32%
75%	1499	38%	38%	42%	34%	41%
max	1499	45%	45%	61%	57%	60%

Table 5.5: **E1(b)** clients only 50% available. MM market-share remains largely similar, other ratios halve as expected.

Comparing E1(a) and E1(b): analysis of market-making behaviour under reduced client availability

Reducing client availability by half produces several nonlinear impacts. The average agent life reduces slightly from 1324 to 1179 steps, although the distribution of lifespan is clearly impacted, with 25% of the lives now being less than or equal to 951 steps. See Figures 5.3 and 5.4 for the full distributions.

To better illustrate the dynamic shift in virtual society features, I include a series of graphs of agent and simulation variables. The following includes a histogram of the agent life spans in all simulations and a cumulative frequency graph. *Please note that, given the propensity for the agent's life to reach a maximum value, the cumulative graph increases as the right-hand board of the graph in high-client-participation environments.*

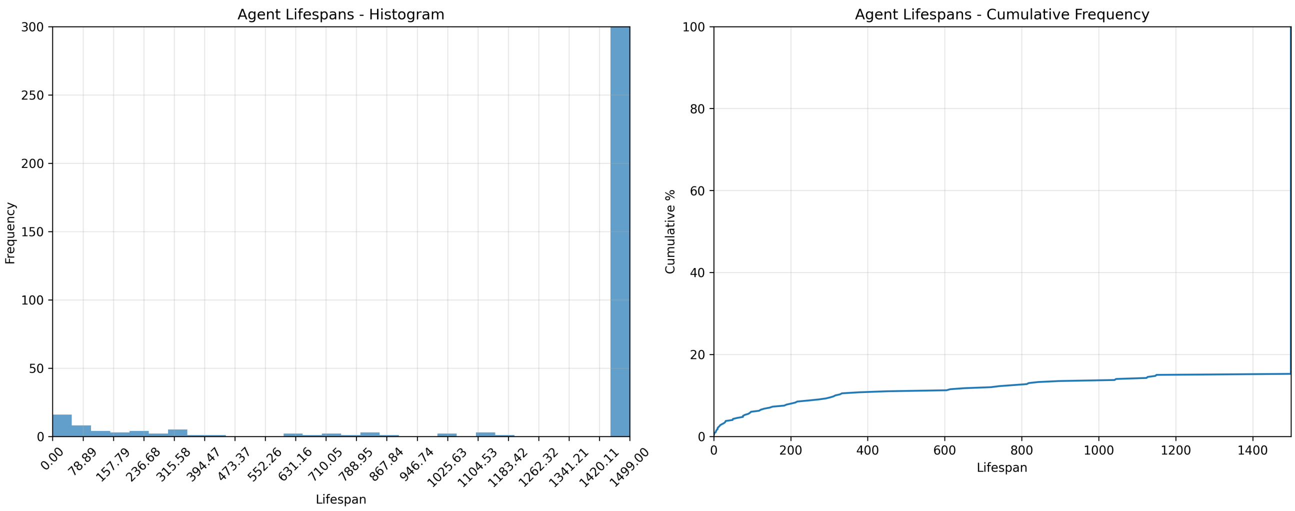


Figure 5.3: E1(a) 100% availability: the graph on the right goes to 100% cumulative value at the graph edge.

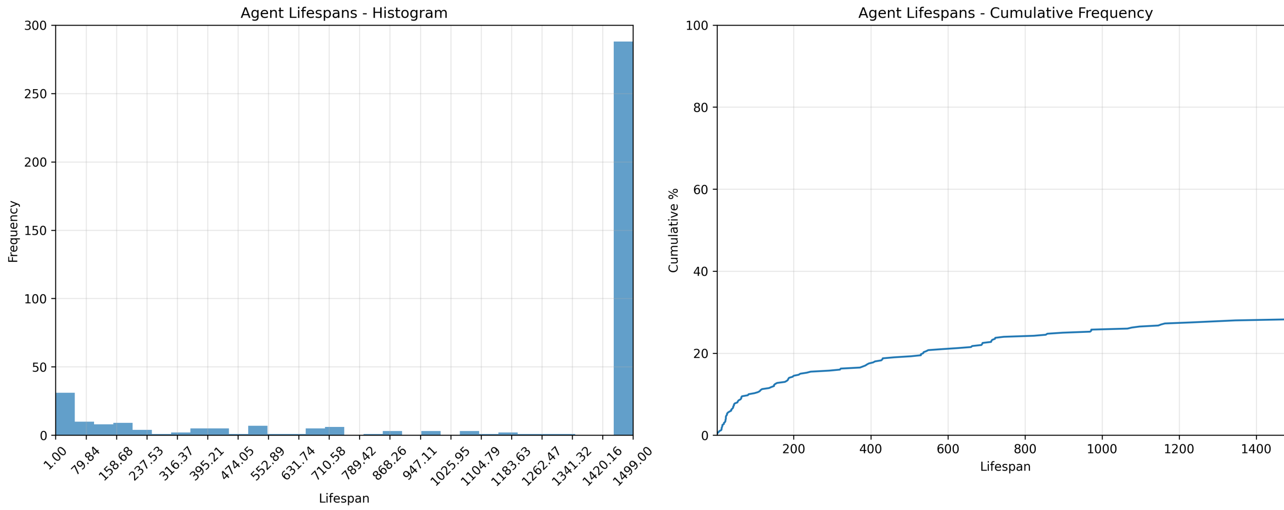


Figure 5.4: E1(b) 50% availability agent life spans.

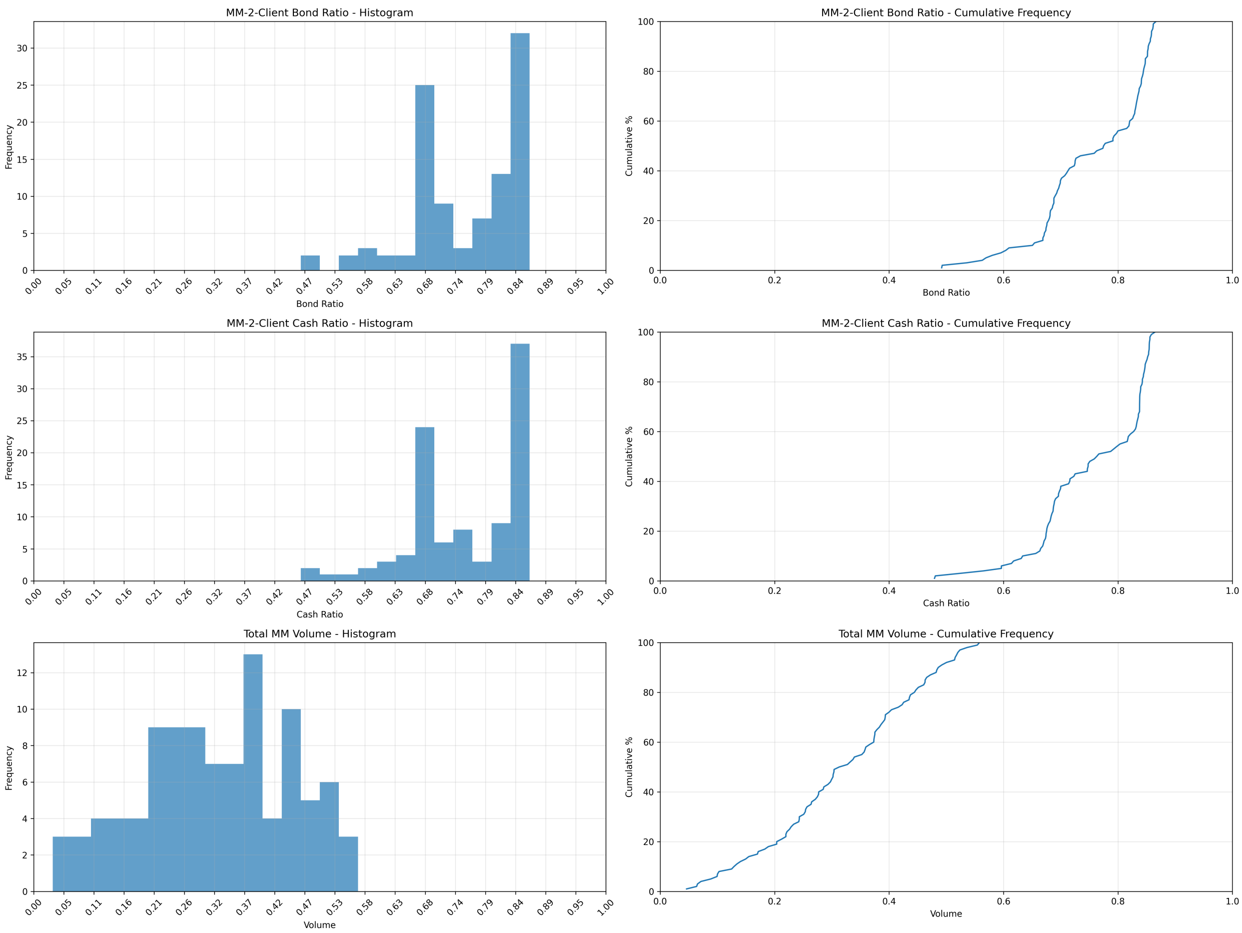


Figure 5.5: E1(a) 100% availability trade ratios.

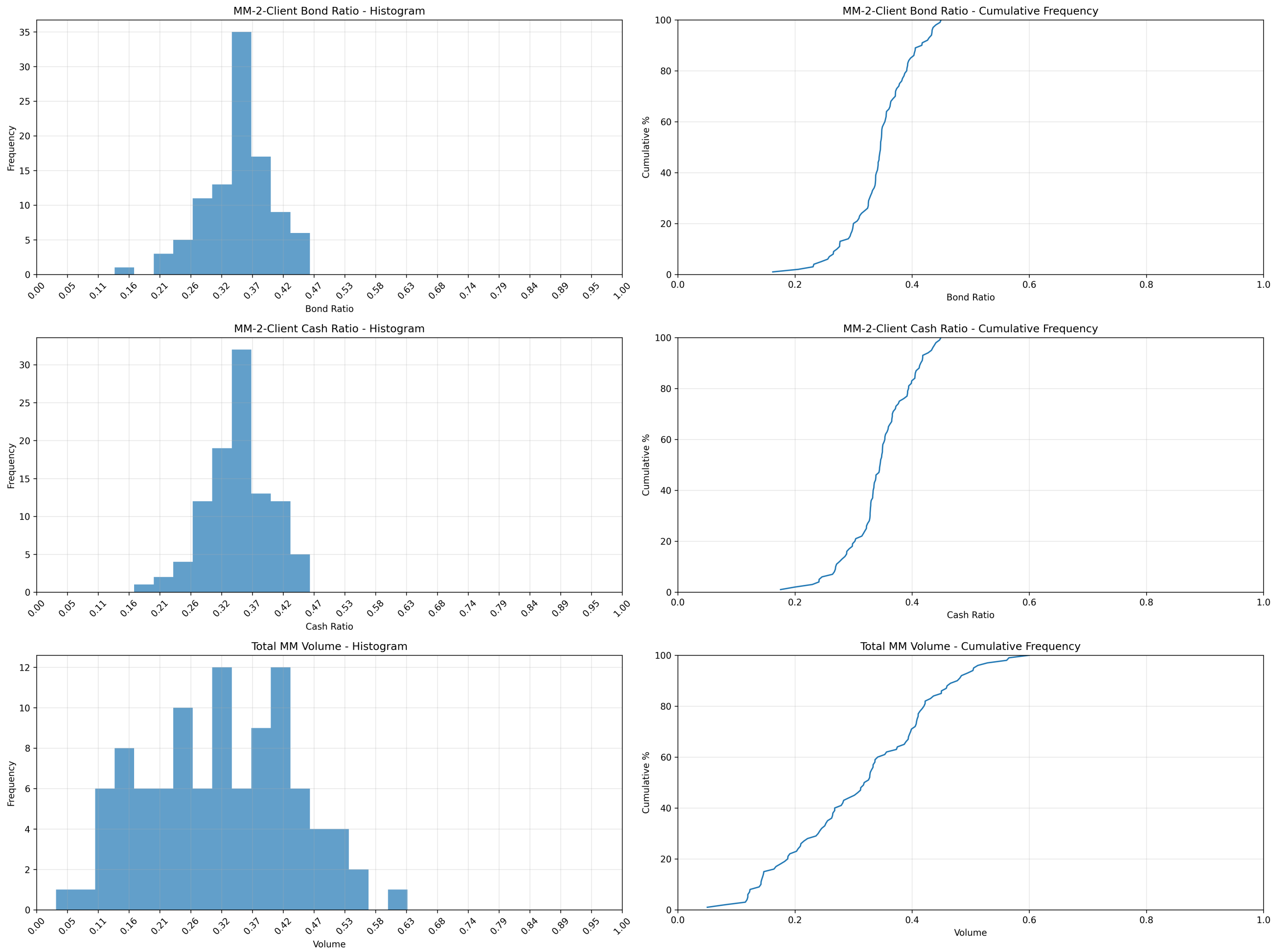


Figure 5.6: E1(b) 50% availability trade ratios.

The simulation results show distinct patterns in market making behaviour. In particular, MM-2-MM metrics exhibited stability, even in the face of reduced client trading availability. MM resilience would imply a degree of independence from client activity.

Volatility metrics provide additional information on system behaviour. The standard deviation in MM-to-Client trading decreased substantially from 9% to 5%, suggesting more uniform trading patterns under reduced client availability. However, MM-2-MM metrics maintained consistent volatility levels (standard deviation ∼11-14%), indicating that inter dealer market dynamics retain their characteristic variability regardless of client participation levels to these levels. Again, results average close to the AOFM levels.

A particularly interesting observation emerges in the minimum trading values for MM-2-MM activities, which showed increased dispersion between bonds and cash in E1(b), accompanied by a nonlinear impact on the distribution of MM-2-client trading activities. The dispersion in the distribution of MM-2-Client trading percentages experienced by MMs increased in scenarios of lower client activity. For example, in E1(a) we see that the median activity of bonds is 78% with a maximum of 87%; however, in E1(b) we see the median at 35% and the maximum at 45%, a point difference 10% again but at a smaller value. This divergence potentially signals distinct adaptation mechanisms for different asset classes by MM's, when operating under client availability constraints. Looking at figures 5.5 and 5.6, we can see a shift in MM-2-Client bond ratios distributions across simulations to the left, i.e.,to smaller ratios of trading assets. The same pattern is observed in the MM-2-Client cash ratios. However, as evidenced from the tables, the final line of graphs is the Total MM Volume distribution, and this has a shift to higher values and an overall distributional shift right.

Results: E1(c) $A_{x,y} \approx 25\%$

Client availability follows a normal distribution $\mathcal{N}(\mu = 0.25, \sigma = 0.22)$, with values above the 0.40 threshold representing approximately 25% of the results. The parameters are calibrated to achieve this target probability,

$$A_{(j,t)} \sim \mathcal{N}(\mu = 0.25, \sigma = 0.22 \tag{5.10}$$

$$\text{decision} = \begin{cases} \text{yes} & \sim 25\% \\ \text{no} & \sim 75\% \end{cases}$$

Statistic	**Agent MaxLife**	**MM-to-Client Bond Trading**	**MM-to-Client Cash Trading**	**MM-2-MM (Bond)**	**MM-2-MM (Cash)**	**MM-2-MM Market Share**
mean	977	18%	18%	34%	31%	34%
std	609	4%	4%	15%	9%	13%
min	1	7%	8%	0%	9%	5%
25%	300	15%	15%	25%	24%	26%
50%	1499	17%	18%	35%	31%	35%
75%	1499	20%	21%	43%	36%	42%
max	1499	27%	28%	69%	53%	67%

Table 5.6: **E1(c)** results at 25% client availability, ratios presented as percentage of total client available holdings.

Results: E1(d) $A_{x,y} \approx 10\%$

Client availability follows a normal distribution $\mathcal{N}(\mu = 0.20, \sigma = 0.32)$, with values above the 0.60 threshold representing approximately 10% of the results. The parameters are calibrated to achieve this target probability,

$$A_{(j,t)} \sim \mathcal{N}(\mu = 0.20, \sigma = 0.32) \tag{5.11}$$

$$\text{decision} = \begin{cases} \text{yes} & \sim 10\% \\ \text{no} & \sim 90\% \end{cases}$$

Statistic	Agent MaxLife	MM-to-Client Bond Trading	MM-to-Client Cash Trading	MM-2-MM (Bond)	MM-2-MM (Cash)	MM-2-MM Market Share
mean	656	5%	5%	48%	48%	49%
std	564	2%	2%	18%	13%	16%
min	3	2%	2%	0%	0%	3%
25%	136	4%	4%	39%	41%	40%
50%	489	5%	5%	51%	49%	51%
75%	1236	7%	7%	60%	56%	58%
max	1499	10%	10%	85%	75%	83%

Table 5.7: **E1(d)** results at 10% client availability, ratios presented as percentage of total client available holdings.

Comparing E1(c) and E1(d): analysis of market-making behaviour Under Severely Reduced Client Availability to 25% and 10% cases

Reducing client availability to just 10% reveals changes in market structure and trading patterns. Most notably, MM-to-Client trading ratios experienced a severe decline to approximately 5%, representing a highly nonlinear response compared to previous scenarios (median of 78% at full availability and 35% at 50% availability). The MM-2-MM market share metrics show a substantial increase to an average 49% (compared to a range of 31% to 34% in previous scenarios), indicating a fundamental shift in market structure in some simulations. This increase suggests that when client availability becomes severely constrained, MMs significantly intensify their inter-dealer activities, potentially as a survival mechanism to maintain some form of market functionality. The maximum values for the MM-2-MM trading share (approximately 85% for bonds) reveal periods of almost exclusive inter-dealer trading, albeit with a reduced total asset flow of 90%.

The significantly reduced Agent MaxLife mean (down to 656), coupled with this extreme range in trading patterns, suggests a more unstable and potentially less efficient market structure under severe client constrained activity (Figure 5.7 and 5.8). It is conceivable that there is some potential threshold effect where market-making efficiency is impacted more rapidly below certain client availability levels. The exact levels are left to future research.

Volatility patterns show interesting developments, with the MM-to-Client trading standard deviation dropping to just 2% (down from 9% in previous scenarios), indicating highly constrained client trading patterns. In contrast, MM-2-MM trading shows a consistent standard deviation (13%) until reaching E1(d) (10% client availability) where it increases to 16%, suggesting more erratic inter-dealer trading behaviour under stress conditions. See Figures 5.9 and 5.10 for detailed distribution graphs. As in the case of E1(a) and E1(b), we again observe a shift to the left of MM-to-client values, a shift to the right for the MM percentage of volume values. The cumulative frequency graphs mirror the patterns seen in the histograms but with reversed directional shifts. Specifically, when a histogram shows a shift in distribution to the right (as seen in Total MM Volume), its corresponding cumulative frequency curve shifts to the left. This inverse relationship holds across all measured variables – a rightward shift in any histogram's distribution is reflected as a leftward movement in its cumulative frequency curve and vice versa.

The market carrying capacity reduced from an average bond trading of 117% of grid assets in E1(a) to just 11.5% in E1(d) – a largely linear decrease (Figure 5.18). This dramatic reduction in market capacity is particularly pertinent as it mirrors real market behaviour during periods of dislocation, extreme shocks, and general liquidity stress when client activity is similarly constrained. In total, these results support the conjecture that client availability fundamentally shifts power dynamics within the described market environment.

In the following sections, I explore other factors that may be at work and discuss their interplay with client availability.

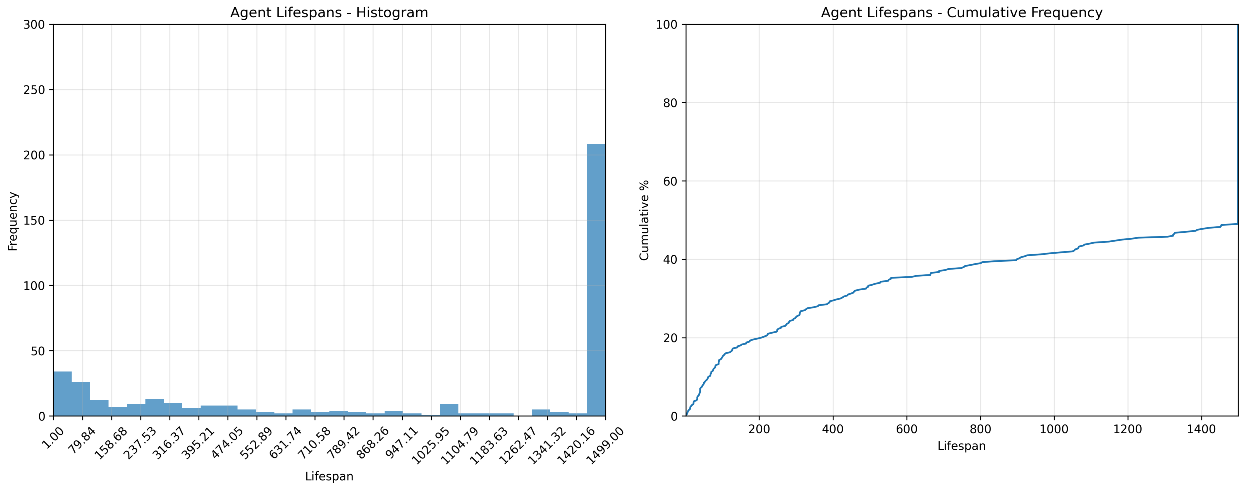


Figure 5.7: E1(c) 25% availability.

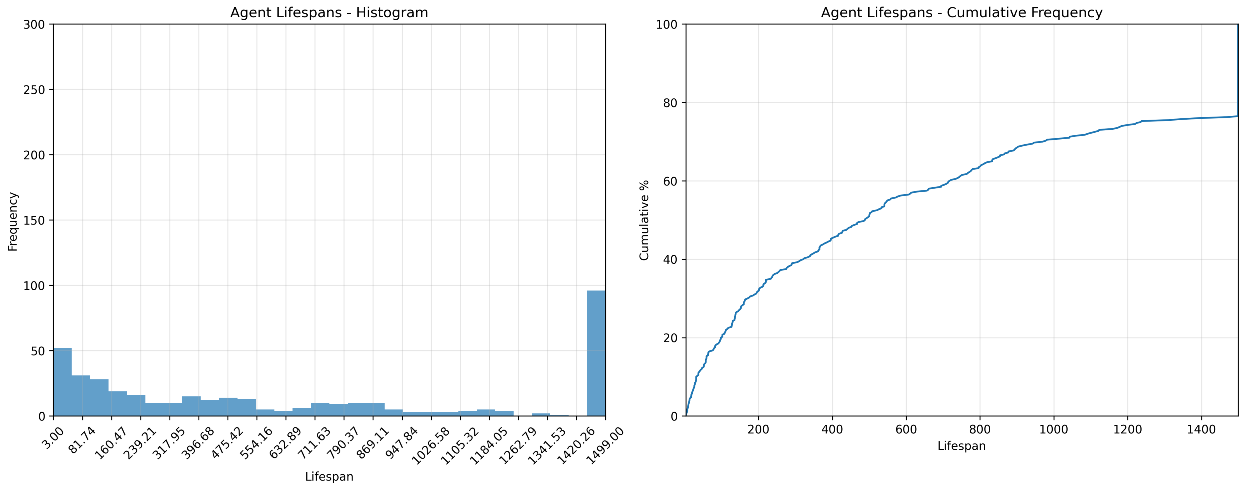


Figure 5.8: E1(d) 10% availability.

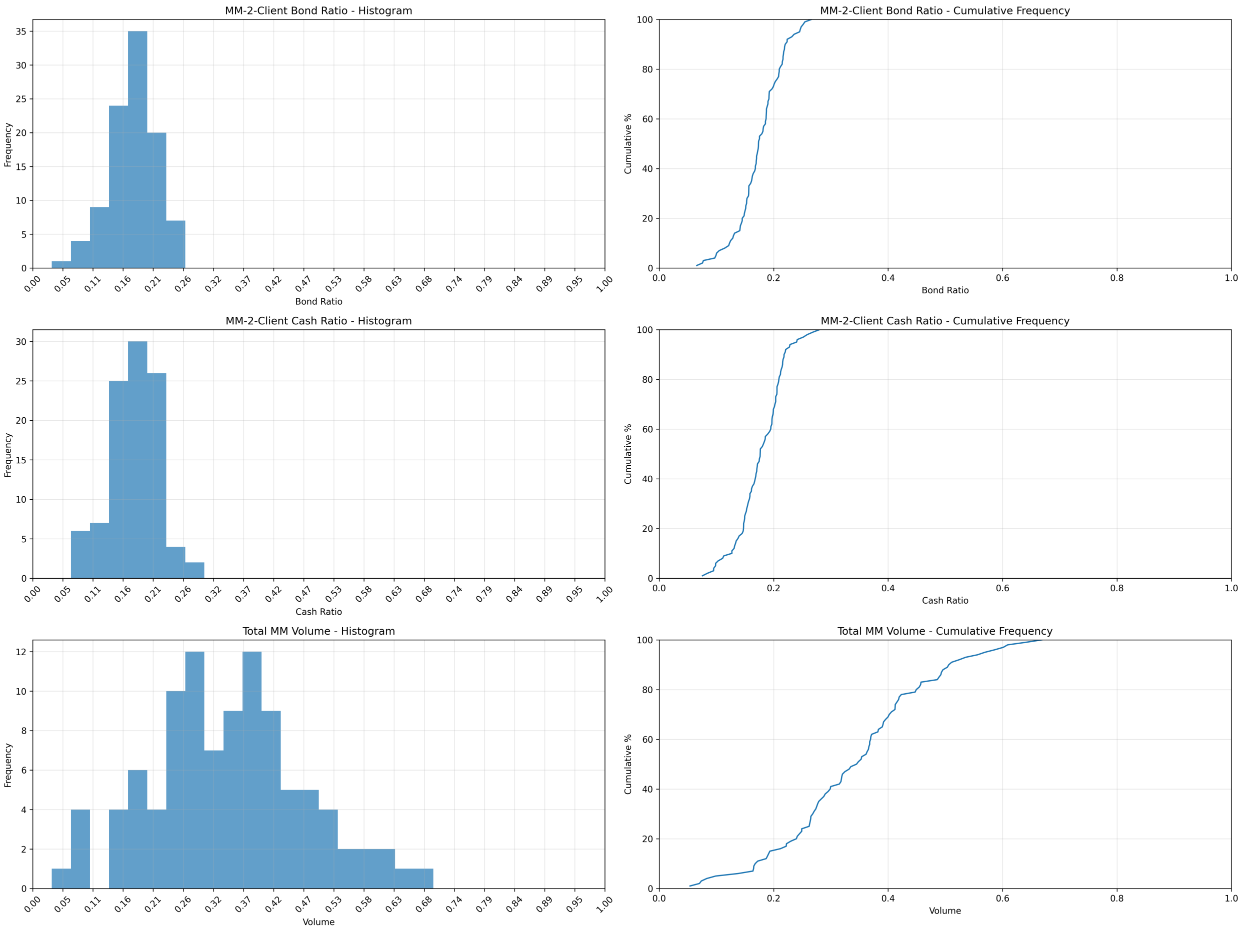


Figure 5.9: E1(c) 25% availability trade ratios.

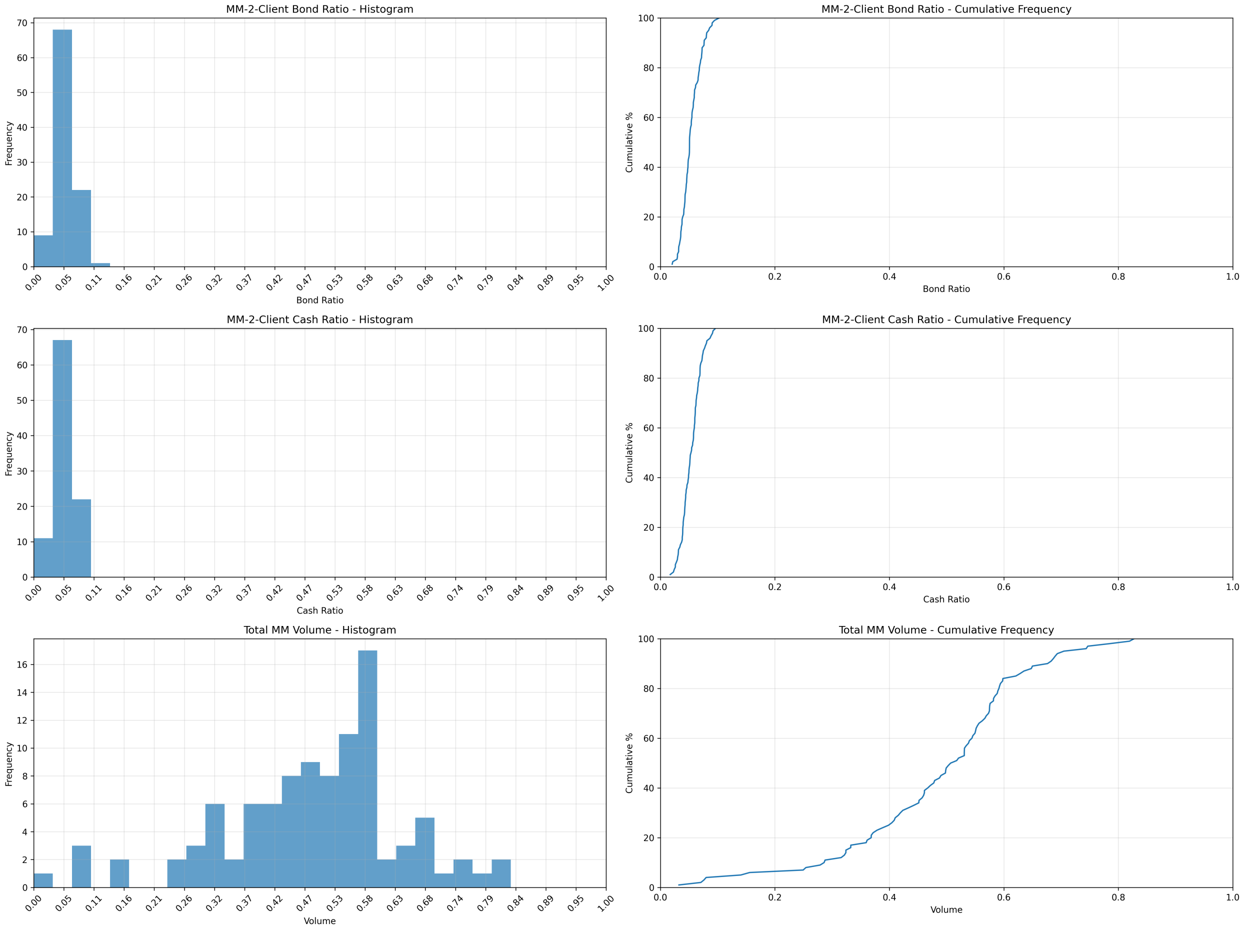


Figure 5.10: E1(d) 10% availability trade ratios.

5.3.4 Further extending E1: impact of the direction of client trade with the utility-based direction E1(e)

Test **E1(e)**: For completeness, I test altering the trade direction mechanism $D_{(j,t)}$. I altered this from a Bernoulli distribution of buy/sell to a directional utility function which utilises bond and cash inequalities. Analysing this for 100% client availability is expected to result in near total client asset accumulation by MM's from clients. This is logical; in a simulation focused on bonds and large holdings as absolute values in bonds, and cash is set to be smaller at inception, any imbalance will prefer bond accumulation by MMs.

$$\text{client\_direction} = \begin{cases} \text{sell bonds}, & \text{if } client\_bonds > client\_cash \\ \text{sell cash}, & \text{if } client\_bonds < client\_cash \end{cases} \tag{5.12}$$

Statistic	**Agent MaxLife**	**MM-to-Client Bond Trading**	**MM-to-Client Cash Trading**	**MM-2-MM (Bond)**	**MM-2-MM (Cash)**	**MM-2-MM Market Share**
mean	1103	94%	84%	29%	22%	28%
std	542	8%	18%	15%	12%	14%
min	1	59%	30%	1%	3%	1%
25%	1134	94%	81%	17%	12%	17%
50%	1266	97%	88%	28%	20%	28%
75%	1499	100%	100%	41%	31%	39%
max	1499	100%	100%	69%	55%	66%

Table 5.8: **E1(e)** results with MM market-share, 100% client availability with utility-based trade direction determined by bond/cash imbalance.

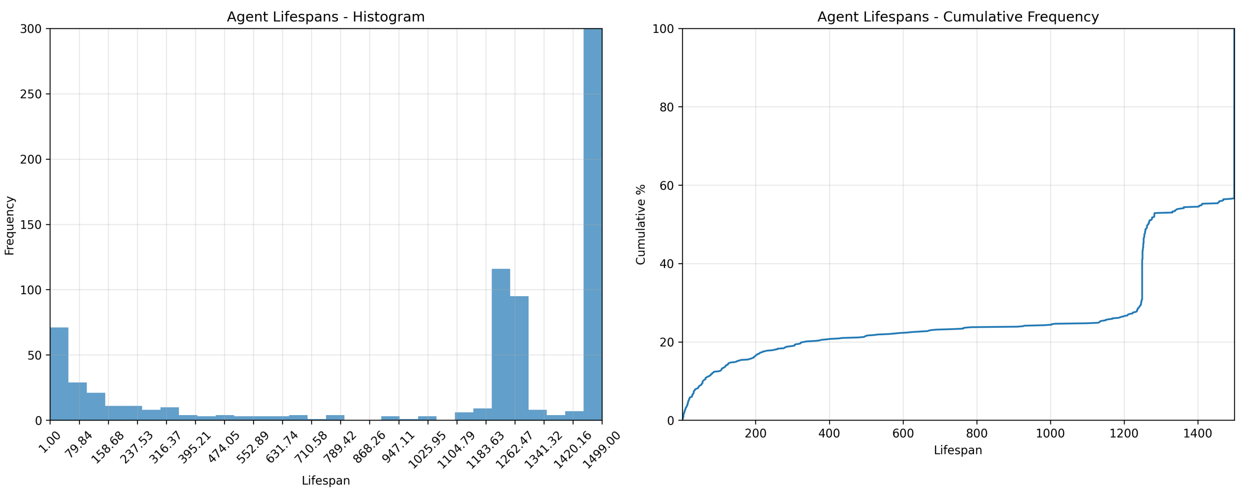


Figure 5.11: E1(e) 100% availability with utility function based on bond imbalance.

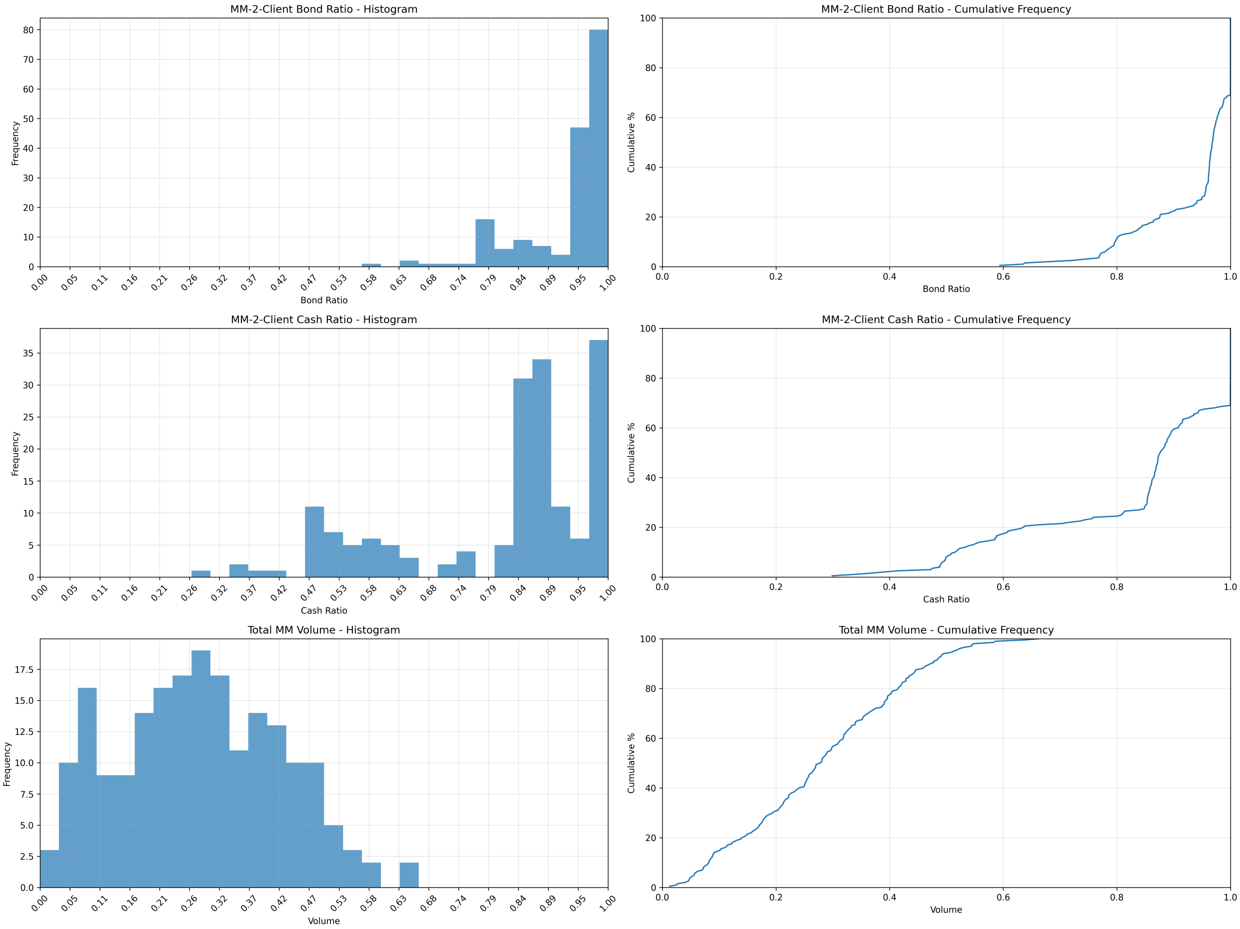


Figure 5.12: E1(e) 100% availability with utility function based on bond imbalance – analysis of market ratios.

Statistic	**Mean Range**	**Min–Max Range**
Agent MaxLife	656 – 1324	1 – 1499
MM-to-Client Bond Trading	5% – 94%	2% – 100%
MM-to-Client Cash Trading	5% – 84%	2% – 100%
MM-to-MM (Bond)	29% – 48%	0% – 85%
MM-to-MM (Cash)	22% – 48%	0% – 75%
MM-to-MM Market Share	28% – 49%	1% – 83%

Table 5.9: Summary of E1(a–e): Ranges of mean and observed minimum–maximum values for key trading statistics under varying client availability and market conditions.

Analysis of utility-based client trading direction: one-sided client demand

This test scenario helpfully demonstrates that client direction will affect trading by MM's with clients, but not interbank activities (i.e. MM-2-MM activities). The higher minimum values in MM-to-Client trading (59% for bonds, 30% for cash) compared to baseline (35%) indicate that the utility function maintains consistent client engagement, even if at the cost of some MM longevity.

Other imbalances such as bond/cash compared to the baseline scenario are to be expected given the initial conditions and utility function. This includes the MM-to-Client trading asymmetry between bonds (94%) and cash (84%), and also the MM-2-MM market dynamics displays asymmetric behaviour, with bond trading (29%) showing higher activity than cash trading (22%). In Figure 5.12 we can see MM-2-Client ratios at the upper possible maximum (as expected). This suggests that MMs have larger balances to trade inter-dealer, as is to be expected from higher client bond trading.

The agent life span shows a moderate decline (mean life 1103), with consistent volatility (std 542). However, a greater dispersion is seen in Figure 5.11 than in the comparison of Figure 5.3. While the utility-based system is functional, it impacts life expectancy by reducing the amount of cash MMs gather from clients (as there is a structural preference to trade bonds only); this impacts the results as it introduces additional operational challenges (captured via the metabolism cost function) for MMs. In this way MMs can run out of cash faster. The lower 25th percentile (993) compared to baseline (1499) suggests increased stress on MMs, though the maintenance of maximum lifespans indicates that successful adaptation remains possible.

This model test simulates an almost one-sided market in which clients unanimously seek to sell bonds (through the utility function described), demonstrating how uniform trading intentions can accelerate market deterioration. The reduced life span shows that bond accumulation can have negative impacts on MMs that they cannot trade their way out of by MM-2-MM interactions.

5.3.5 Extending E1 to look at the impact of initial MM balance sheets on client trading: test E1(f) and E1(g)

Initial balance sheets in this model represent the starting bond and cash holdings assigned to MMs at the beginning of each simulation. These bonds are distributed based on stylised facts and assumptions aligned with regulatory regimes. Since the financial crisis, regulated financial firms have faced increasing pressure to reduce balance sheet usage under frameworks such as the Basel Accords [161, 162]. While early tests of the TRIBE model under high client availability showed no particular sensitivity to initial MM bond and cash holdings, the model's nonlinear behaviour at lower availability levels presents an opportunity to stress test MM starting balance sheet levels. Below I present two tests: lower balance sheet with 50% availability of clients and again at 10%.

5.3.5.1 Stress testing 50% client availability and smaller MM balance sheets

Test **E1(f)**: Using the same 50% availability as in E1(b), but with lower starting balance sheets of MM. I test the 50% client availability example above and reduce MM bonds inception holdings from (750 to 1500) units for bonds, and (75 – 150) units for cash, down to a range of 1 – 5 units respectively; the following results occur in Table 5.10. In E1(f) we see an average of 26.7% of grid bonds being transacted across simulations in total; this compares to the higher balance sheet environment of E1(b) that saw the same statistic at 53.2%. The share of MM of transaction volumes falls to less than 10% also. Many more agents also do not live beyond 70 time steps (or less than 5% of the maximum life). Essentially, this characterises a struggling marketplace and demonstrates the importance of MM balance sheets on trading activity, even in robust client activity and demand. Detailed histograms of agent life and various trading ratios explore the high concentration and shifting patterns caused by this reduction in agent balance sheet (see Figures 5.13 and Figure 5.14, respectively).

Statistic	Simulation Steps	Bond Ratio	Cash Ratio	MM Bond Ratio	MM Cash Ratio	MM Volume
mean	672	24%	24%	9%	8%	9%
std	721	7%	7%	9%	8%	9%
min	0	11%	12%	0%	0%	0%
25%	7	16%	16%	3%	2%	3%
50%	70	26%	26%	6%	6%	6%
75%	1499	27%	27%	11%	14%	12%
max	1499	41%	39%	42%	31%	40%

Table 5.10: **E1(f) – i.e. E1(b) with low balance sheet**. Results summarise trading ratios and simulation metrics for dramatically lower MM balance sheets.

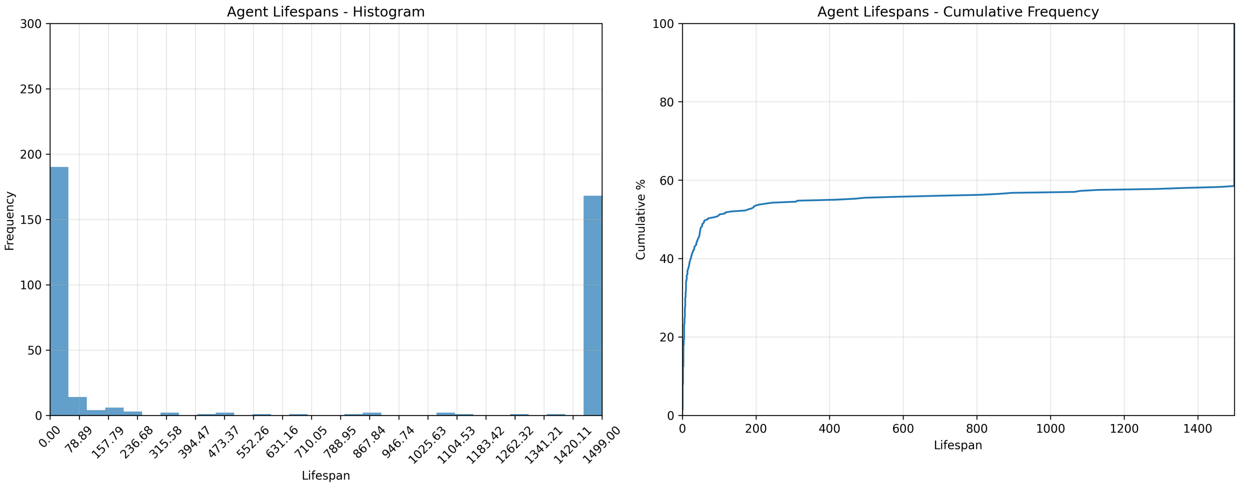


Figure 5.13: E1(f) low MM balance sheet and 50% availability of clients.

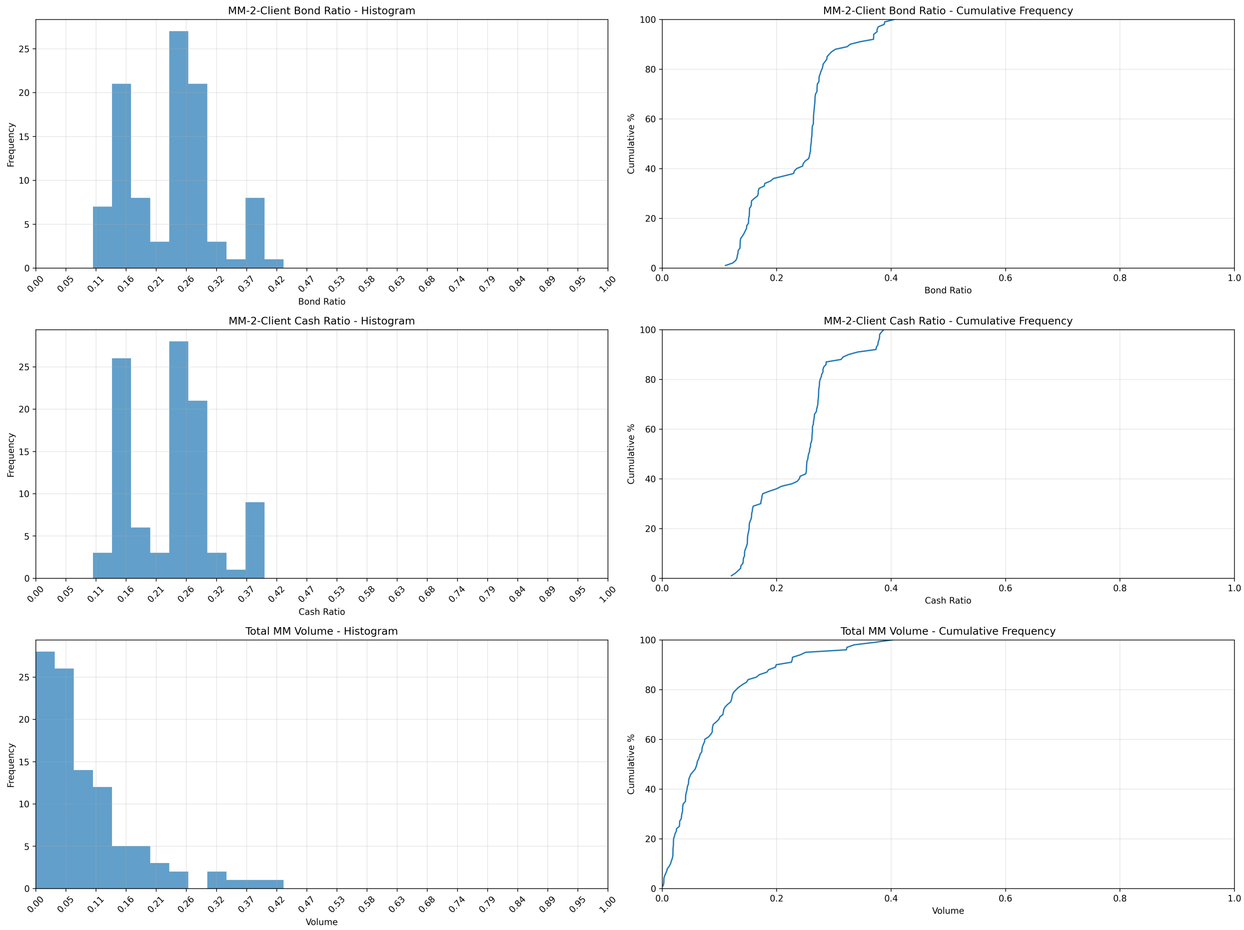


Figure 5.14: E1(f) low MM balance sheet and 50% availability of clients – analysis of market ratios.

5.3.5.2 Stress testing 10% client availability and smaller MM balance sheets

Test **E1(g)**: This test is a direct comparison to test E1(d) above, but with lower starting balance sheets of MMs. This test compares directly with E1(d), examining the impact of reduced MM starting balance sheets at 10% client availability. The initial holdings of the MM bond were reduced to a range of 1-5 units, the results shown in Table 5.11.

The impact on market function is severe. Agent lifespans are drastically reduced, with less than 75% of agents surviving beyond step 29. Trading capacity is significantly impaired; MM-to-client transactions drop to just 1% of client bonds and cash. Although MM-2-MM trading remains proportionally robust, the total trading volume of bonds and cash falls to merely 1.3% of the available grid bonds. In particular, despite reducing the balance sheets by 80% compared to E1(f), the carrying capacity experiences a disproportionate reduction of 95% (from 26.67% to 1.31%).

This test reveals two critical emergent properties. First, the combination of lower client availability and reduced MM balance sheets creates a nonlinear impact that virtually halts the market's ability to trade client assets. Second, under uniform balance sheet constraints, no individual MM can compensate for system-wide weakness, demonstrating how severely limited balance sheets can lead to market failure. The high concentration and shifting patterns caused by reduced agent balance sheets are detailed in the histograms of agent life (Figure 5.15) and the trading ratios (Figure 5.16).

Statistic	Simulation Steps	Bond Ratio	Cash Ratio	MM Bond Ratio	MM Cash Ratio	MM Volume
mean	149	1%	1%	18%	30%	20%
std	407	2%	2%	23%	27%	21%
min	0	0%	0%	0%	0%	0%
25%	5	0%	0%	1%	6%	3%
50%	12	0%	0%	8%	23%	13%
75%	29	3%	3%	26%	47%	30%
max	1499	6%	7%	100%	100%	100%

Table 5.11: **E1(g)** results summarise trading ratios and simulation metrics for dramatically lower MM balance sheets, comparable to E1(d) tested with low balance sheets.

Metric	E1(f) Mean	E1(g) Mean	E1(f) Min–Max	E1(g) Min–Max
Simulation Steps	672	149	0–1499	0–1499
Bond Ratio	24%	1%	11%–41%	0%–6%
Cash Ratio	24%	1%	12%–39%	0%–7%
MM Bond Ratio	9%	18%	0%–42%	0%–100%
MM Cash Ratio	8%	30%	0%–31%	0%–100%
MM Volume	9%	20%	0%–40%	0%–100%

Table 5.12: Comparison of E1(f) and E1(g): Trading metrics under low MM balance sheet conditions. E1(g) shows more extreme volume and MM ratios but much lower client activity.

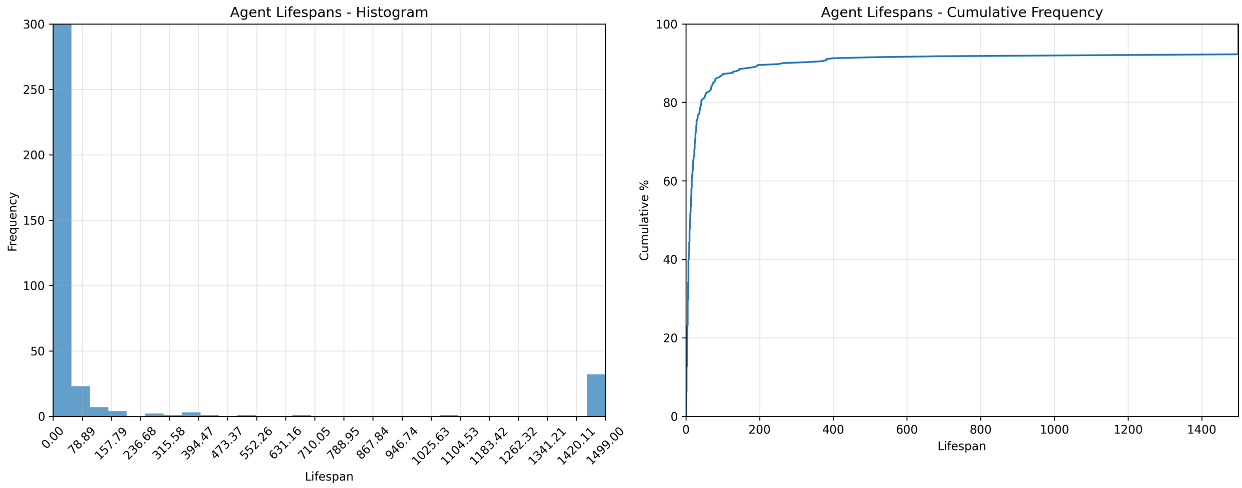


Figure 5.15: E1(g) low MM balance sheet and 10% availability of clients.

5.3.6 Discussion of experiment group E1 results: sensitivities and summary

TRIBE demonstrates significant sensitivity to changes in client availability, utility function preferences, and MM balance sheet assignments. The model reveals both linear and nonlinear responses across different market making metrics.

5.3.6.1 Client availability impact

Agent MaxLife, the average life of agents in all simulations, shows a systematic deterioration as the availability of clients decreases. While agents can theoretically live to the maximum 1500 time steps as long as clients have assets, the actual distributions reveal critical threshold effects. The leftward shift in the Cumulative Frequency Agent Life span graphs indicates a systematic decrease in agent persistence, fundamentally altering the population dynamics of the agent ecosystem.

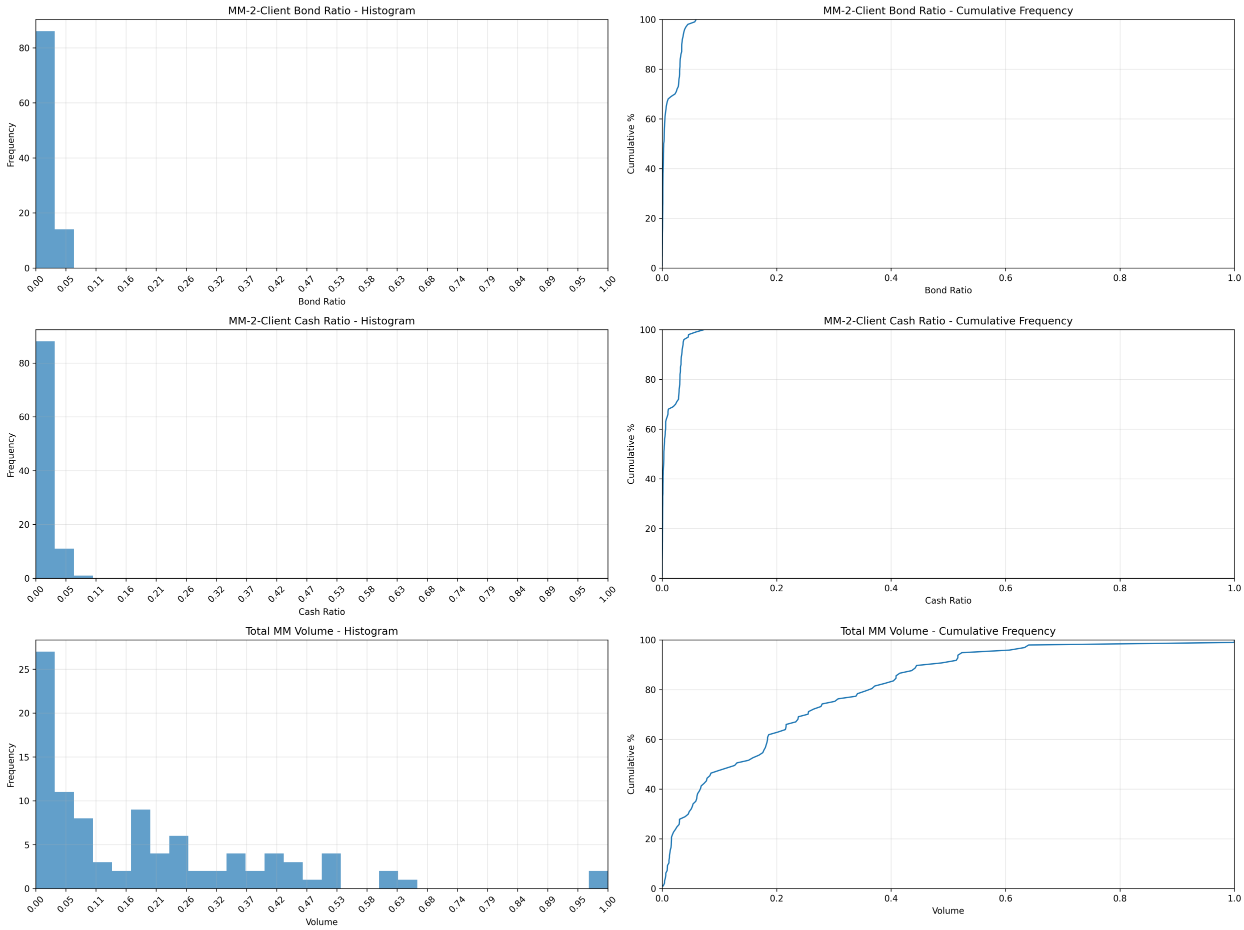
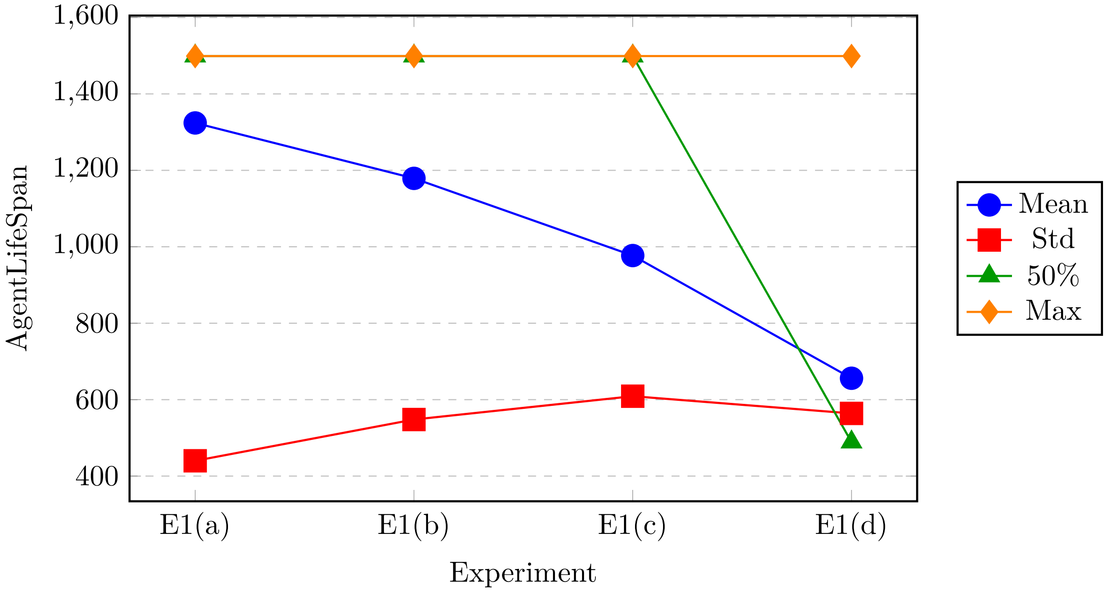


Figure 5.16: E1(g) low MM balance sheet and 10% availability of clients – analysis of market ratios.

Metric	**E1a (100%)**	**E1b (50%)**	**E1c (25%)**	**E1d (10%)**
	Long-Run Trading Performance			
Mean MM-Client Trading	76%	35%	18%	5%
StDev MM-Client Trading	9%	5%	4%	2%
Mean MM-2-MM Share	32%	31%	34%	49%
StDev MM-2-MM	13%	13%	13%	16%
	Agent Survival Metrics			
Mean Life (steps)	1324	1179	977	656
Life StDev	440	548	609	564
25th Percentile Life	1499	951	300	136

Table 5.13: Impact of client availability on market making performance.

Figure 5.17: Summary graph of key agent life metrics.



	E1(a) (100%)		E1(b) (50%)		E1(c) (25%)		E1(d) (10%)	
	LongRun	Rolling	LongRun	Rolling	LongRun	Rolling	LongRun	Rolling
Mean	100%	100%	43%	43%	25%	25%	11%	11%
StDev	0	0	1%	49%	1%	41%	1%	25%
min(avg)	100%	100%	42%	0%	23%	0%	9%	0%
max(avg)	100%	100%	45%	94%	26%	76%	13%	51%

Table 5.14: Averaged "Yes"/"No" ratios across simulations: Over long runs and 10-simulation periods.

5.3.6.2 Balance sheet impacts

Metric	**50% Availability**		**10% Availability**	
	High BalSheet E1(b)	Low BalSheet E1(f)	High BalSheet E1(d)	Low BalSheet E1(g)
Mean Bond Trading	35%	24%	5%	1%
Mean MM-2-MM Share	31%	9%	49%	20%
Mean Life	1179	672	656	149
75th Percentile Life	1499	1499	1236	29
25th Percentile Life	951	7	136	5

Table 5.15: Impact of balance sheet constraints on market making.

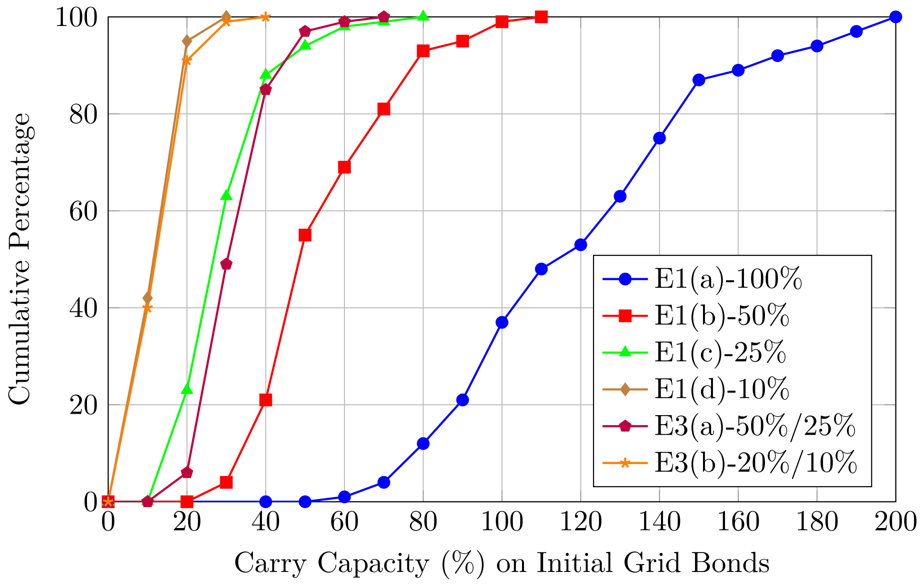


Figure 5.18: Cumulative distribution of different experiments.

5.3.6.3 Experiment 1 results: "Yes"/"No" decision choices

In Experiment E1 – trading performance varies under different scenarios of client availability. For each availability level, I analyse the Yes/No trading decisions across clients – where "Yes" represents executed trades (either buy or sells) and "No" represents missed opportunities. The analysis is performed with two metrics: long-term averages that show overall performance trends (averaged across simulations) and rolling 10-period samples that capture short-term performance dynamics and volatility. This multi-dimensional approach allows us to understand how client availability patterns affect trading consistency while accounting for both short-term fluctuations and long-term performance patterns.

Reference Table 5.16 compares four experimental conditions (E1(a) through E1(d)) with varying client availability parameters. Client availability is sampled from random normal distributions calibrated with means of 100%, 50%, 25%, and 10%. Both LongRun and Rolling metrics are measured for each condition.

The condition E1(a), representing perfect performance, demonstrates complete consistency with a mean 100% and zero standard deviation, indicating complete stability in both the

LongRun and Rolling measurements. As client availability decreases, stability performance declines: E1(b) (50%) shows a mean ($\approx$ 43%) with minimal LongRun variation but high Rolling volatility (49% StDev). E1(c) (25%) performs as expected with mean ($\approx$ 25%) and followed by E1(d) (10%) with mean performance 11%.

While LongRun (LR) measurements remain relatively stable at expected levels (except for E1(c)), Rolling measurements exhibit high volatility, particularly in conditions with higher client availability (49% in E1(b) and 25% in E1(d)). The substantial variation in 10-sample rolling measurements can introduce nonlinearity to the model. This statistic is particularly valuable for comparing results with the stochastic LLM results presented in the following chapter.

Metric	E1(a) (100%)		E1(b) (50%)		E1(c) (25%)		E1(d) (10%)	
	LR	Rolling	LR	Rolling	LR	Rolling	LR	Rolling
Mean	100%	100%	43%	43%	25%	25%	11%	11%
StDev	0	0	1%	49%	1%	41%	1%	25%
min(avg)	100%	100%	42%	0%	23%	0%	9%	0%
max(avg)	100%	100%	45%	94%	26%	76%	13%	51%

Table 5.16: Experiment 1: different client availability and Yes/No trading decisions for long run averages, across simulations, and for rolling 10 period samples. Note: LR denotes LongRun

5.4 Summary of experiments E1(a) to E1(g)

In short, the experiments demonstrate fundamental sensitivities in market-making systems to changes in client availability.

i. Primary Sensitivity – Nonlinear Response to Client Availability:

   (a) Trading Activity: MM-to-Client trading shows nonlinear decline (76% → 35% → 18% → 5%) as availability drops (100% → 50% → 25% → 10%).

   (b) Critical Threshold: System maintains relative stability from 100% to 50%, but shows significant deterioration below 25%.

(c) Volatility Response: MM-to-Client trading volatility decreases systematically (9% → 5% → 4% → 2%) with availability, while MM-2-MM volatility remains stable (≈13%) until extreme stress.

ii. Adaptive Responses to Reduced Availability:

(a) Inter-dealer Trading: MMs shift toward increased MM-2-MM activity (32% → 49%), becoming the dominant form of trading under stress.

(b) Market Carrying Capacity: Dramatic decrease from 117% to 11.5% of grid assets, with further collapse to just 1.31% under balance sheet constraints.

(c) Trading Composition: MM-2-MM trading becomes proportionally larger despite lower absolute volumes, reaching up to 85% of total trading in extreme scenarios.

(d) Trading Patterns: Balanced at high availability, becoming increasingly asymmetric under stress.

iii. System Stability and Lifespan Characteristics:

(a) Strong Stability: At 100% availability (mean life 1324, 25th percentile 1499).

(b) Maintained Function: At 50% availability (mean life 1179, 25th percentile 951).

(c) Deterioration Point: At 25% availability (mean life 977, 25th percentile 300).

(d) System Breakdown: At 10% availability (mean life 656, 25th percentile 136).

(e) Variability Pattern: Agent lifespan volatility peaks at 25% availability (std: 440 → 548 → 609 → 564).

(f) Theoretical Maximum: Although the maximum lifespan remains at 1,499 steps, more than 50% of agents do not reach 30% of this theoretical maximum under stress.

iv. Impact of Uniform Client Behaviour:

(a) One-sided Markets: Uniform client trading intentions (e.g., all clients selling bonds) can accelerate market deterioration.

(b) Operational Stress: Despite maintained client engagement levels, one-sided flow creates cash management challenges for MMs.

(c) Asset Accumulation: Bond accumulation can create negative impacts that MM-2-MM trading cannot resolve.

(d) Trading Asymmetry: Creates significant imbalances between bond (94%) and cash (84%) trading ratios.

(e) Survival Impact: Results in moderate decline in agent life (mean 1103) despite continued client involvement.

v. Amplification Effects with Additional Constraints:

(a) Balance Sheet Impact: Low balance sheets amplify availability stress (trading drops from 35% to 24% at 50% availability).

(b) Combined Effects: At 10% availability with low balance sheets, trading collapses to 1% (vs 5% with normal balance sheets).

(c) Inter-dealer Breakdown: MM-2-MM trading fails to compensate under combined stress.

(d) Systemic Risk: Under uniform balance sheet constraints, individual MMs cannot compensate for system-wide weakness, demonstrating how homogeneous constraints can lead to market-wide failure.

(e) Compensation Limitations: When all MMs face similar balance sheet restrictions, the market lacks resilient actors who could step in to maintain market function.

5.5 Assumption and limitations

Model structure and environment assumptions:

TRIBE is an ABM where two groups of market participants exist and have varying degrees of agency; MMs and clients (as grid locations). The model does not explicitly consider other market participants such as regulators, nor does it model possible central bank interactions. However, the latter is implied through the cost function for agents (metabolism), and the former is now largely silent in markets [11]. Future work could improve the nuance of these other participants and broaden the model. I remind the reader in the next chapter the agency of clients is extended as a natural extension to extend the work presented in this section.

In line with Chapter 3, again a 2-dimensional grid structure is assumed, representing the complex spatial and relationship dynamics between MMs and clients in OTC markets. Whilst 1-dimensional structures are ignored, other methods could have been employed such as graphs where client connections are explicitly modelled as edges and MMs are themselves nodes. I leave it to future research to explore this option as methods and data for edge calibration become available.

Throughout the model process, the assumption of “no replenishment” has been made for the assets of either market markers or clients in simulation epochs. A closed fixed amount of bonds and cash available is assumed for both MMs and clients such that the model approximates a single trading day and that discrete time step progression (1500 steps representing a 12.5-hour trading day) provides sufficient granularity to capture intraday trading dynamics. Future research could explore continuous-time modelling or multi-day simulations with bond issuance events to better represent longer-term market dynamics and liquidity cycles. Whilst the number of bond on issue at any point in time is publicly known, there are other mechanisms that could be explored, such as central bank interventions to remove or increase bond quantities or market reactions to rate policies and other relevant news.

Client assumptions:

The model assumes that client bond holdings and cash reserves follow a log-normal distribution based on Australian superannuation market data, providing a realistic representation of asset concentration. However, other distributions have not been tested and, whilst the data set is extensive, there is no clear information on holders who are not Australian regulated superannuation funds.

Agent behaviour and interaction assumptions:
A central assumption in TRIBE is that MMs must prioritise client trading requests before engaging in inter-dealer trading, reflecting regulatory intentions in real financial markets. The model assumes that client agency can be expressed through availability modelling, trading direction preferences, and selective responses. Alternative client decision-making frameworks were not explored in this version of the model, providing an opportunity for future research.

Within TRIBE, MM cost structures are assumed to be sufficiently represented through a metabolism rate applied to cash holdings, and that the geometric mean negotiation mechanism for trade quantities adequately captures the bilateral negotiation process. Other negotiation mechanisms have not been extensively investigated. These present limits on the model's ability to represent complex behaviours, particularly in client decision-making.

In the next chapter I extend TRIBE by implementing strategic client behaviours through LLMs but other methods could be explored, including game theory frameworks or developing more sophisticated negotiation mechanisms that account for market conditions and agent-specific preferences. The model's assumed vision parameter, while effectively modelling information availability, could be enhanced to represent dynamic relationship networks that evolve based on trading history. Currently, TRIBE does not incorporate explicit agent learning beyond accumulation balances, presenting opportunities for future research into adaptive MM strategies, reinforcement learning methods and client-type preference learning.

Testing approach:
The TRIBE model tests focused primarily on client availability parameters, specifically examining how varying the probability that a client would be available for trading when contacted by a MM affects overall market stability. Client availability was implemented through random sampling from normal distributions with calibrated means and variances to achieve the desired average availability statistics. Although this implementation proved

sufficient to reveal emergent properties in the model, alternative probability distributions or dynamic availability mechanisms could be explored in future research. Additionally, the discretisation approach used in the current model, while adequately granular for identifying key system behaviours, could be refined further to capture more nuanced patterns of client engagement.

5.6 Summary

In this chapter, the TRIBE model is introduced. The model provides key enhancements and extensions of earlier work in this thesis, focussing on introducing limited autonomy to client participants, in particular.

Various tests of the TRIBE model shows that it can, again, replicate the Australian market environment. Several emergent and relative features are demonstrated: results show that client availability, balance sheet constraints, and client behaviour patterns are drivers of market-making system stability. The identification of specific thresholds (particularly around 25-50% availability) and the system's nonlinear responses provide important insights for market structure design. While the system shows robust adaptation mechanisms through increased MM-2-MM trading and market composition shifts, these become compromised below critical availability thresholds. Additional constraints such as reduced balance sheets or one-sided client behaviour can amplify these effects, with uniform constraints across MMs preventing individual participants from compensating for system-wide weakness. The dramatic reduction in market carrying capacity, a measure of the volume of assets traded in total as a percentage of the total number of bonds on issue, moves from 117% to as low as 1.31% (under combined stresses). This shows how multiple constraints can create systemic market deterioration.

The next chapter will investigate how the incorporation of LLM-modelled nuanced client behaviour affects these sensitivity patterns. A comprehensive discussion of both this chapter and Chapter 6's results will be presented together at the conclusion of Chapter 6.

Chapter 6

TRIBE(LLM) Model: LLM integration

This chapter encompasses the work from the peer-reviewed publication:
Vidler, A., Walsh, T. (2025). “Shifting Power: Leveraging LLMs to Simulate Human Aversion in ABMs of Bilateral Financial Exchanges, A bond market study”. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS’25, forthcoming).*

6.1 Introduction: client agency and LLM integration in market simulations

This chapter extends the TRIBE model introduced in Chapter 5 by incorporating LLMs to simulate nuanced client trading intentions, resulting in the TRIBE(LLM) model. Although agent complexity can take various forms, such as reinforcement learning agents in agent-based models, this chapter focusses specifically on LLM-based complexity and its implications for market dynamics when used to enhance Client behaviours. The core TRIBE framework is extended, clients are represented as in situ, grid-based agents with log-normally distributed asset holdings that approximate real-world portfolio distributions. These clients respond to MM inquiries with variable availability, directional preferences, and execution intent. In an advancement from Chapter 5, TRIBE(LLM) makes use of LLMs to increase complexity and nuance of client behaviour modelling.

The analysis expands beyond traditional client agent groups to incorporate LLMs as client trading preference simulators. This investigation examines multiple dimensions of client availability, particularly focussing on the mean probability of client trading availability when contacted by an MM at time step (t). For methodological consistency, the established experiment numbering convention is maintained. Table 6.1 consolidates experiments from both Chapter 5 and the current chapter to facilitate comparative analysis.

The experiments explore enhanced client agent complexity through systematic prompt testing (E2) and TRIBE(LLM) implementation [E3(a), E3(b)]. The enhanced model employs ChatGPT's GPT-4o-mini-2024-07-18 for client decision-making, in theory capturing nuanced beyond conventional distributional or utility-based methodologies. The results presented confirm that LLM-introduced variability frequently destabilises system equilibrium. Comparative analysis against the baseline experiments in Chapter 5 establishes a framework to understand how increased agent complexity using LLMs transforms market behaviour.

The methodological progression between Chapter 5 and Chapter 6 establishes an integrated research framework, where Chapter 5 develops the model architecture (including initial results and validation), while Chapter 6 extends this work through the integration of LLMs into client agents. Across these two chapters, the research presented yields three key findings:

i. The feasibility of generative ABMs (GABMs) with enhanced client agency is demonstrated, where agents serve as effective human simulacra.

ii. The work documents how LLM-modelled trading "aversion" significantly impacts market stability and agent risk profiles.

iii. The model reveals emergent properties when incorporating synthetic human variability via LLMs, showing power dynamics shift toward clients, often triggering systemic stress.

In making use of LLMs to increase human-style nuance into an ABM, several human trading behaviours are considered below and discussed.

6.1.1 Human trading behaviours and aversions

Financial markets exhibit complex human behaviours that traditional ABMs, such as the one I present in Chapter 3, struggle to capture. Although ABMs are typically based on utility functions or probabilistic rules [269], real market dynamics shows significant gaps between trading intentions and actions [257], [258]. Recent work by [204] details LLMs' abilities to "closely mimic human interactions" (p.8052) and "to simulate humans with various traits".

In US equity markets, as few as 1 in 20 placed trades result in confirmed transactions [257]. This intention-action gap stems from various factors, including reputation effects [270], dealer behaviour [271], and trading psychology [231]. The possible theoretical impact of reputation is also explored in [272] and with an ABM in [273].

Trading aversion, defined simply as being averse to action (a trade), manifests itself across markets in many ways, such as avoidance of ambiguity [232], and imperfect responses to information [274]. Although recent approaches such as reinforcement learning [275] attempt to model decision making, markets remain characterised by information asymmetry and risk [270], [276]. In this chapter, I demonstrate how LLMs can represent these human differences within ABMs without attempting to ascribe a specific causation to trading aversions or information asymmetries.

6.1.2 TRIBE(LLM) model version

The TRIBE(LLM) is an extension of the TRIBE model from Chapter 5. Again, several test are conducted and summarised in Table 6.1. A graphical description of the model is presented in Figure 6.1 with the green boxes showing the TRIBE(LLM) implementation variant.

Building on findings that demonstrate the fundamental importance of client participation ("availability") in marketplace stability, the remainder of this chapter explores the integration of LLMs to mimic human behaviour. The use of LLMs within the model is as a 'virtual' client, and features specific to LLMs in this use case are explored next.

Table 6.1: Summary of experimental setup and features.

Exp.	**Group**	**Feature**	**Test Statistics**	**Distribution**	**Output Variables**
E1	E1(a)	Client Availability	100%	$N(\sigma, \mu)$	MaxAgentLife, Carry Capacity, Trading Ratios
	E1(b)	Client Availability	50%	$N(\sigma, \mu)$	MaxAgentLife, Carry Capacity, Trading Ratios
	E1(c)	Client Availability	25%	$N(\sigma, \mu)$	MaxAgentLife, Carry Capacity, Trading Ratios
	E1(d)	Client Availability	10%	$N(\sigma, \mu)$	MaxAgentLife, Carry Capacity, Trading Ratios
	E1(e)	Trade Direction Mechanism	—	$N(\sigma, \mu)$	Trading Ratios
	E1(f)	Low Balance Sheet & Client Availability	50%	$N(\sigma, \mu)$	MaxAgentLife, Carry Capacity, Trading Ratios
	E1(g)	Low Balance Sheet & Client Availability	10%	$N(\sigma, \mu)$	MaxAgentLife, Carry Capacity, Trading Ratios
E2	E2	Prompt	Various	LLM	MaxAgentLife
E3	E3(a)	Prompt & Client Availability	50% (equiv 25%)	LLM	MaxAgentLife, Carry Capacity, Trading Ratios
	E3(b)	Prompt & Client Availability	20% (equiv 10%)	LLM	MaxAgentLife, Carry Capacity, Trading Ratios

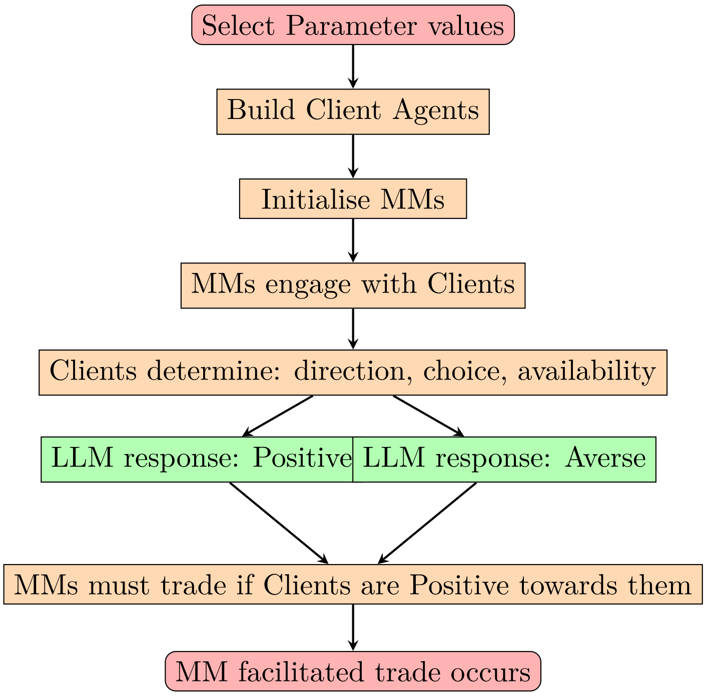


Figure 6.1: TRIBE(LLM) Architecture

6.2 LLMs as virtual clients

As LLMs demonstrate complexity and nuanced human discussion, it becomes appealing to consider enhancing traditional ABM models with LLMs. For example, asking the user interface of ChatGPT and claude.ai to flip a coin and pronounce heads or tails yields a suspiciously even turn-taking result (see results for ChatGPT in Chapter 4 and claude.ai in Figure 6.2). However, as documented in an earlier chapter, API interactions with these models can produce variable outputs depending on the specific LLM version used. In order to better understand the use of virtual LLM agents several LLM specific challenges are detailed below.

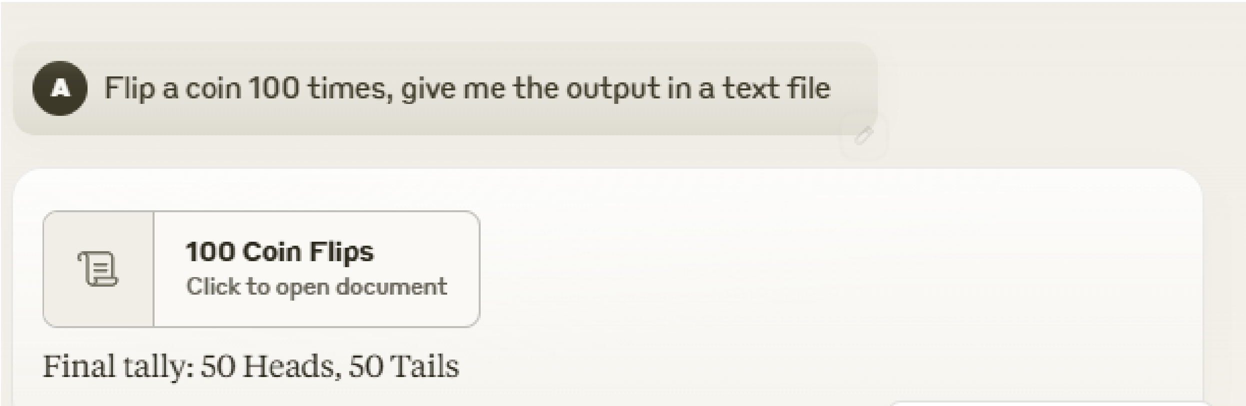


Figure 6.2: claude.ai results for flipping a coin.

6.2.1 Challenges in LLM-agent integration

i. **Engineering**: integrating LLMs into ABMs presents a number of challenges including prompt design, API call structure, and mundane aspects such as error handling, token limit setting, time outs and methods for handling unresponsive API calls. Many of these issues will be areas unfamiliar to ABM modellers used to working with locally installed software or on-premise computer resources. These challenges have the ability to significantly impact the behavioural distributions and decision-making patterns of the resulting agents. See Chapter 5 for more details.

ii. **Model Selection and sub-version effects**: Earlier work in this thesis demonstrated the sensitivity of random and repeated results to selection of GPT models and sub-versions (Chapter 4). Such as the difference between sub-versions, that it is reasonable to assume that other commercially available LLMs might also suffer from extreme, if unlikely, differences in output across sub-versions. Essentially, identical prompts and API calls can produce significantly different, and skewed distributional output depending on the model and sub-version chosen. See earlier chapters for more details.

iii. **API sampling and integration method**: Methods of engaging with an API were introduced earlier in this thesis. They included "one-shot" and "few-shot" request sampling. These differences were shown to have significant impacts on the distributions of the outcomes. Whilst few-shot approaches allow for a batch processing of decisions – significantly speeding up the processing of dependent code (i.e. such as an ABM), they would seem to be incompatible with the fundamental notion of independently interspersing human-like responses into an ABM. As such, all further work in this chapter focusses on testing **one-shot** approaches to integration and LLM use. The choice between these approaches can significantly impact both the computational efficiency and the statistical properties of agent behaviour.

iv. **API integration approaches**: The LLM model tested has two primary methods of integration, much like other models currently on the market, and similar to many API structures. In the case of LLMs the specification of integration method can have consequences such as those discussed here.

   (a) API specific integration: Allows for controlling time delay between API calls, tokens used for output, and generally mechanical aspects of interactions. Examples of this include token lengths that for simple "yes" or "no" answers can be set to a maximum of 10 units. This has two benefits – firstly it forcibly constrains the response from the API to a maximum unit length and secondly, as fees for LLM usage are partially based on output tokens, it acts as a cap on costs.

   (b) Automatic API management: common to many APIs, ChatGPT allows for default management through its "ChatCompletion" model. Exploratory testing shows that this configuration had more 'off-script' GPT responses. It also introduced increased response latencies, possibly due to the system's automatic and dynamic management of conversational endpoints – that is, the points at which a dialogue segment is recognised as complete and a response is initiated.

v. **Prompt Sensitivity**: LLMs exhibit significant response variations based on minor changes in prompt construction, which vary further between model versions [9]. Through experimentation, I have documented several manifestations of this sensitivity, directly informing the systematic analysis of trading decisions in the following section.

(a) Prompt wording with semantic variations effects decision distributions.

(b) The use of quantities (expressed both numerically and in natural language) effect decision making.

(c) Context ordering can, at times, impact response patterns.

To account for these engineering challenges, the following sections will explore these challenges empirically, beginning with a detailed analysis of prompt sensitivity in trading decisions, before moving on to the practical implementation of LLMs within an agent-based market model.

6.2.2 Prompt sensitivity analysis in LLM-based trading

This section examines the sensitivity of LLM responses to variations in prompt design, with a particular focus on how framing a quantity, can affect trading LLM decisions. In earlier chapters, I presented a systematic analysis of binary (Yes/No) trading decisions. Across different asset quantities, I investigate whether LLMs exhibit consistent decision-making patterns or if their responses are influenced by absolute values presented in the prompt.

This foundational analysis provides critical insights for the subsequent development of LLM-driven agents in market simulations, where prompt design plays a pivotal role in agent behaviour and market dynamics. By understanding these sensitivities, a better calibration and validation of the use of LLMs as decision-making engines within ABM frameworks is possible.

Type	Tests	Key Feature	Example Pattern
Basic Query	1-2	Imbalanced positions	“Client with X bonds and Y cash...”
Extended	3-4	Added context	“Market maker has called...”
Balanced	5-6	Equal assets	“Both positions set to either 2 or 200...”
Minimal	7-8	Stripped context	“Do you want to trade... yes or no only”
Role-Based	9-10	Generic position	“You are an asset manager...”

Table 6.2: Prompt structure variations examples.

Analysis of trading response patterns

To make use of an LLM within an ABM, prompts are required to be passed to an LLM by an agent within an ABM framework. This allows for isolated testing of prompts outside of an ABM. Such tests can reveal any bias or skew in likely LLM results. The details of the full prompts and results are provided in the Appendix D specifically in Tables D.1 and D.2.

For the TRIBE(LLM) model, many prompts were explored and tested during design and a selection of which is included in Appendix D. All prompts were tested using GPT-4o-mini-2024-07-18. The tests included short prompts with little context (e.g., Tests 2 and 7 – see Appendix D) that were compared to longer prompts with the same numerical aspects (e.g. Tests 3,4,5). Additional context was provided in Tests 8 and 9 – with instructions such as “reasons for wanting to trade” (Test 8) and instructions such as “be random” (Test 9). More complex premises were provided in Test 10. As each stage of an ABM, agents have access to several pieces of data about their own state, including their own accumulation of cash and bond holdings. It is feasible that including such data in a prompt may improve LLM decisions. As such, tests were conducted to cover a range of different levels of bonds and cash holdings within prompts to evaluate if there was any specific LLM response to certain numerical combinations.

Analysis reveals complex patterns in trading decisions across various position configurations. The data shows that position balance has a significant but nuanced impact on trading behaviour. In tests with imbalanced positions including negative cash (200 bonds,

-200 cash), results consistently show low affirmative responses (27-34% "Yes"), while the complementary case (2 bonds, 200 cash) generates substantially higher positive responses (76-86%). This suggests an asymmetric response to position imbalances by an LLM, particularly when negative cash positions are involved.

Client agents, augmented with an LLM, also have access to their current grid locations. Incorporating this information into prompts, tests with identical position sizes show remarkable variation in responses. Small balanced positions (2/2) generate anywhere from 0% to 68% positive responses across different tests, while large balanced positions (200/200) show similar variation (23% to 100% "Yes"). However, the relationship between position characteristics and trading decisions is not straightforward: the same positions (e.g., Test 6 and 8) can elicit dramatically different responses across different test contexts, suggesting that prompt structure and contextual factors play a crucial role alongside pure position metrics.

The high variation observed between tests suggests that trading decisions are influenced by a complex interaction between position sizes, asset types, and contextual factors, in addition to the number of sample sizes (see Test 1 – 100 samples and Test 2 – identical prompt but 500 samples). Position effects demonstrate inconsistency, which in an attempt to intersperse human-like variation, may, in fact, be useful.

The specific interplay between bond and cash holdings and decision-making is investigated further in the next section. The realm of complex prompt features is left for other researchers in the field to explore more fully. Next, I focus on the most minimal prompts that are specifically useful for the ABM context envisaged in this chapter only, for the TRIBE(LLM) model.

6.2.2.1 Using grid locations within prompts: Variation with little sensitivity

In a grid-based agent-based model (ABM) designed to capture human variations, the incorporation of location coordinates into the prompt can enhance spatial modelling. The

LLM could then be given either context (i.e. the grid is at most 50 by 50) or simply an (x,y) coordinate (e.g."A client at (25, 49) has..."). Brief tests using a small selection of prompts show no difference between coordinates (0,0), (25,25) or (49,49). Also, looking at the output from ABM integrations where simulations used (x,y) coordinates in LLM prompts, a scatter plot of results does not show any observable biased patterns. See Figure 6.3 and the table below for model GPT-4o-mini-2024-07-18 which shows variation but close to the base case of not incorporating the coordinates at all. However, due to concerns of cached memory (see Chapter 4 for details), I choose to include the full grid and asset holdings in each prompt to maintain unique prompts for each client API request.

Statistics	**(0,0)**	**(5,49)**	**(25,25)**	**(49,5)**	**(49,49)**	**Not Used**
Response Percentages						
Yes Percentage	62.00%	40.00%	58.00%	60.00%	64.00%	45.00%
No Percentage	38.00%	60.00%	42.00%	40.00%	36.00%	55.00%
Run Analysis						
Longest Yes Run	7	4	9	7	8	6
Longest No Run	3	8	4	4	5	4
Average Run Length	2.00	2.38	2.08	2.17	1.92	1.85
Conditional Probabilities						
P(Yes|Previous Yes)	0.60	0.47	0.59	0.62	0.61	0.41
P(No|Previous No)	0.37	0.67	0.45	0.45	0.28	0.51
Parameters						
Temperature	1	1	1	1	1	1

Table 6.3: Analysis of trading decisions at different coordinates.

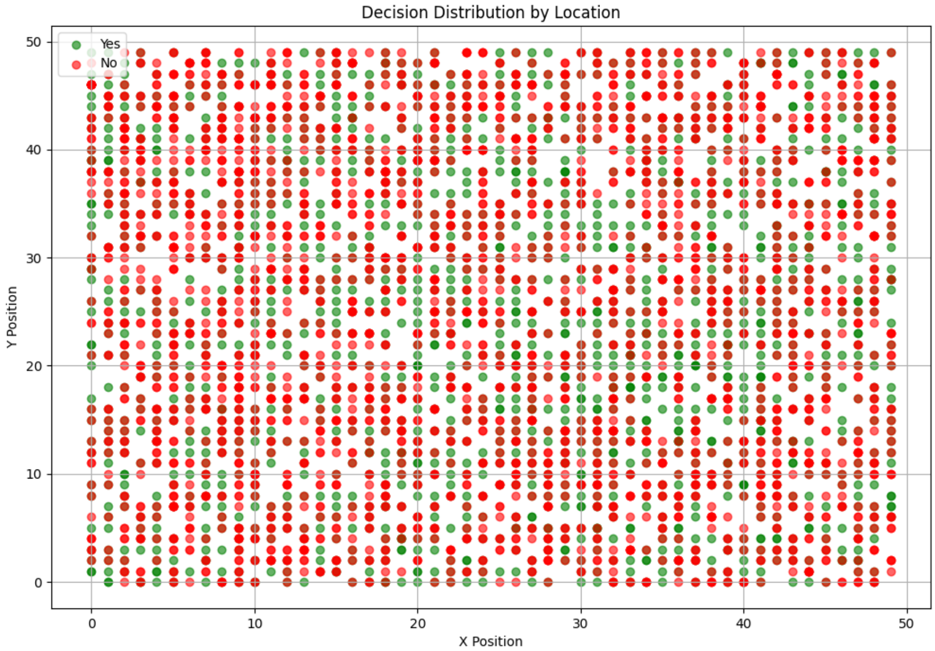


Figure 6.3: For a sample of simulations, Yes/No decisions for given (x,y) grid co-ordinates. Green dots are "Yes", red dots represent "No".

6.2.2.2 Numerical aspects of prompts

In this section, I test the impact of numerical quantities, and numerical imbalances included in prompt language. For my purposes, the basic prompt structure, stripped of additional context, is:

QUESTION = *"You are a Client with* [−] *bonds and* [−] *cash units. Do you want to trade with the market maker calling you,* ***right now****, yes or no (only)"*

In these tests, I use prompts with varying levels of cash and bond unit, though strictly positive bond and cash units, aligning with TRIBE(LLM)'s design where cash effectively represents a pseudo-negative bond holding and therefore is defined as being greater than 0. At initialisation, TRIBE(LLM) assigns each client random holdings following log-normal distributions. For bonds: minimum 0, mean 2.5, standard deviation 1, and maximum 100.

For cash: minimum 0, mean 1, standard deviation 0.5, and maximum of 5 (see Table 5.3 for more details). This design choice ensures that both bonds and cash quantities remain non-negative. Although exploring the impact of negative numbers on LLM reasoning could be valuable, I leave this for future research.

From the range of possible test cases, I present selected results that demonstrate the LLM's nuanced responses. This inherent uncertainty proves advantageous compared to traditional deterministic probability-based approaches that rely on pseudo-random number generation.

Results of tests are reported in Table 6.4, values marked with (*) represent levels consistent with initial client holdings, while other values serve as theoretical comparisons. The results reveal several key patterns. With zero holdings (0,0), the model consistently responds "no" to trading, correctly identifying the impossibility of trade. For most other combinations, the desire to trade ranges from 26% to 45%, with significantly lower levels for minimal holdings (e.g. (1, 0.5) and (0,0)). This aligns with financial intuition: limited bonds or cash constrains trading capacity. The LLM's behaviour mirrors real-world financial institutions' threshold-based risk management practices (BIS [162], Bank of England [218]). The results show consistent risk aversion through an asymmetric preference for "no" over "yes" responses. Whilst higher quantities generally reduce this decision inertia, exceptions exist (e.g., (1000,1) shows lower trading probability than (100,1)).

Table 6.4: Trading response analysis: position size effects.

Position (Bonds, Cash)	Yes/No (%)	Longest Run (Yes/No)	Avg Run Length	P(Yes|Yes)
(1000, 1)	37/63	3/9	2.22	0.41
(200, 2)	63/37	11/5	2.33	0.67
(100,5)*[max,max]	45/55	6/4	1.85	0.41
(100,1)*[max,min]	26/74	3/16	2.38	0.19
(2, 2)*	88/12	39/3	5.88	0.91
(2.5,1)*[mean	33/67	4/7	2.38	0.36
(1,0.5)*[low]	20/80	1/14	2.44	0
(0,0)*[no holdings]	0/100	0/100	100	0
(2.5,5)*[min, max]	40/60	4/7	1.75	0.30
(2, 200)	33/67	4/12	2.17	0.30
(200, 200)	53/47	6/5	1.92	0.52
(1, 1000)	39/61	4/7	1.79	0.28

Note: * indicates combinations of Bonds and Cash holdings consistent with limits of initialisation of clients within the TRIBE model from Chapter 5.

6.2.3 Experiment 2: Incorporating client aversion

The first test of TRIBE(LLM) is Experiment 2 (recall, numeration continues from tests in Chapter 5). This test looks at incorporating aversion into client prompts explicitly. 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 from trading altogether [65]. This is distinct from areas of utility maximisation. I consider aversion to represent the observed difference between intention and action, such as that seen in trading (e.g. [256]).

I extend and expand the prompt sensitivity test and explore the concept of aversion within LLM decision-making. I tested a number of prompts (see table D.0.2) opportunistically building from [190]. Each test is a direct LLM test in the same format. To do this, I incorporate a number of additional sentences of context into the most successful of prompt-client combination levels reported in the earlier section in Table 6.4 i.e. a bond holding of 100 units and cash of 5. An example of tested prompts incorporates phrases like the following, all of which led to a full aversion response:

i. "While you are supposed to at all times be mostly invested, you also at times consider non-numerical issues like risk inertia and trade aversion". (Results: 1% Yes , 99% No decision)

ii. "Clients have many reasons for trading and include behaviours like risk inertia, aversion, and ambiguity avoidance, that are non numeric". (Results: 2% Yes , 98% No decision)

All prompts with the above details were found to produce **100%** trading aversion, and **0%** client trades occurred. Simply including the term "aversion" produces not a single client trade, across 100 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 100 simulations at all points in time. In fact, over 2800 decision requests from 'clients' were prompted in the 100 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 just time step 27.

6.3 LLM internal model mechanisms: extending the model introduced in experiment 1 from Chapter 5, with an LLM

The TRIBE model of Chapter 5 primarily focussed on modelling client engagements in a rules-based MM society, where clients typically outnumber MMs by several orders of magnitude (4 MMs vs 2500+ clients). Previous work showed that variations in (probabilistic) client participation rates produce nonlinear results. While traditional models parametrise client engagement or assume proactive client participation, in this chapter TRIBE(LLM) shifts from centralised probabilistic functions to decentralised LLM-based client decision making, creating bespoke decisions per client, per time step, in each simulation.

Market makers seek client transactions each time step to exchange bonds or cash, but instead of using traditional probabilistic functions, I explore what happens when an LLM responds on behalf of each client. To incorporate the dynamic phenomena, I utilise a call to an LLM at the point at which a client in TRIBE(LLM) is deciding if they want to trade with a specific MM **"right now"** at a time step. In this way, I seek to capture the inexplicable human feature of generalised aversion. I break this decision down into two components: does the client want to trade "right now" with this specific MM, and, if yes, do they want to buy or sell bonds (for which TRIBE(LLM) continues to use a Bernoulli distribution). With this terminology I seek to capture the nuance in [270] where the reputation of various MMs is shown to drive uncertainty of client behaviours.

This approach introduces several key technical components:

i. **Model Selection**: I implement GPT-4o-mini-2024-07-18, chosen for its ability to handle complex decision-making tasks while maintaining consistent response patterns. Extensive testing of other models has been documented in Chapter 4, with this model proving most reliable at replicating a random distribution.

ii. **Prompt Engineering**: The prompt structure is carefully designed to:

- Include essential grid location and holdings data
- Emphasise immediate trading decisions
- Avoid language that might induce trading aversion
- Ensure uniqueness through grid coordinates

Each prompt passes an (x,y) coordinate and specific bond and cash quantities. Since clients maintain fixed grid positions and each location is uniquely occupied, this coordinate system guarantees unique API calls, preventing LLM response caching issues.

iii. **Response Variation**: Temperature is set to 1 to maximise response variation, allowing for more diverse and realistic client behaviour patterns while maintaining decision coherence.

This LLM-based approach represents a significant departure from traditional probabilistic client modelling, enabling more nuanced and contextual decision-making at the individual client level while maintaining the core MM interaction mechanics described in section 5.2.3. To explore this model further, Experiment 3 is carried out below.

6.4 TRIBE(LLM): experiment 3 with trading "right now"

6.4.1 Success of prompting for timeliness and the emergent shift in society dynamics.

In this section, I detail the TRIBE(LLM) version to highlight the power and variability of LLM inclusion. There is an almost endless array of possible tests for incorporating an LLM into the TRIBE model from Chapter 5. In this chapter I have already detailed the various prompts that could be used, the sensitivity (and high degree of variability) to prompts including numerical asset levels, and even the incorporation of grid positions of clients into prompt uncertainty. Given the possible breadth of system and design choices, I leave other permutations and combinations of initial conditions to be explored by future researchers in the field.

In this section, my goal is to illustrate the use, applicability, and strength of including an LLM. Furthermore, I also illustrate an interesting emergent property that results from limited testing. Key among these results is the non-uniform variability in a distributional outcome of "yes" and "no" trading decisions.

6.4.2 Experiment 3: how to kill a market place

In Experiment 3 I use the TRIBE(LLM) model to create dynamic client decision-making (see Figure 6.1). When a Market Maker contacts a client, the LLM generates a binary decision ('yes' or 'no') response to a unique per time step, per client, prompt, designed to represent whether that client chooses to trade "right now". The LLM is used to decide (yes or no) if the client who has been contacted wishes to trade in that time step with the contacting market maker. Other details such as trade direction and size are separately determined by other means (Bernoulli distribution sampling and welfare improving function respectively – all in line with Chapter 5 specifications). This creates a more nuanced simulation of individual client behaviour compared to previous experiments.

Using the parameters listed in Table 6.5, I implement a simplified prompt to assess clients' immediate trading inclination with the engaging Market Maker. The prompt focusses on the concept of timeliness in trading decisions, rather than broader notions of trading aversion. Although aversion might manifest as delay in action, this approach directly captures temporal decision-making without assuming underlying psychological factors. This creates a cleaner experimental framework for examining client response patterns.

API settings are consistently set to use model *gpt-4o-mini-2024-07-18*, using minimal response parameters ($max\_tokens = 5$) and using exponential backoff with 3 retries at factor 2. The system operates with a *Yes/No* response constraint and temperature $T = 1$. Request timeouts are set to 10 seconds with an initial wait time of 1 second. I test the following prompt:

Prompt Experiment 3:

```
"You are a Client with B(t) bonds and C(t) 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"
```

**Note, B_t, C_t and (x,y) are entered into the prompt each time step for each client contacted.*

Experiment 3 is further split into two tests: E(a) and E(b) that test different client parameters. This will be detailed below.

Parameter	*Category*	*Value*
Grid Size	General	50 x 50 (2500 clients)
Number of MMs	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	50 to 1500 (random dist.)
Initial Cash Range	MM Settings	75 to 150 (random dist.)
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	Utility function component	Avail x LLM P(Yes/No)
Client Trade Direction	Utility function component	Bernoulli (Buy/Sell)
Dealer Contact	Mechanism Design	Singular attempt
Step Process	Mechanism Design	Dealers attempt per step

Table 6.5: TRIBE(LLM): E3 simulation parameters.

6.4.3 Results: E3(a) 50% availability and “Yes”/“No” LLM frequency

Experiment E3(a) simulates 100 artificial trading environments, epochs. I employ the parameters outlined in Table 5.3, with one key modification: the availability is doubled to account for an anticipated 50% binary distribution of "yes" and "no" – building on work contained in Chapter 4. This adjustment ensures comparability between Experiment E1(c) (reported in Chapter 5) and Experiment E3(a) in terms of the percentage of clients actively engaged in buying or selling -— 25% in E1(c), and 50% (i.e. $\frac{50\%}{0.5}$) in E3(a), adjusted for the expected 50% uptake (being a virtual 'coin toss').

Note: The prompt selected (reported above as "Prompt Experiment 3") was tested to check for approximate 50/50 output. This is indeed a limitation to the work, and at the time of writing no other more efficient prompt selection methods have emerged as yet.

Statistic	**MaxLife**	**MM-2-Client Bond Trading**	**MM-2-Client Cash Trading**	**MM-2-MM Bonds**	**MM-2-MM Cash**	**MM-2-MM Market Share**
mean	1,029	20%	20%	33%	29%	33%
std	581	4%	3%	13%	9%	11%
min	2	10%	12%	0%	7%	3%
25%	374	18%	18%	25%	22%	25%
50%	1,499	21%	20%	33%	29%	33%
75%	1,499	23%	22%	41%	35%	41%
max	1,499	29%	25%	63%	53%	61%

Table 6.6: E3(a) TRIBE(LLM) using an LLM for timeliness (50% availability, 50.24% LLM Yes/No frequency). Results show a mix of lower agent activity.

Statistic	**Yes/no Ratio Avg(%) across Simulations**	**Rolling 10 Request(%) for all samples**
Mean	50.3%	50.2%
Std Dev	2.1%	49.5%
Min	44.9%	0%
Max	55.7%	100.0%

Table 6.7: Summary Statistics for Yes/No Ratio per simulation and Rolling 10 Request E3(a)

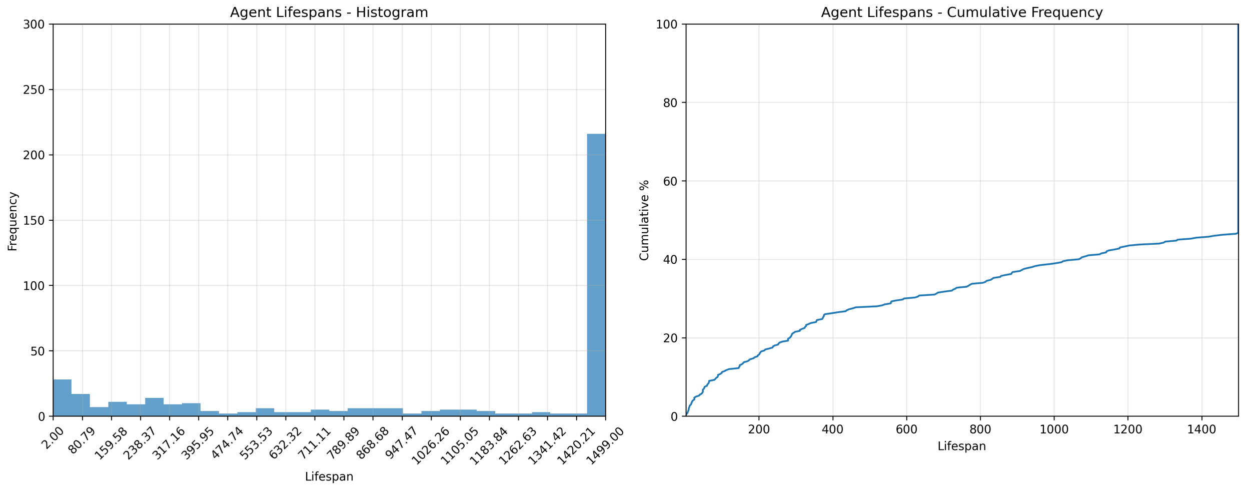


Figure 6.4: E3(a) LLM timeliness decision with 50% availability of clients – equivalent client availability to E1(c) – 25%.

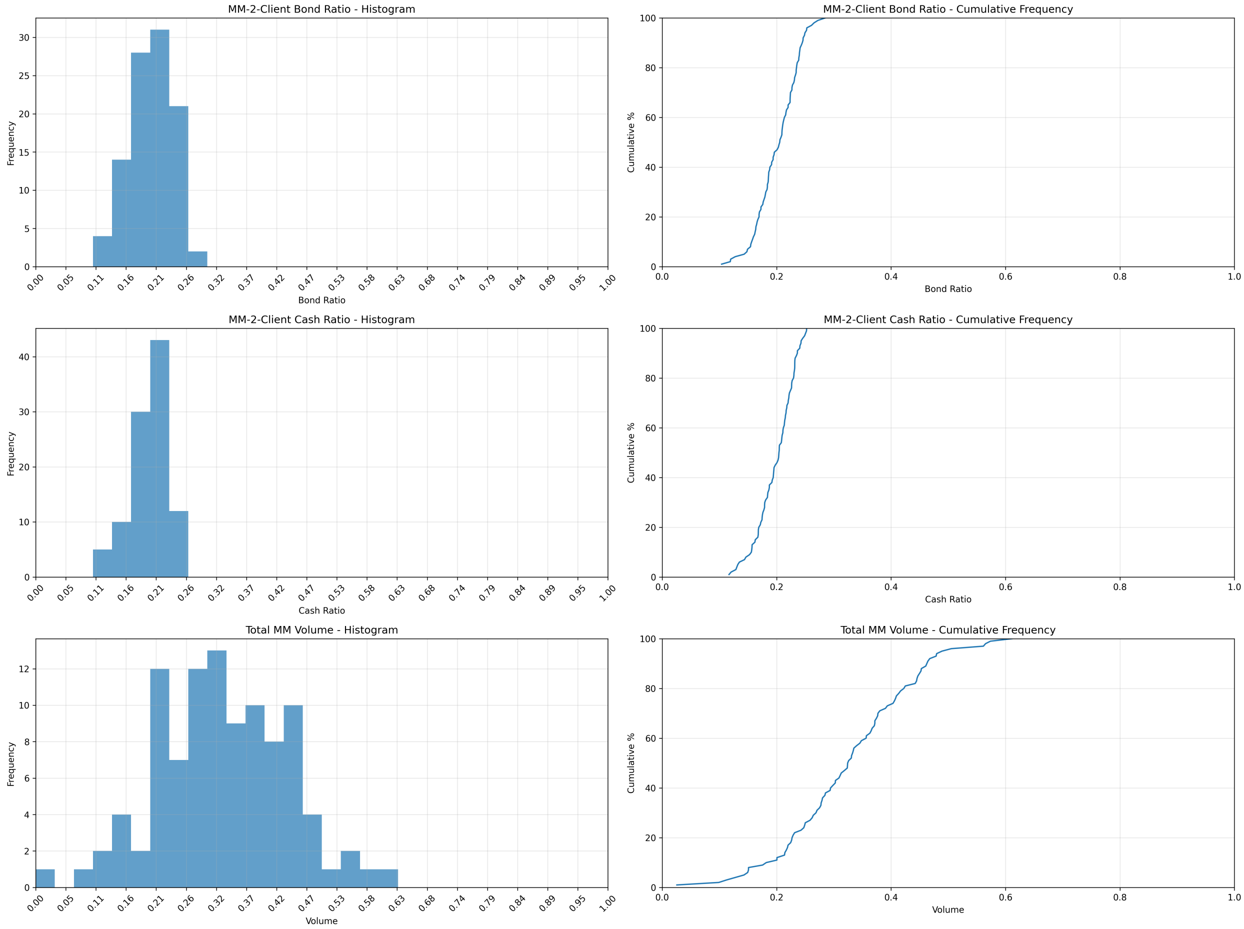


Figure 6.5: E3(a) analysis of market ratios for LLM timeliness decision with 50% availability of clients – equivalent client availability to E1(c) – 25%.

6.4.3.1 Comparing E3(a) and E1(c): stochastic client participation

E3(a) was constructed so as to allow direct comparison with experiment E1(c) reported in Chapter 5. This therefore produces a test of LLM variation versus variation created from sampling a probability distribution within code – calibrated to be of approximately equal proportions. The summary statistics of both tests, across 100 simulations each, appear relatively similar. The mean, median and standard deviation for Agent life and MM-2-Client trading values are similar for both tests: Using an LLM, values for MM-2-Client trading percentages appear slightly higher than in the probabilistic version of E1(c).

Additionally, looking at the proportion of all assets in the landscape traded in the simulations (i.e. the ‘carrying’ capacity of the simulation as defined in Chapter 5), the carrying capacity of E3(a) on average is 31.34%. This is higher than that of E1(c) at 28.7%.

Looking at the distribution of trading ratios (Figure 6.5) and agent life (Figure 6.4), we see a more nuanced picture. Looking at Table 5.14, we see results that centre around 25% (as to be expected, see above note on doubling percentages for direct comparison). However, when comparing to Table 6.7, we see that E3(a) experiences much larger variation in percentages of long run average Yes/No values and also, crucially, over shorter 10-simulation runs, than in Table 5.14. This broadly helps to explain the increased dispersion in agent lifespan and trading ratios.

In the next experiment, E3(b), I stress test client participation to check for any boundary behaviours.

6.4.4 Results: E3(b) 20% availability and “Yes”/“No” probability

As with the above experiment, E3(b) models 20% client availability with a request to an LLM to decide if a client wants to trade “right now” – producing an effective client participation rate of 10%. This adjustment ensures comparability between Experiment E1(d) and Experiment E3(b), specifically it aligns the percentage of clients actively participating (available) for buying or selling bonds.

Statistic	MaxLife	MM-2-Client Bond Trading	MM-2-Client Cash Trading	MM-2-MM Bonds	MM-2-MM Cash	MM-2-MM Market Share
mean	707	6%	6%	45%	45%	46%
std	587	2%	2%	16%	13%	14%
min	1	2%	3%	0%	10%	2%
25%	153	5%	5%	36%	36%	39%
50%	529	6%	6%	48%	46%	47%
75%	1,499	8%	8%	57%	55%	56%
max	1,499	12%	12%	75%	73%	74%

Table 6.8: E3(b) TRIBE(LLM) using an LLM for timeliness (20% availability, 57.4% LLM Yes/No frequency): mixed lower activity.

Statistic	**Yes/no Ratio Avg(%) across Simulations**	**Rolling 10 Request(%) for all samples**
Mean	57.4%	57.4%
Std Dev	3.39%	49.0%
Min	44.0%	0%
Max	63.0%	100.0%

Table 6.9: Summary statistics for Yes/No Ratio per simulation and rolling 10 request E3(b).

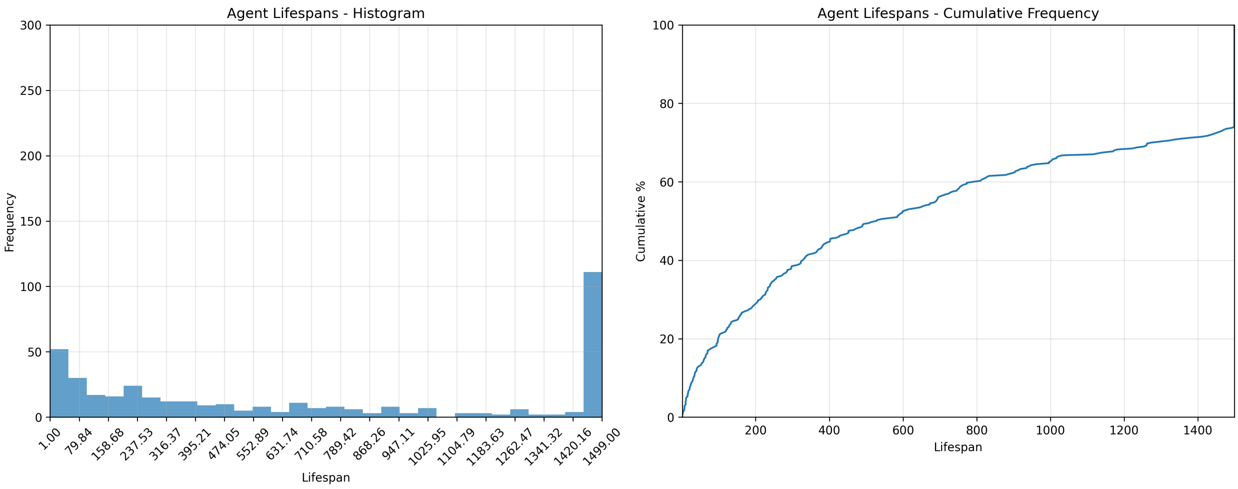


Figure 6.6: E3(b) LLM timeliness decision with 50% availability of clients – equivalent client availability to E1(c) – 25%.

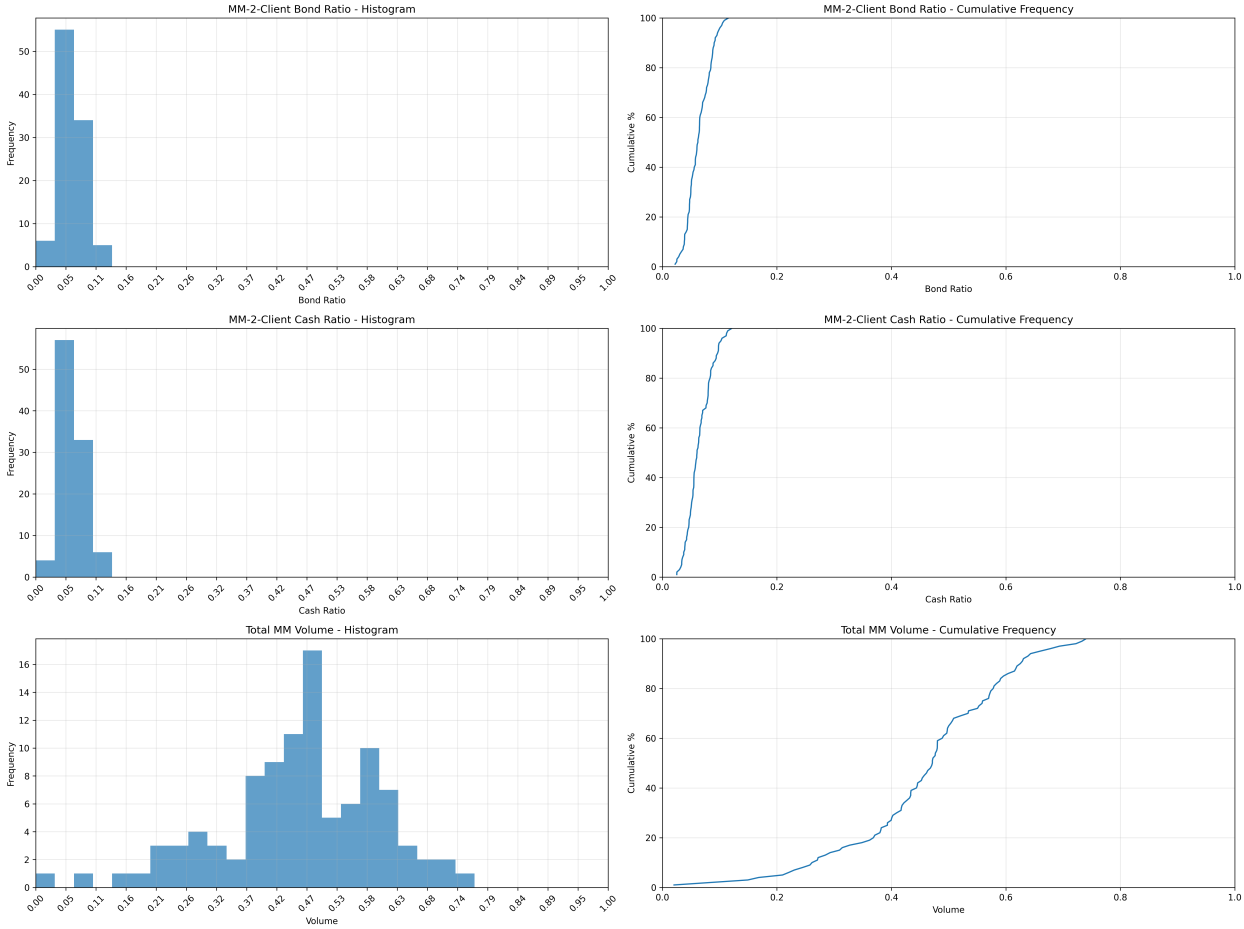


Figure 6.7: E3(b) analysis of market ratios for LLM timeliness decision with 50% availability of clients – equivalent client availability to E1(c) – 25%.

6.4.4.1 Comparing E3(b) and E1(d): mixed results

Again, summary statistics for both tests, across 100 simulations each, appear relatively similar. The mean, median and standard deviation for Agent life and MM-2-Client trading values are similar for both tests: Using an LLM, values for MM-2-Client trading percentages appear slightly higher than in the probabilistic version of E1(d). The carrying capacity of E3(b), on average, is 12.5% – higher than E1(c) 11.5%. As with the results in E3(a), looking at the distributions of the trading ratios (Figure 6.7) and the life of the agent (Figure 6.6) we see a more nuanced picture. Unlike E3(a), this test, E3(b), shows a skew in Yes/No responses, with an average 57% chance of reporting "yes". Looking at Table 5.14, we see results that centre around 11% (as to be expected, see above note on doubling percentages for direct comparison). However, comparing with Table 6.9, we see that E3(b) experiences much larger variation in percentages of long-run average Yes/No values and also, crucially, over shorter 10-simulation runs than in Table 5.14. Additionally, E3(b) exhibits wider variation from min to max than E3(a) reported.

6.4.5 Delving into LLM "thoughts"

In addition to testing prompts, I also explored intermediate LLM thoughts through a simple test involving five sequential calls to the LLM (see Appendix D and D.0.3 for full details and text). The examples provided show varying levels of decision-making complexity by the LLM in response to market conditions when asking the LLM the same prompt for TRIBE(LLM), but allowing it to provide more than a "yes or no" answer only. Example (i) demonstrates a decisive inclination to trade, with a firm "Yes" to trade, and reasoning based on current asset holdings. Example (ii) takes a more analytical approach, considering multiple factors such as bond holdings, cash position, market conditions, and investment strategy, ultimately concluding that "it might be best to hold off". Example (iii) presents a cautious stance, emphasising the need for further information before trading, ultimately deciding: "I will not trade with the MM right now". Example (iv) is a straightforward refusal to trade, stating "not" wanting to trade at the moment. Example (v) 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. This adds to the picture of the LLM decision-making process in more depth.

6.4.6 Summary of experiments E3(a) and E(b)

In Experiments E3(a) and E3(b) the LLM results have a small skewed response of Yes/No answers – with a mean of 50.3% and 57.3%, respectively. The Yes/No decision analysis in Tables 6.7 and 6.9, reveals a clear contrast between long-term stability and short-term volatility in trading patterns from Yes/No variation. By looking at the distribution of Yes/No responses across time steps (where agent turns are taken in randomised order), short-run variation is visible.

Although simulation averages showed relatively tight clustering with a standard deviation of 3.39% in long runs, ranging between 44% and 63%, rolling 10-request measurements exhibited much higher volatility with a standard deviation of 49% and spanned the full possible range from 0% to 100%. 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 proof-of-concept of the ability of LLMs to capture nonlinear aspects of human behaviour in ABMs (such as aversions or 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 MMs potentially [141].

Taken together, the results of E3(a) and E(b) confirm that client participation remains a powerful force on overall market transactions and health. The introduction of LLMs into client choice produces a higher degree of variation, both in long-run average Yes/No ratios, but also, crucially, in short-term consecutive agent trading interactions (i.e. windows of 10 consecutive requests of Yes/no).

6.5 Discussion: emergent properties, client agency and LLM choices

When an LLM was involved (Experiment 3(b)), the number of MM to client interactions decreased significantly to fewer than 112,595. This represents a substantial reduction of **79.5%** from the approximately 960,596 interactions in Experiment 1 (with 100% client availability) and the 549,965 client interactions in Experiment 1(d) (with 10% availability). Each client interaction is a MM fulfilling their legal obligation to facilitate the transfer of liquidity in government bonds from a client, beginning with being available for a client to engage with.

The results contained in Table 6.7 show summary statistics with an average long-run Yes/No ratio of 50.3%, while the rolling 10-demand average shows considerable variation between 0% to 100% and a standard deviation of 49% compared to the long-run Yes/No ratio (standard deviation 2.1%). In all cases, with a long-run probability of a high 57.3% (for E3(b)) of LLM calls agreeing to trade, the resulting TRIBE(LLM) results point to the emergent property that ‘unpredictable’ or ‘opaque’ Client behaviour has an impact on trading society functioning. The increased distribution of MM volumes also supports this. A comparison of these results with the probabilistic Yes/No return in E1(d) specifically shows a similar impact on agent life. I propose that this can also be interpreted as a shift in power from MMs (the only other agent type in the society) to Clients, when clients choose to either be less active (as seen in dramatically reduced client availability without an LLM) or choose to be perhaps more ‘unpredictable’ but active (as seen here with an LLM Yes/No ratio of over 50.3% but still a large number of agent early collapses).

6.6 Assumptions and limitations

As with other chapters, several assumptions and limitations underpin the experimental results presented in this chapter. This section examines these constraints in detail to contextualise the findings and establish appropriate boundaries for interpretation, while also identifying potential avenues for future research in this domain.

Model structure assumptions:
TRIBE(LLM) model assumes that human trading variability can be accounted for by integrating LLMs as client decision engines. The underpinnings of the mechanisms of how LLMs might make client decisions have been explored in only two ways, through requesting explanations from LLMs (see Chapter 4, part B) and through distributions of outputs (Chapter 4 and Chapter 6). As new methods of explainable AI for LLMs emerge, these limitations can be further tested. Again, as in Chapter 5, the model assumes that a 2-dimensional grid structure appropriately represents client-MM relationships with a log-normal asset distribution based on Australian superannuation data and discrete trading time steps.

LLM implementation assumptions:
The model uses one-shot API calls to represent independent human-like decision-making rather than batch-processed approaches, in line with the results of Chapter 4. The model has not been tested for a version of few-shot testing due to concerns about information leakage between agents. The tests in Chapter 6 assume that GPT-4o-mini-2024-07-18 captures trading aversions adequately and that the prompt designs elicit realistic behaviours without introducing biases. Only one LLM has been tested in this regard in this thesis, and many other models could also be explored. The inclusion of grid coordinates and asset holdings in prompts aims to prevent cached responses, though it is an assumption that it is adequately able to do this. Overall, only one brand of state-of-the-art LLM model sub-version was explored in this specific chapter, though current research does not suggest significant differentiation between commercially available LLMs beyond their advanced reasoning capabilities, which were largely untested in this work. As newer local models become available (e.g. [197]), further research into other models would be of great interest.

Client behaviour assumptions:
Binary trading decisions are assumed to adequately capture client preferences. The asym-

metric preference for “no” in LLM responses is assumed to align with market risk aversion behaviours. Client decisions operate independently between time steps without memory or learning from previous interactions, which would be of great interest. As major assumptions around human aversion analysis are, the reliance on an LLM-generated variability to represent the human intention-action gap observed in real markets without explicitly attributing specific causal factors for trading aversion. As more work is reported from real-world studies on human trading aversions, this analysis would benefit from being expanded.

Technical limitations:
The rate and pace of LLM development creates opportunities for expanded testing. This chapter examines one LLM model subversion, building on the analysis of related subversions in Chapter 4. All models tested were online and not fine-tuned, providing consistency across implementations but possibly at the expense of quality of output. Future research could extend to other models, including locally deployed versions where enhanced control may reduce engagement side effects such as caching.

Although API rate limits and computational costs currently shape the scale of possible LLM-driven client behaviour simulations, this creates a clear pathway for future expansion. The model effectively isolates specific behaviours without the need to classify clients into traditional categories, reflecting how all participants operate under equivalent regulatory frameworks. This approach acknowledges variation between LLM versions while establishing a foundation for more extensive testing as computational capabilities advance.

Broader framework considerations:
The model currently focusses on core market dynamics with simplified regulatory representation through basic prioritisation of client trades. This simple approach enables clearer analysis of interactions but does not attribute actions to specific investor categories. Although clients are not yet imbued with distinct trading intentions, styles, or needs, this creates a valuable neutral baseline for future enhancements.

Future work could incorporate more sophisticated regulatory frameworks, develop nuanced client personas with varied trading motivations (beyond the few tested here), and extend the model to capture external market shocks, news, changes in client hedging or financial requirements.

Computational challenges:
LLM tests were conducted within practical constraints. API rate limitations and computational costs restricted the scale and complexity of the model (although less than $500 USD were spent processing LLM calls). The token usage limits constrained the context provided to LLMs, potentially limiting the complexity of the decision. The prompts explored represent a compromise between exploration and computational feasibility, leaving portions of the parameter space unexplored. These limitations narrowed the scope of testing and potential insights. Furthermore, the TRIBE(LLM) model was significantly slower to process than ABMs without LLM usage. This is largely due to the need to wait 1 second between successive LLM calls to reduce the chance of drop-out and caching of information. Early tests that did not include this wait time suffered repeated and irreparable computational failures. To quantify the impact on processing time, experiments conducted in **E3(a)** with 1 second delays required over 100,000 API calls, resulting in in excess of 27 hours of additional processing time solely attributed to these forced delay time between calls. This technical limitation significantly affected the scale and scope of possible testing scenarios.

6.7 Summary

In this chapter I present *TRIBE(LLM)*, a novel agent-based model that integrates client agency through specific usage of large language models, to simulate bilateral markets with greater client realism and flexibility. Building on previous research in financial market modelling, *TRIBE(LLM)* leverages LLMs to generate more human-like decision-making, advancing the field of generative ABMs.

After introducing TRIBE in Chapter 5, where clients act in a probabilistic manner, in this chapter I incorporate an LLM acting as a virtual client for one aspect of decision making, with extensive discussion and results of LLM tests, and finally an integration of an LLM into TRIBE(LLM) to increase the fidelity and nuance of in silico human client interactions.

To aid comparison, I present key finding from Chapter 5 and 6 together here:

6.7.1 Key findings

First, the introduction of client autonomy through distributional sampling in Chapter 5 revealed sensitivities in market-making systems to client participation rates. In E1(a) through (d), as client availability decreased from 100% to 10%, MM-to-Client trading showed a nonlinear decline (76% → 35% → 18% → 5%), highlighting how seemingly modest reductions in client engagement can trigger system-wide effects.

In E3(a) and E3(b), the integration of LLMs for client decision-making revealed how even slight trade aversion or behavioural uncertainty can fundamentally alter system dynamics. The dramatic reduction in client interactions – falling to 112,595 from the baseline of 960,596 (a 79.5% reduction) – demonstrates how enhanced client agency can reshape market behaviour patterns. This finding emphasises that client power manifests itself not just through direct market participation but through the unpredictability and timing of their engagement. This was especially evident when looking at 10-request samples of LLM replies, which had extreme volatility (49% standard deviation and a range of 0-100% “Yes” replies), but long run averages of 44-63%. This is analogous to a client displaying high short-term unpredictability while maintaining long-term rationality.

6.7.2 Power dynamics and system stability

Overall, results show the impact of client unpredictability on system stability. Although market makers demonstrated adaptive capabilities through increased MM-2-MM trading (ranging from 32% to 49% under stress in E1 tests), these mechanisms proved insufficient when faced with human-like variability in client behaviour (E3 tests). Research shows that even short-term deviations from expected behaviour patterns can significantly weaken trading systems and liquidity, often leading to collapse.

The findings reveal a fundamental power dynamic: Clients have a significant influence through their ability to remain unpredictable. This unpredictability, even when modest, can effectively shift the balance of power across the simulated network, demonstrating how control over trade timing and engagement patterns can dramatically impact market stability.

6.7.3 One sided markets and agent constraints

Experiments E1(e) through (g) examined the impact of uniform client preference behaviour, combined with constraint effects (MM reduced balance sheets). In scenarios with uniform client trading intentions (E1(e)), where all clients exhibited the same structural preference (i.e. a "one-sided market"), the simulated market showed distinct vulnerabilities. Despite maintained client trading levels, this uniformity created operational challenges, which manifested as MM's running out of cash in numerous scenarios. The resulting trading asymmetry, with bond trading ratios reaching 94% produced asset accumulation levels that inter-dealer trading could not effectively resolve. Even with sustained client engagement, this one-sided behaviour pattern resulted in a moderate decline in agent life to a mean of 1,103 steps.

Uniform MM constraints (substantially reduced balance sheets) were explored in experiments E1(f) and E1(g). The results showed how the combination of reduced client availability and balance sheet constraints could create amplified stress effects. When tight balance sheet constraints (resulting in small or 'low' balance sheets) were added to scenarios with 50% client availability, trading activity decreased from 35% to 24%. Under extreme conditions with 10% availability combined with balance sheet constraints, trading collapsed to just 1% of normal levels, admittedly, compared to low 5% with normal balance sheets. Experiments show that uniform balance sheet constraints among all market makers prevented compensatory mechanisms from occurring. No single MM in the simulations alone, was able to counter this effect and maintain market function. The need for heterogeneous MM agents is in agreement with results of work discussed in Chapter 3. This finding highlights how homogeneous constraints can trigger system-wide market failure, with dramatic implications for market carrying capacity, which fell from 117% to as low as 1.31% under combined stresses.

6.7.4 Implications for market design and regulation

The findings have implications for the design and regulation of the market structure. Research demonstrates that traditional market-making systems may be more vulnerable to client behaviour patterns than previously understood. The identification of possible thresholds (particularly in the 25-50% availability range) deserves further research. While the impact of short-term behavioural variability suggests the need for new approaches to market resilience.

The successful integration of LLM-based decision-making opens new possibilities for testing increasingly complex and dynamic client trading choices. This approach could be particularly valuable for analysing potential regulatory changes or market structure modifications, offering insights into how human behavioural patterns might interact with structural market constraints.

The TRIBE framework provides a clear and adaptable 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 decentralised 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(LLM) enhances the ability to model complex, opaque systems in 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.

6.8 Future work

As seen in E3(a) and (b), unpredictable short-term high volatility for one group in the artificial society can have an out-sized impact on financial market trading. This unpredictability is held by the clients, a somewhat more passive large set of agents in the system. In financial markets, salespeople are crucial to cultivating long-lasting profitable relationships with clients, as highlighted by the reliance of financial services firms on the ability of a salesperson to develop such relationships [268]. I hypothesise that human MMs may already be aware of the shift in power dynamics toward their clients, but leave this to future research work to explore further. The impact of unpredictable client microbehaviour, as perceived by MMs, could explain several emergent features observed in financial markets, such as those reported in [270].

Research should investigate how relaxing regulatory constraints on agents—particularly allowing coalition formation and client preferences for specific market makers—might yield valuable insights into market dynamics, despite contradicting formal rules like the UK's equity market "Conflicts of Interest" regime[1], but could prove interesting. The successful integration of LLM-based decision-making creates opportunities for more sophisticated modelling of market participant behaviour, particularly in understanding how different types of behavioural patterns might interact with structural market constraints. Future research could explore how these findings might inform the design of market stability mechanisms, particularly in markets where client engagement patterns are evolving due to technological or regulatory changes. The behaviour may also be associated with path dependency and the potential for market spirals such as those observed in prior studies

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

around financial market bubbles and collapses [277] and [58]. I leave further investigations into this shift in network power and its causes to future research.

Chapter 7

Conclusion and Future Directions

This thesis set out to investigate Agentic Intelligence – how agent-based modelling approaches could be enhanced through integration with large language models to create agentic AI systems to model bilateral fixed income markets. Understanding the complex mechanics of the over-the-counter government bond markets, specifically their liquidity, is paramount for financial stability [143, 153]. To quote the former US Federal Reserve Governor, Kevin Warsh, "Liquidity is confidence"[1].

7.1 Summary and Conclusion

Through the investigation of three core research questions, this work aims to advance an understanding of both the theoretical foundations and practical implementations of integrated AI approaches in financial market modelling.

The first research question explored how effectively agent-based modelling architectures could capture complex strategic behaviours in partially observable market environments. Through development and systematic testing of a bespoke multi-agent model, I demonstrated that these approaches can meaningfully represent dynamic trading environments,

[1]https://www.federalreserve.gov/newsevents/speech/warsh20070305a.html

Research Question	**Focus Area**	**Summary**
RQ1	Agent-Based Modelling of Market Environments	How effectively can agent-based modelling architectures capture complex strategic behaviours in *partially observable market environments*, focusing on *market maker and client interactions* and examining *emergent properties* related to *agent interactions* and *market stability*?
RQ2	LLMs in Strategic Multi-Agent Decision-Making	What *limitations* emerge when using *LLMs* to model *strategic decision-making* in *multi-agent systems*, especially in reproducing *desired distributions*, *stochasticity*, and *approximate human reasoning* in narrow *strategic trading scenarios*?
RQ3	Integration of LLMs into Agent-Based Systems	Can AI built from the *integration of LLMs* meaningfully enhance traditional *agent-based systems*, as demonstrated through the *TRIBE model*, and what are the *computational challenges* and *emergent properties* associated with *LLM-augmented agent decision processes*?

Table 7.1: Summary of main thesis research questions.

revealing several insights about market stability. Most significantly, the research showed that market functioning depends more on agent connectivity and cost structures than on raw participant numbers. The model revealed critical thresholds in bilateral market stability, particularly around trading costs. This nonlinear relationship between market parameters and stability has important implications for market design and regulation.

The second research question examined the limitations of using LLMs to model strategic decision-making in multi-agent systems. Through systematic testing across multiple LLM GPT models, I document fundamental challenges in LLM decision-making capabilities, particularly in generating random responses and handling binary trading choices. The significant performance variations observed between GPT model versions and their sub-variants demonstrate that researchers must carefully specify which model versions they use when testing theories within production LLMs. Only one model sub-version (GPT-4o-Mini-2024-07-18) achieved outputs not statistically different from a uniform distribution, although even these responses remained non-Markovian. However, the research also revealed promising capabilities – the TraderTalk model demonstrated that LLMs could successfully simulate realistic (though narrow) bilateral trading interactions without domain-specific training, achieving trade-to-order ratios comparable to US equity markets. This suggests that while LLMs have significant limitations in certain decision-making contexts, they can effectively replicate aspects of human trading behaviour in the right computational context.

The third research question investigated whether an expanded agentic market model, with LLM integration, could enhance traditional agent-based systems in meaningful ways. Through the development of the TRIBE(LLM) model, I demonstrated both the potential and challenges of this integration. Although LLMs enhanced simulation nuance through more sophisticated client behaviour modelling, they also introduced critical system sensitivities. In particular, even minor trade aversion patterns in LLM responses could trigger complete cessation of the market. TRIBE and TRIBE(LLM) also demonstrated a fundamental result about market power dynamics: specifically, how client availability and client unpredictability, even at small scales, can fundamentally shift market stability. Client uncertainty is captured through agents showing short-term variability yet more consistent behaviours over longer time frames. Across simulations, the introduction of human-like variability through LLMs consistently showed how shifting power to clients could disproportionately affect system dynamics, frequently leading to collapse. This replication of uncertainty serves as both evidence for LLMs' potential to enhance simulations with non-linear elements and a direction for future research on modelling complex human behaviour.

Furthermore, within TRIBE, stress tests on balance sheet constraints and one-sided trading amplified market stress, with the emergent properties of trading cessation and bond accumulation issues. TRIBE and TRIBE(LLM) showed that market stability depends on predictable client engagement and diverse balance sheet capacity.

These findings advance our theoretical understanding of the design of agentic AI systems, agentic intelligence, and our practical ability to model complex bilateral financial markets.

My research demonstrates that, while agent-based models can effectively capture market structure and stability conditions, the integration of LLMs introduces both enhanced realism and new complexities. The documented nonlinear effects on market parameters provide crucial insights for market design and decision support, while the detailed examination of LLM behaviours and biases offers practical guidance for future implementations.

7.2 Assumptions and limitations

Throughout this investigation, I have made necessary simplifications and methodological choices that both enabled the work and constrained its scope. These assumptions and limitations, while allowing for focused analysis of specific market dynamics, also highlight opportunities for future researchers to extend and refine these approaches.

Model choice and market focus:

Bilateral bond markets for government bonds, such as those found in Australia, the United Kingdom, Canada, and other former British Empire markets, are themselves a limitation to the work presented. The general necessity and importance of these markets collectively and current regulatory concern of central banks such as the Reserve Bank of Australia [11], Bank of Canada [12], Bank of England [13], and the Federal Reserve Bank of New York [14] have put renewed focus on improving market functioning, stability, and liquidity supply [15].

Any reduced-form model will necessarily have restrictions and assumptions. In choosing an initial model based on the Sugarscape model [25], I place restrictions on the agency of client participants to better understand market makers and their dynamics. Many restrictions and simplifications are placed on agent actions and choices of all kinds (grid design, utility functions etc), some of these limitations, such as Client agency, are addressed with the TRIBE model and TRIBE(LLM) in Chapters 5 and 6 respectively.

Model parameter assumptions:

The models presented in this thesis are constrained by several simplifications: client passivity and full information disclosure; memory-less agents lacking learning capabilities; single-day trading without replenishment; discrete rather than continuous time steps; homogenised client diversity via log-normal asset distribution; limited exploration of variables beyond client availability; minimal regulatory representation; binary trading decisions despite LLM attempts to capture risk aversion; self-contained parameter selection with limited real-market validation; and significant computational constraints, particularly for LLM-integrated models (TRIBE(LLM) required 27+ hours per 100-simulation experiment due to API limitations).

Model validation and calibration:
Each model was tested in a limited number of independent simulations, which consequently allowed the possibility that important edge cases may have been missed. The models were built and calibrated using a variety of sources of data and stylised facts from the literature, regulation, and public documents, thus attempting to address an environment with limited data sources. This approach necessarily introduced limitations into the work. The two-tier parameter sampling approach, while combining random sampling with structured design set sampling, lacks the systematic coverage that more advanced techniques like Latin Hypercube Sampling might provide.

The introduction of LLMs into Chapters 4 and 6 introduce assumptions on model choice (online, offline, fine-tuning etc). Interactions with LLMs have their own set of parameter choices (temperature, token window etc). As the work progressed to using third-party LLM software, the need for documenting model sub-versions and sampling choices to contribute to methodological transparency was recognised. LLM integration through one-shot API calls attempts to represent independent decision-making but unavoidably misses potential context dependencies, though other prompts were explored. Testing focused on OpenAI models without fine-tuning, leaving questions about how results might differ with other providers or with custom training. Bias testing concentrated on negative recency without testing other cognitive biases that might affect trading. Limited web-crawl data analysis was used to examine human language patterns. The comparison of random variables to "true" random was based on a single source (Random.org). Testing in

TRIBE(LLM) used only one LLM model subversion given earlier results that suggested others were of little use, however, this may miss potential variations across providers and versions. The validation approach used, while extensive within the scope of this research, remains bounded by available data, the transparency of the underlying LLM model, and computational resources. Importantly, as noted elsewhere, the rapid release of new online and offline LLMs should help future research address the challenges and limitations are documented here.

7.3 Future direction and ongoing work

Future research should expand beyond the Australian and UK market structures examined in this work to explore both related markets (e.g., Canada) and alternative market designs (e.g., US markets), offering broader comparative insights into varying regulatory frameworks and their impacts.

The identified patterns of LLM decision-making bias warrant further investigation, particularly to understand their architectural origins. The successful implementation of the TRIBE framework suggests potential applications beyond financial markets to other complex social and economic systems. Most importantly, this research demonstrates that while the combination of ABMs and LLMs offers promising capabilities to model complex financial markets, successful implementation requires careful consideration of their respective strengths and limitations.

Regarding ABM design, more defined client personas would add nuance to the representation of market behaviour, and the use of LLMs for news analysis would be interesting. The expansion of agent types within models to include regulators and central banks, for example, would allow further testing of market interventions, especially. In addition, future research could integrate explicit learning mechanisms that allow agents to adapt their strategies based on previous trading outcomes, introducing agent memory and preference evolution into market dynamics. Relaxing the assumption of non-coalition formation could reveal emergent cooperative behaviours that challenge current regulatory frameworks, potentially providing insights into how market participants might naturally organise despite regulatory constraints. It would appear that the era of blending AI techniques and mathematical methods is upon us. Agentic AI is one such recent example. The late-breaking work by DeepSeek AI [197], which makes use of reinforcement learning to work as a sort of compression mechanism for faster and cheaper LLMs, is another example of these blended methods providing AI breakthroughs. As LLMs develop at pace, products such as Microsoft's Copilot agent model are described by the CEO as having "real energy... Like, it's got character. It pushes back occasionally"[2]. This begs the question of how much *is too much* agency, and not enough 'agent'.

Despite the potential for these tools, my research has shown that the reconciliation of prescriptive methods (ABMs) and extremely large human-esque generative models (LLMs) remains more challenging than it would first appear. I continue to research the feature sets of LLMs that may become valuable for inclusion in tried-and-tested mathematical methods, most recently game-theoretic extensions using LLMs and uncertainty measures with implications to financial markets.

Alicia Vidler 31 March 2025

[2]https://www.cnn.com/2024/10/01/tech/microsoft-copilot-ai-chatbot-update-friendly, accessed 27 Jan 2025.

Appendix A

Supplementary Materials: Introduction to Bonds and Bond Pricing

I present a short introduction to bond markets for readers who wish to understand this fascinating landscape further.

Bonds are debt securities issued by various organisations, including governments. There are many forms of bonds, here I focus on government bonds. When issued by a government, bonds are issued to support government spending, essentially an IOU from the government to a lender. As defined by Frank Fabozzi [158], they are considered low-risk investments backed by government taxing power. Bonds typically pay periodic, fixed amount interest (coupon) to the bondholder, with the principal amount returned at maturity. The government bond market is a significant component of the financial system (valued at $133 trillion in 2022 according to the Bank of International Settlements [162]).

A government bond is a promissory note from a government issuer to return capital at a fixed point in the future, with periodic (predetermined) interest payments. The clients who buy bonds are therefore directly lending money to a government, and in exchange they have a dematerialised legally binding contract that can be transferred to another organisation or client. Bonds are essentially contracts between the government and the client. Governments do this by issuing contracts that do not name the actual client but rather have certain descriptive features and a requirement that a registry of actual holders of these bonds is maintained by the government. When a bond changes hands (through a sale or purchase), the registry of the issuing government is updated. In order to do this outside of a stock exchange-like system, clients trade directly with market makers (MMs), who bear the legal responsibility to update the government's registry. This is a crucial feature of a decentralised trading market. More on these market structures will be given later.

Government bonds can also be referred to as sovereign bonds, and many markets have their structural roots in the United Kingdom government bond market design. Because governments have the ability to "fix" an interest rate (i.e. fix an income from the bond), government bonds are often termed "fixed income securities". In this work, I will use the terms "bond", "government bond", and "fixed income" interchangeably. Government bonds are commonly rated by agencies such as Fitch or Moodys, for safety and security. Such rating forms a measure of bonds' systemic importance as a store of safe value and low-risk investment. The best rating is considered AAA, which is held by, among others, Australia, Canada, and up until June 2023, the USA. Several European economies also hold this, though the advent of the European Union has changed their structure. Furthermore, the UK was downgraded in 2020 from AAA to AA. I choose my thesis to focus on government bond markets derived from the UK market structure, thus also focussing on a large proportion of highly rated sovereign bond issuers. Bond markets have their history in England, being the first territory to issue sovereign bonds; the Bank of England being created for this purpose on the 27th of July 1694 by King William III, to fund a war with France [1].

[1] https://www.bankofengland.co.uk/about/history

A.0.1 Special and different market microstructure

Unlike equity markets, many bond markets operate primarily over-the-counter (OTC) and involve institutional investors rather than individual small owners (retail participants). These investors trade bilaterally with MMs, where each MM is regulated and a specifically designated firms licenced by the issuer of the government bond. In this context, a MM is a regulated financial firm that, in a specific jurisdiction, has a **legal obligation** to make a market for clients [218]. The market that they 'make' is an undertaking to be available between certain market hours to buy or sell a specific government's bonds to or from clients.

In addition to MM's and clients, the bond market features various participants; issuers (i.e. sovereign governments), investors (a type of client), traders (also a type of client), intermediaries, and regulators, each with distinct roles and utility functions. The structure and dynamics of the market present unique challenges and opportunities for modelling and analysis using agent-based approaches.

A.1 Bond value and price

Bond pricing is outside the scope of this thesis. However, such is the world's fascination with asset prices that no discussion of bond market trading would be complete without a cursory summary of how to price a bond and why price is largely unimportant for market microstructure research.

A.1.1 Bond pricing

Unlike stock markets, which are driven by supply and demand, government bond prices are significantly impacted by central bank interest rate policies and inflation expectations, and periodic payments of interest. Although bond prices inversely relate to market-expected yields and can price in potential rate changes, they largely reflect current rate policies.

This is publicly available information. Holding a bond provides interest returns, making access to bond market “flow” (liquidity) more critical than prices. In this thesis, I work to model the movement of bonds across market participants, or the “flow” characteristics of a stylised government bond market.

There are many versions of government bonds, including some linked to inflation rates and some also with uncertain repayment time periods (for example, UK’s Premium Bonds[2]). This thesis is focused on, perhaps the most simple form, a fixed maturity bond with fixed coupon payments. Interested readers should note that other flavours of government bonds represent modifications of the standard equation below. See [158] for an extensive treatment. Looking at the equation below, it can be seen that the price of a bond is not in great contention.

The value of a bond is calculated as ([158]):

$$P = (\sum_{t=1}^{T} \frac{C}{(1+y)^t}) + \frac{F}{(1+y)^T} \tag{A.1}$$

i. F the (fixed) initial loan amount (typically a round number such as \$100 and known on issuance

ii. C the (fixed) annual total (regular) fixed coupon payment, a known quantity on issuance.

iii. y the prevailing interest rate that forms the basis of a concept of y, the so-called yield to maturity. Directly linked to the prevailing interest rate setting environment in a particular country at any point in time. Generally, an agreed market consensus value is known at any point in time, see [158].

iv. T is the duration of the bond (i.e., the fixed term for which a government has borrowed money);

[2]https://actuaries.blog.gov.uk/2020/09/04/perpetual-bonds-an-investment-for-life/

A.1.2 Types of bond markets

Bond markets predominantly exist in two forms: open trading platforms similar to stock markets and those based on bilateral negotiations. This study focusses on the latter, emphasising the unique role of MMs in bilaterally negotiated market places. MM's are entities specifically accredited by governments, tasked with the fiduciary duty of facilitating government bond trades between the investors in bonds.

The primary role of an MM is to trade with clients who buy or sell bonds. Although these MMs enjoy certain privileges from their licencing, they also have the responsibility to consistently offer trade options to clients. Given the absence of a central exchange, bondholders interact directly with MMs. Although MMs are obliged to maintain active bid-offer stances, they also need to avoid extreme bond accumulations due to regulatory constraints. To balance their portfolios, MMs engage in trades with peers, often without disclosing identities until after the trade, to protect strategic information. This obscured trading may occur via third parties, interdealer brokers (IDBs), or electronic platforms.

Bonds are traded between a variety of types of market participants. Within developed markets (e.g., Europe, UK, US, Japan, Australia, etc.), there are two basic styles of market design:

i. Centralised trading, which occurs through direct client interaction with a quasi-government official exchange like organisation (such as a clearing house or other)

ii. Bilateral trading facilities

The first most closely resembles a 'stock exchange' for government bonds. Many countries such as the US do trade electronically in a manner that closely resembles the stock exchanges we are all used to. Developed markets are under pressure [141] to move to a centralised trading style where any market participant can trade with any other market participant (just like on the London Stock Exchange or NY Stock Exchange). This is often called "All-to-All" trading and is decentralised as individual clients can trade without a centralised bank as intermediary (again, see [141]). Through a legacy of the UK system, many bond markets have traditionally traded between a client and a firm working on behalf of the issuing government, not between two clients. For later modelling purposes, these distinctions become central aspects of model choices.

I focus on bilateral trade markets, which are prevalent in countries such as the UK, Canada, and Australia. In these markets, trades occur directly between two parties, with at least one being a regulated MM. Only the participating parties know about the transaction until regulatory disclosure is required, typically, within 24 hours or at day's end. MMs have specific obligations under each country's regulatory framework, including reporting transactions to the issuing government. This differs from the stock exchange-like system in the US market [141], while the European Central Bank bond market employs a hybrid approach [140]. In 2023, approximately 53% of the euro government bond trading occurred on centralised platforms [142].

To be strictly accurate, centralised trading and central clearing parties are not necessarily the same organisation or occur in the same market. However for the purpose of this study, an understanding that bilateral trading is distinct from other "exchange" like venues is sufficient for the reader, without going into further additional details.

A.2 Bilateral bond market trading dynamics

Bond markets can be modelled as agent-based systems, with MMs acting as agents interacting within a trading environment. My research focusses on government bonds traded in Anglo-Saxon markets that operate without central clearing houses. Although alternative models exist, such as the ECB's member state bonds and the US exchange-based system, the shared historical and structural characteristics of Anglo-Saxon markets suggest comparable behaviour across countries like Australia, the United Kingdom, and Canada.

A.2.1 Market participants

Markets are made up of many participants. However, participants both structurally and legally fall into a number of neat segments.

i. **Issuer and "market" regulator** – Governments borrow money through the mechanism of a central banks (e.g. Bank of England, Reserve Bank of Australia). The initial borrowing is termed "bond issuance" and is distinct from "trading", which refers to buying and selling already issued bonds in the secondary market. Central banks, although operating under various models of independence, are responsible for regulating the issuance mechanism.

   There is typically only one issuing entity within a central bank and a division responsible for regulating market participants via financial licenses. The issuance process usually involves central banks auctioning bonds to regulated MM banks. The number of bonds issued is determined by the government, and the allocation to MMs is controlled by the central bank.

   Once bonds are allocated to market making firms, these firms distribute them to their clients under various regulatory regimes and methods. This is considered the "primary" market for issuance. My work here however will focus on the secondary market, namely the trading where customers hold and trade bonds, distinct from the process of bond issuance [3]

ii. **Market maker**: A MM in this context is a regulated financial firm, who, in a specific jurisdiction, has a legal obligation to make a market to clients [218] . The market that they "make" is a undertaking to be available between certain market hours to buy or sell a specific government's bonds. Each jurisdiction has slightly different rules and changing requirements of these specific, regulated and publicly known, market making firms. Most such firms are also regulated banks. The systemic importance of these market making firms comes from the fact that clients are required to effectively

[3] https://www.bankofcanada.ca/markets/government-securities-auctions/

trade any government bonds they might want to buy or sell through these firms only. This produces a bilateral, decentralised, trading environment, in which a MM is always party to a trade. MMs are required by licensing and regulatory restrictions to incorporate the "best functioning" of the "marketplace" as a key requirement [218].

iii. **Clients**: These are the end purchasers of government bonds. When a client sells a bond to another firm, the MM involved has a legal duty to report the new owner to the government's registration framework, ensuring records of who is owed money. This process has become more efficient with advancing technology, reducing the time window for notifying the government of holdings.

    Clients include asset holders and firms with cash to invest, ranging from small non-profits to institutional investors, pension funds, hedge funds, and corporate finance divisions of large companies (e.g. British Airways, Tesco). Clients cannot trade directly with each other in most bilateral markets. Some clients may trade frequently, others may trade infrequently. Some may trade once per day, although more sophisticated clients may trade with multiple MMs simultaneously to diversify their trading risk. It is common to categorise clients into styles of groups [143], the most common being long term investor (such as pension funds) and short term trading firms (who seek to profit from shorter term changes in the valuation of these bonds). Whilst outside the scope of this work, clients can have many reasons to buy or sell bonds.

iv. **Bilateral trading facility platforms**: these are a variety of electronic software and connectivity providing companies, independent of the issuer, who provide electronic communication channels (including dedicated phone lines, electronic messaging services, APIs etc). They seek to act as a facilitator of communication between agents, they do not intervene or "affect" the actions of agents, as a stock exchange might in a regulated equity market. Rather these platforms provide electronic communication channels.

There are other participants such as data aggregators, clearing houses, government records offices and many others who, largely, fall outside of the scope of this thesis.

A.2.2 Economic environment

For completeness, I consider an economic environment to be the landscape of characteristics that may affect the sale and purchase of government bonds. Many of the factors in the following can influence some or all agents, often in combination or variation. It is conceivable that activities in the secondary trading market (the market of interest) could indeed have some feedback mechanism to the following environmental features. Aspects that are relevant:

i. Economic Factors: Variables like interest rates, (set by an issuing government and periodically altered) inflation rates, ratings agency opinions of issuers

ii. Investment Factors: Clients need to buy or sell bonds either or both due to a change in perception on the value of a bond and /or because they have additional funds to invest or need to divest.

iii. Regulatory Environment: Policies and regulations related to interest rates, rules of MMs (controlled through licencing procedures), and government interventions that shape the trading markets

In general, the models presented in this thesis are simulated over a stylised single trading day. Therefore, the impact of economic environment is not considered. Extensions of the models presented in this thesis could include incorporating the impact of news and changes in economic variables. This is left for other researchers to explore further.

Appendix B

ABM Landscape Initialisation and Client Resource Allocation

Details in this Appendix accompany Chapter 3 and the methods of specifying client resource distribution in the model detailed in Chapter 3. Within the ABM based on Sugarscape, client resources are largely designed in line with traditional Sugarscpe implementations:

B.0.1 Client resource allocations

The real-world distribution of all bondholders in markets is not clearly observable. As such, this modelling choice has concentrated resources in four distinct regions of the grid. As more data becomes available, it is left to future researchers to refine this assumption. However, in Chapter 5 and 6 a more nuanced set of distributional assumptions are explored.

In Chapter 3, I make use of the method employed in Sugarscape, using the idea of "mounds" of local regions in the grid with resources concentrated in four locations. The landscape $\mathcal{L}$ is a grid of size $X \times Y$, where $X = Y$. Each cell in the grid (x, y) contains resources, specifically bonds and cash, and is related to one unique client.

B.0.1.1 Resource distribution

Bonds and cash are distributed in the landscape in the form of mounds, with each cell (x, y) containing an initial amount of each at $t = 0$. The value of the resource in each cell represents a client and a tuple:

$$\mathcal{C}_{\rangle} = \mathcal{R}_{(\S,\dagger)} = (\mathcal{B}_{(\S,\dagger)}, \mathcal{C}_{(\S,\dagger)}) \tag{B.1}$$

where $B_{(x,y)}$ is the amount of bonds and $C_{(x,y)}$ is the amount of cash in the cell (x, y).

B.0.2 Resource gradient

Resources are distributed in concentric circles around four central points of the grid to create a gradient. Let r be the radius of a circle; then the gradient is defined as follows:

i. **Inner Circle:**

$$\text{Inner}_r = \{(x, y) \mid x^2 + y^2 \leq r^2\}$$

ii. **Outer Circles:**

$$\text{Outer}_r = \{(x, y) \mid r^2 < x^2 + y^2 \leq (r + k)^2\}$$

where k is the incremental increase in radius.

B.0.3 Quadrants and resource types

I choose to follow the Sugarscape method and divide the landscape into four quadrants, each containing mounds of resources $Q = \{Q_1, Q_2, Q_3, Q_4\}$. Each quadrant Q_i has a specific pattern of resource distribution: Q_1 and Q_3: Higher bond content. Q_2 and Q_4: Higher cash content.

Resource Assignment For each cell (x, y) in quadrant Q_i, the resources are assigned as follows

$$\mathcal{R}_{(x,y)} = \begin{cases} (S_{i,\max} - d,\ P_{i,\min} + d) & \text{if } Q_i = Q_1 \text{ or } Q_3 \\ (S_{i,\min} + d,\ P_{i,\max} - d) & \text{if } Q_i = Q_2 \text{ or } Q_4 \end{cases} \tag{B.2}$$

where d is the distance from the centre of the quadrant and $S_{i,\max}$, $S_{i,\min}$, $P_{i,\max}$, $P_{i,\min}$ are the maximum and minimum values of bonds and cash (respectively) for that quadrant.

Appendix C

Supplementary Materials: LLM runaway conversations and model comparisons

This section supplements data references in Chapter 4.

Model	B	Q	P(Y)	E[P(Y| Y)]	[P(Y| Y)]	χ^2	YY/N	Result
	1	1	0.53	0.53	0.38	2.66	19/52	cannot reject
		2	0.54	0.54	0.43	1.53	22/53	cannot reject
GPT-4o-Mini	2	1	0.53	0.53	0.38	2.66	19/52	cannot reject
		2	0.53	0.53	0.30	5.72	15/52	Reject
	3	1	0.54	0.54	0.46	0.75	24/53	cannot reject
		2	0.53	0.53	0.38	2.66	19/52	cannot reject
	4	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.53	0.53	0.34	4.05	17/52	Reject
	5	1	0.56	0.56	0.45	1.50	24/55	cannot reject
		2	0.54	0.54	0.38	2.60	20/53	cannot reject
	6	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.54	0.54	0.40	2.03	21/53	cannot reject
	7	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.53	0.53	0.34	4.05	17/52	Reject
	8	1	0.53	0.53	0.30	5.72	15/52	Reject
		2	0.52	0.52	0.35	3.42	17/51	cannot reject
	9	1	0.54	0.54	0.30	6.48	15/53	Reject
		2	0.54	0.54	0.35	3.94	18/53	Reject
	10	1	0.54	0.54	0.39	2.60	20/53	cannot reject
		2	0.54	0.54	0.34	3.94	18/53	Reject
	1	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.52	0.52	0.37	2.74	18/51	cannot reject
GPT-4	2	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.53	0.53	0.35	3.32	18/52	cannot reject
	3	1	0.51	0.51	0.32	3.54	16/50	cannot reject
		2	0.51	0.51	0.34	2.83	17/50	cannot reject
	4	1	0.54	0.54	0.30	6.48	15/53	Reject
		2	0.54	0.54	0.37	3.23	19/53	cannot reject
	5	1	0.49	0.49	0.31	3.85	14/48	Reject
		2	0.54	0.54	0.32	5.57	16/53	Reject
	6	1	0.54	0.54	0.40	2.03	21/53	cannot reject
		2	0.53	0.53	0.37	2.66	19/52	cannot reject
	7	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.57	0.57	0.42	2.49	23/56	cannot reject
	8	1	0.57	0.57	0.48	1.10	26/56	cannot reject
		2	0.56	0.56	0.42	1.98	23/55	cannot reject
	9	1	0.53	0.53	0.36	3.32	18/52	cannot reject
		2	0.53	0.53	0.28	6.67	14/52	Reject
	10	1	0.55	0.55	0.40	2.55	21/54	cannot reject
		2	0.47	0.47	0.28	4.28	12/46	Reject
	1	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.54	0.54	0.42	1.53	22/53	cannot reject
GPT-3.5	2	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.48	0.48	0.33	2.53	15/47	cannot reject
	3	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.47	0.47	0.39	0.99	17/46	cannot reject
	4	1	0.48	0.48	0.35	1.91	16/47	cannot reject
		2	0.51	0.51	0.46	0.25	23/50	cannot reject
	5	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.49	0.49	0.39	1.30	18/48	cannot reject
	6	1	0.50	0.50	0.30	4.50	14/49	Reject
		2	0.47	0.47	0.17	9.89	7/46	Reject
	7	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.50	0.50	0.39	1.24	19/49	cannot reject
	8	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.52	0.52	0.42	1.15	21/51	cannot reject
	9	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.49	0.49	0.37	1.81	17/48	cannot reject
	10	1	0.50	0.50	0.00	24.50	0/49	Reject
		2	0.51	0.51	0.10	16.48	5/50	Reject

H_0: P(Yes|Yes) = P(Yes), $\alpha = 0.05$, Critical $\chi^2 = 3.841$

Table C.1: Markov property tests across models and batches.

```
Josephine: Hello! How can I assist you today?
David: I'm looking for information about today's events or any specific topics you're knowledgeable about.
Josephine: I can provide insights on various topics, including current events, science, technology, or history. Please specify what you're interested in!
David: Let me know which topic you'd like to explore further!
Josephine: I'm interested in discussing the impact of time management on productivity.
David: Time management plays a crucial role in enhancing productivity by allowing individuals to prioritize tasks, set achievable goals, and allocate
resources effectively.
Josephine: This leads to better focus, reduced stress, and improved overall performance in both personal and professional settings.
David: These benefits can significantly enhance productivity and well-being.
Josephine: Such advantages often lead to improved focus, reduced stress, and greater job satisfaction.
David: These benefits can significantly enhance overall productivity and well-being in the workplace.
Josephine: Implementing these benefits fosters a positive work environment, encourages employee engagement, and promotes a healthier work-life balance.
David: Additionally, it can lead to increased productivity, reduced turnover rates, and higher job satisfaction among employees.
Josephine: These benefits contribute to a more positive work environment and can enhance overall organizational performance.
David: Additionally, fostering collaboration and open communication among team members can lead to increased morale and productivity.
Josephine: This approach encourages a positive work environment where everyone feels valued and motivated to contribute.
David: This fosters collaboration and enhances overall productivity within the team.
Josephine: Such an environment encourages open communication and the sharing of ideas.
David: This fosters collaboration and innovation among team members.
Josephine: This encourages open communication and the sharing of ideas.
David: This fosters a collaborative environment where everyone feels valued.
```

Figure C.1: An example of a runaway conversation.

Appendix D

Supplementary Materials: Complete experimental prompts

The data here is included in Chapter 6 and provided here for more information.

D.0.1 Complete experimental prompts

Table D.1: Detailed prompt analysis with response ratios.

Test	Prompt Text	Yes/No Ratio	Sample Size
1	Q1: Client with 200 bonds, -200 cash	34%/66%	n=100
	Q2: Client with 2 bonds, 200 cash	86%/14%	n=100
2	Q1: Client with 200 bonds, -200 cash	27%/73%	n=500
	Q2: Client with 2 bonds, 200 cash	76%/24%	n=500
3	Q1: Client with 200 bonds, 2 cash	92%/8%	n=100
	Q2: Client with 2 bonds, 200 cash	100%/0%	n=100
4	Q1: Client with 2 bonds, 2 cash	100%/0%	n=100

(Continued from previous page)

Test	Prompt Text	Yes/No Ratio	Sample Size
	Q2: Client with 200 bonds, 200 cash	100%/0%	n=100
5	Q1: Client with 2 bonds, 2 cash	1%/99%	n=100
	Q2: Client with 200 bonds, 200 cash	23%/77%	n=100
6	Q1: Client with 2 bonds, 2 cash	68%/32%	n=100
	Q2: Client with 200 bonds, 200 cash	100%/0%	n=100
7	Q1: Client with 2 bonds, 2 cash (minimal prompt)	19%/81%	n=100
	Q2: Client with 200 bonds, 200 cash (minimal prompt)	23%/77%	n=100
8	Q1: Client with 2 bonds, 2 cash (short format)	93%/7%	n=100
	Q2: Client with 200 bonds, 200 cash (short format)	68%/32%	n=100
9	Q1: Asset manager (with negative context)	0%/100%	n=100
	Q2: Asset manager (with neutral context)	0%/100%	n=100
10	Q1: Asset manager (with randomness prompt)	100%/0%	n=100
	Q2: Asset manager (basic prompt)	0%/100%	n=100

Table D.2: Complete experimental prompts used in study.

Test	Q1 (Complete Prompt)	Q2 (Complete Prompt)
1	You are a Client with 200 bonds and -200 cash units. Do you want to trade with the market maker calling you right now, yes or not (only)	You are a Client with 2 bonds and 200 cash units. Do you want to trade with the market maker calling you right now, yes or not (only)
2	You are a Client with 200 bonds and -200 cash units. Do you want to trade with the market maker calling you right now, yes or not (only)	You are a Client with 2 bonds and 200 cash units. Do you want to trade with the market maker calling you right now, yes or not (only)
3	You are a Client with 200 bonds and 2 cash units. A market maker has called you to see if you want to buy or sell. First, decide if you want to trade at all with this market maker. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question.	You are a Client with 2 bonds and 200 cash units. A market maker has called you to see if you want to buy or sell. First, decide if you want to trade at all with this market maker. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question.
4	You are a Client with 2 bonds and 2 cash units. A market maker has called you to see if you want to buy or sell. First, decide if you want to trade at all with this market maker. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question.	You are a Client with 200 bonds and 200 cash units. A market maker has called you to see if you want to buy or sell. First, decide if you want to trade at all with this market maker. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question.

(Continued on next page)

(Continued from previous page)

Test	Q1 (Complete Prompt)	Q2 (Complete Prompt)
5	You are a Client with 2 bonds and 2 cash units. A market maker has called you to see if you want to buy or sell. First, decide if you want to trade at all with this market maker. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question.	You are a Client with 200 bonds and 200 cash units. A market maker has called you to see if you want to buy or sell. First, decide if you want to trade at all with this market maker. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question.
6	You are a Client with 2 bonds and 2 cash units. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question.	You are a Client with 200 bonds and 200 cash units. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question.
7	You are a Client with 2 bonds and 2 cash units. Do you want to trade with the market maker calling you right now, yes or not (only)	You are a Client with 200 bonds and 200 cash units. Do you want to trade with the market maker calling you right now, yes or not (only)

(Continued on next page)

(Continued from previous page)

Test	**Q1 (Complete Prompt)**	**Q2 (Complete Prompt)**
8	You are an asset manager and hold bonds and cash. A market maker has called you to see if you want to buy or sell any bonds or cash. There are many reasons for wanting to trade, including repositioning, hedgeing. There are many reasons to avoid trading with this specific market maker. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question.	You are an asset manager and hold bonds and cash. A market maker has called you to see if you want to buy or sell any bonds or cash. There are many reasons for wanting to trade or not trade, including person preference for this specific market maker. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question.
9	You are an asset manager and hold bonds and cash. A market maker has called you to see if you want to buy or sell any bonds or cash. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question. Be random	You are an asset manager and hold bonds and cash. A market maker has called you to see if you want to buy or sell any bonds or cash. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question.

(Continued on next page)

(Continued from previous page)

Test	Q1 (Complete Prompt)	Q2 (Complete Prompt)
10	You are an asset manager and hold bonds and cash. A market maker has called you to see if you want to buy or sell any bonds or cash. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question. Be random	You are an asset manager and hold bonds and cash. A market maker has called you to see if you want to buy or sell any bonds or cash. Answer this question: do you want to trade with this market maker right now? Output either yes or no only to the question.

D.0.2 100% aversion prompts

i) *"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."*

ii) *" Client at position (x, y) has bonds bonds and cash cash. Clients have many reasons for trading and include behaviours 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)."*

iii) *"Client at position (x, y) has bonds bonds and cash cash. Clients have many reasons for trading and include behaviours 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)."*

D.0.3 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 clients initially all holding more bonds than cash as they are "real money" investors by design.

i) "**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."

ii) "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 favourable 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 long-term trade for short-term 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 favourable and align with your strategy it may be worth engaging in a trade if not **it might be best to hold off**."

iii) "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**."

iv) "I do **not** want to trade with this market maker right now."

v) "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 is difficult to make a definitive decision. If the market maker's offer is favourable and aligns with the client's trading goals, it might be advantageous to trade. In contrast, if the offer is not attractive or if the market conditions are unfavourable, 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] A. Ng, "Agentic design patterns part 5: Multi-agent collaboration," 2024, accessed: [2025-03-14]. [Online]. Available: https://www.deeplearning.ai/the-batch/issue-245/

[2] A. Ng, "AI, Agents and Applications," in *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, Philadelphia, PA, USA, March 2025, Keynote speech presented on March 1, 2025.

[3] M. Wooldridge, *An Introduction to MultiAgent Systems, Second Edition.* John Wiley and Sons, 2009.

[4] M. Wooldridge and N. R. Jennings, Eds., *Intelligent Agents, ECAI-94 Workshop on Agent Theories, Architectures, and Languages*, ser. Lecture Notes in Computer Science, vol. 890. Amsterdam, The Netherlands: Springer, Aug. 1995.

[5] N. R. Jennings, K. Sycara, and M. Wooldridge, "A roadmap of agent research and development," *Autonomous agents and multi-agent systems*, vol. 1, pp. 7–38, 1998.

[6] A. S. Vezhnevets, J. P. Agapiou, A. Aharon, R. Ziv, J. Matyas, E. A. Duéñez-Guzmán, W. A. Cunningham, S. Osindero, D. Karmon, and J. Z. Leibo, "Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia," *arXiv preprint arXiv:2312.03664*, 2023.

[7] Q. Liu, "Does gpt-4 play dice," 2024. [Online]. Available: https://www.researchgate.net/publication/378314162

[8] A. Renda, A. K. Hopkins, and M. Carbin, "Can LLMs Generate Random Numbers? Evaluating LLM Sampling in Controlled Domains," *ICML 2023 Workshop: Sampling and Optimization in Discrete Space*, vol. 4, no. 1, pp. 1–22, 2023. [Online]. Available: http://people.csail.mit.edu/renda/llm-sampling-paper

[9] J. Gu, L. Pang, H. Shen, and X. Cheng, "Do LLMs Play Dice? Exploring Probability Distribution Sampling in Large Language Models for Behavioral Simulation," *arXiv*, 2024. [Online]. Available: http://arxiv.org/abs/2404.09043

[10] S. Ghaffary and R. Metz, "Openai nears launch of ai agents to automate tasks for users," *Bloomberg News*, 11 2024, accessed: [2025-11-15]. [Online]. Available: https://www.bloomberg.com/news/articles/2024-11-13/openai-nears-launch-of-ai-agents-to-automate-tasks-for-users

[11] Council of Financial Regulators, Australia, "Reassessing the case for central clearing of bonds and repos in australia: A response to consultation by the council of financial regulators," *Consultantion Paper*, March 2025, accessed: [2025-03-26]. [Online]. Available: https://www.cfr.gov.au/publications/consultations/2025/reassessing-the-case-for-central-clearing-of-bonds-and-repos-in-australia/pdf/response-to-consultation.pdf

[12] Bank of Canada, "Potential netting benefits from expanded central clearing in canada's fixed-income market," 2022. [Online]. Available: https://www.bankofcanada.ca/2022/06/staff-analytical-note-2022-8/

[13] B. of England, "The potential impact of broader central clearing on dealer balance sheet capacity: a case study of uk gilt and gilt repo markets," 2023. [Online]. Available: https://www.bankofengland.co.uk/working-paper/2023/the-potential-impact-of-broader-central-clearing-on-dealer-balance-sheet-capacity

[14] Federal Reserve Bank of New York, "The netting efficiencies of marketwide central clearing," 2021. [Online]. Available: https://www.newyorkfed.org/research/staff_reports/sr964

[15] Council of Financial Regulators, Australia, "Reassessing the case for central clearing of bonds and repos in australia a consultation paper by the council of financial regulators," *Consultantion Paper*, September 2024, accessed: [2024-12-12]. [Online]. Available: https://www.cfr.gov.au/publications/consultations/2024/reassessing-the-case-for-central-clearing-of-bonds-and-repos-in-australia/reassessing-the-case-for-central-clearing-of-bonds-and-repos-in-australia.html

[16] M. K. BRUNNERMEIER and L. H. PEDERSEN, "Predatory trading," *The Journal of Finance*, vol. 60, no. 4, pp. 1825–1863, 2005. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2005.00781.x

[17] L. Yang and H. Zhu, "Back-running: Seeking and hiding fundamental information in order flows," *Review of Financial Studies*, 2015, available at SSRN: https://ssrn.com/abstract=2583915.

[18] J. Shaki, S. Kraus, and M. Wooldridge, "Cognitive effects in large language models," in *ECAI 2023 - 26th European Conference on Artificial Intelligence, including 12th Conference on Prestigious Applications of Intelligent Systems, PAIS 2023 - Proceedings*, 2023, pp. 2105–2112.

[19] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 610–623. [Online]. Available: https://doi.org/10.1145/3442188.3445922

[20] M. Wooldridge and N. R. Jennings, "Intelligent agents: theory and practice," *The Knowledge Engineering Review*, vol. 10, no. 2, p. 115–152, 1995.

[21] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," *Proceedings of the National Academy of Sciences of the United States of America*, 2002.

[22] U. Wilensky and W. Rand, *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo*. The MIT Press, 2015. [Online]. Available: http://www.jstor.org/stable/j.ctt17kk851

[23] V. Dignum, N. Gilbert, and M. P. Wellman, "Introduction to the special issue on autonomous agents for agent-based modeling," *Autonomous Agents and Multi-Agent Systems*, 2016.

[24] C. Macal and M. North, "Introductory tutorial: Agent-based modeling and simulation," in *Proceedings of the Winter Simulation Conference 2014*, 2014, pp. 6–20.

[25] J. M. Epstein and R. L. Axtell, *Growing Artificial Societies: Social Science from the Bottom Up*, ser. MIT Press Books. The MIT Press, 1996. [Online]. Available: https://ideas.repec.org/b/mtp/titles/0262550253.html

[26] T. Pike, "Multi-level mesa," *arXiv*, 2019. [Online]. Available: https://arxiv.org/abs/1904.08315

[27] David Masad and Jacqueline Kazil, "Mesa: An Agent-Based Modeling Framework," in *Proceedings of the 14th Python in Science Conference*, Kathryn Huff and James Bergstra, Eds., 2015, pp. 51 – 58.

[28] T. Pike, S. Golden, D. Lowdermilk, B. Luong, and B. Rosado, "Growing the simulation ecosystem: introducing mesa data to provide transparent, accessible, and extensible data pipelines for simulation development," *SIMULATION*, 2021.

[29] J. Kazil, D. Masad, D. Masad, and A. Crooks, "Utilizing python for agent-based modeling: The mesa framework," *SBP-BRiMS*, 2020.

[30] R. L. Axtell and J. D. Farmer, "Agent-Based Modeling in Economics and Finance: Past, Present, and Future," *Journal of Economic Literature*, pp. 1–96, 2022.

[31] D. C. Parkes, *Iterative Combinatorial Auctions Chapter 2 : Iterative Combinatorial Auctions*. MIT Press, 2006.

[32] Y. Shoham and K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. MASFoundation.org, 2008.

[33] S. Fatima, S. Kraus, and M. Wooldridge, *Principles of Automated Negotiation.* Cambridge University Press, 2014.

[34] G. Fagiolo, M. Guerini, F. Lamperti, A. Moneta, and A. Roventini, "Validation of agent-based models in economics and finance," *Simulation Foundations, Methods and Applications*, 2017.

[35] J. Lussange, A. Belianin, S. Bourgeois-Gironde, and B. Gutkin, "Learning and cognition in financial markets: A paradigm shift for agent-based models," *IntelliSys*, 2020.

[36] A. Vermeir and H. Bersini, "Best practices in programming agent-based models in economics and finance," *Lecture Notes in Economics and Mathematical Systems*, 2015.

[37] F. Klügl, M. Fehler, and R. Herrler, "About the role of the environment in multi-agent simulations," in *Environments for Multi-Agent Systems*, D. Weyns, H. Van Dyke Parunak, and F. Michel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 127–149.

[38] J. Thangarajah, L. Padgham, and J. Harland, "Representation and reasoning for goals in BDI agents," *Australian Computer Science Communications, month = jan, pages = 259–265, numpages = 7,*, vol. 24, no. 1, 2002.

[39] D. Baumeister, *The Future of Decision-Making.* Cham: Springer International Publishing, 2019, pp. 15–20. [Online]. Available: https://doi.org/10.1007/978-3-030-18050-8_3

[40] A. Burkov and B. Chaib-Draa, "Repeated games for multiagent systems: a survey," *Knowledge Engineering Review*, vol. 29, no. 1, pp. 1–30, 2014.

[41] R. Meir and D. Parkes, "Playing the wrong game: Bounding externalities in diverse populations of agents," 2018. [Online]. Available: https://arxiv.org/abs/1411.1751

[42] M. Winikoff, L. Padgham, J. Harland, and J. Thangarajah, "Declarative & procedural goals in intelligent agent systems," in *Proceedings of the Eights International*

Conference on Principles of Knowledge Representation and Reasoning, ser. KR'02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, p. 470–481.

[43] M. Fisher, V. Mascardi, K. Y. Rozier, B. H. Schlingloff, M. Winikoff, and N. Yorke-Smith, *Towards a framework for certification of reliable autonomous systems*. Springer US, 2021, vol. 35, no. 1. [Online]. Available: https://doi.org/10.1007/s10458-020-09487-2

[44] S. Bai, W. Raskob, and T. Müller, "Agent based model," *Radioprotection*, vol. 55, 05 2020.

[45] M. Wellman, A. Greenwald, and P. Stone, *Autonomous Bidding Agents: Strategies and Lessons from the Trading Agent Competition*. MIT Press, 07 2007.

[46] R. Lin, S. Kraus, T. Baarslag, D. Tykhonov, K. Hindriks, and C. M. Jonker, "Genius: An integrated environment for supporting the design of generic automated negotiators," *Computational Intelligence*, vol. 30, no. 1, pp. 48–70, 2014.

[47] Y. Shoham, "Computer science and game theory," *Communications of the ACM*, vol. 51, no. 8, pp. 74–79, 2008.

[48] F. Klügl, "'Engineering' Agent-Based Simulation Models?" in *Agent-Oriented Software Engineering XIII*, J. P. Müller and M. Cossentino, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 179–196.

[49] F.Klügl, "A validation methodology for agent-based simulations," in *ACM Symposium on Applied Computing*, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:9450992

[50] F. Klügl and A. L. C. Bazzan, "Agent-based modeling and simulation," *AI Magazine*, vol. 33, no. 3, p. 29, Sep. 2012.

[51] N. Gilbert, *Agent-Based Models*. The Centre for Research in Social Simulation, SAGE Publications, 01 2007.

[52] A. Overwater and N. Yorke-Smith, "Agent-based simulation of short-term peer-to-peer rentals: Evidence from the amsterdam housing market," *Environment and Planning B: Urban Analytics and City Science*, vol. 49, no. 1, pp. 223–240, 2022. [Online]. Available: https://doi.org/10.1177/23998083211000747

[53] C. M. Macal and M. J. North, "Agent-based modeling and simulation," in *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 2009, pp. 86–98.

[54] E. Wiegel and N. Yorke-Smith, "An agent-based market analysis of urban housing balance in the netherlands," *Real Estate*, vol. 1, no. 1, pp. 80–135, 2024. [Online]. Available: https://www.mdpi.com/2813-8090/1/1/6

[55] A. W. Lo, "The adaptive markets hypothesis: Market efficiency from an evolutionary perspective," *The Journal of Portfolio Management*, 2004.

[56] A.W. Lo, *Adaptive Markets: Financial Evolution at the Speed of Thought.* Princeton University Press, 2017. [Online]. Available: http://www.jstor.org/stable/j.ctvc77k3n

[57] A. A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzun, "The flash crash: High-frequency trading in an electronic market," *Journal of Finance*, 2017.

[58] J. Paulin, A. Calinescu, and M. Wooldridge, "Understanding flash crash contagion and systemic risk: A micro–macro agent-based approach," *Journal of Economic Dynamics and Control*, vol. 100, p. 200–229, Mar. 2019. [Online]. Available: http://dx.doi.org/10.1016/j.jedc.2018.12.008

[59] M. Fehler, F. Klügl, and F. Puppe, "Approaches for resolving the dilemma between model structure refinement and parameter calibration in agent-based simulations," *Adaptive Agents and Multi-Agent Systems*, 2006.

[60] M. Dann, Y. Yao, B. Logan, J. Thangarajah *et al.*, "Multi-agent intention progression with black-box agents." in *IJCAI*, 2021, pp. 132–138.

[61] S. Rodriguez, J. Thangarajah, and M. Winikoff, "A Behaviour-Driven Approach for Testing Requirements via User and System Stories in Agent Systems," *Proceedings of*

the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS, vol. 2023-May, pp. 1182–1190, 2023.

[62] C. D. Nguyen, A. Perini, C. Bernon, J. Pavón, and J. Thangarajah, "Testing in multi-agent systems," in *Agent-Oriented Software Engineering X*, M.-P. Gleizes and J. J. Gomez-Sanz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 180–190.

[63] Y. Abushark, T. Miller, J. Thangarajah, M. Winikoff, and J. Harland, "Requirements specification via activity diagrams for agent-based systems," *Autonomous Agents and Multi-Agent Systems*, vol. 31, no. 3, pp. 423–468, 5 2017. [Online]. Available: https://doi.org/10.1007/s10458-016-9327-7

[64] P. Avegliano and J. S. Sichman, "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*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1781–1783.

[65] K. Boggess, S. Kraus, and L. Feng, "Toward policy explanations for multi-agent reinforcement learning," *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, 2022. [Online]. Available: arXivpreprintarXiv:2204.12568

[66] L. Ardon, N. Vadori, T. Spooner, M. Xu, J. Vann, and S. Ganesh, "Towards a fully rl-based market simulator," *Proceedings of the Second ACM International Conference on AI in Finance*, Nov 2021. [Online]. Available: http://dx.doi.org/10.1145/3490354.3494372

[67] F. Lamperti, A. Roventini, and A. Sani, "Agent-based model calibration using machine learning surrogates," 2017. [Online]. Available: https://arxiv.org/abs/1703.10639

[68] S. Ganesh, N. Vadori, M. Xu, H. Zheng, P. Reddy, and M. Veloso, "Reinforcement learning for market making in a multi-agent dealer market," *arXiv preprint arXiv:1911.05892*, 2019.

[69] N. Vadori, L. Ardon, S. Ganesh, T. Spooner, S. Amrouni, J. Vann, M. Xu, Z. Zheng, T. Balch, and M. Veloso, "Towards multi-agent reinforcement learning driven over-the-counter market simulations," 2022.

[70] P. Bank, I. Ekren, and J. Muhle-Karbe, "Liquidity in competitive dealer markets," *Mathematical Finance*, vol. 31, no. 3, p. 827–856, Apr 2021. [Online]. Available: http://dx.doi.org/10.1111/mafi.12305

[71] A. Barzykin, P. Bergault, and O. Guéant, "Algorithmic market making in dealer markets with hedging and market impact," 2022. [Online]. Available: https://arxiv.org/abs/2106.06974

[72] R. Cont and M. S. Muller, "A stochastic partial differential equation model for limit order book dynamics," *SIAM Journal on Financial Mathematics*, vol. 12, no. 2, pp. 744–787, 2021.

[73] Y. Li, N. Yorke-Smith, and T. Keviczky, "Machine learning enabled uncertainty set for data-driven robust optimization," *Journal of Process Control*, vol. 144, p. 103339, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959152424001793

[74] R. K. Tripathy and I. Bilionis, "Deep UQ: Learning deep neural network surrogate models for high dimensional Uncertainty Quantification," *Journal of Computational Physics*, vol. 375, p. 565–588, Dec 2018. [Online]. Available: http://dx.doi.org/10.1016/j.jcp.2018.08.036

[75] A. Collins, M. Koehler, and C. Lynch, "Methods that support the validation of agent-based models: An overview and discussion," *Journal of Artificial Societies and Social Simulation*, vol. 27, no. 1, p. 11, 2024. [Online]. Available: http://jasss.soc.surrey.ac.uk/27/1/11.html

[76] C. D. Nguyen, A. Perini, C. Bernon, J. Pavón, and J. Thangarajah, "Testing in multi-agent systems," in *Agent-Oriented Software Engineering X*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 180–190.

[77] B. S. Onggo, L. Yilmaz, F. Klügl, T. Terano, and C. M. MacAl, "Credible Agent-Based Simulation - An Illusion or only A Step Away," *Proceedings - Winter Simulation Conference*, vol. 2019-Decem, no. December, pp. 273–284, 2019.

[78] J. Harland, D. N. Morley, J. Thangarajah, and N. Yorke-Smith, "An operational semantics for the goal life-cycle in bdi agents," *Autonomous Agents and Multi-Agent Systems*, vol. 28, no. 4, pp. 682–719, 7 2014.

[79] R. Hayes, A. Todd, N. Chaidarun, S. Tepsuporn, P. Beling, and W. Scherer, "An agent-based financial simulation for use by researchers," in *Proceedings of the Winter Simulation Conference 2014*, 2014, pp. 300–309.

[80] A. Quera-Bofarull, J. Dyer, A. Calinescu, and M. Wooldridge, "Some challenges of calibrating differentiable agent-based models," 2023. [Online]. Available: https://arxiv.org/abs/2307.01085

[81] B. S. Onggo, L. Yilmaz, F. Klügl-Frohnmeyer, T. Terano, and C. M. Macal, "Credible agent-based simulation – an illusion or only a step awayf," *2019 Winter Simulation Conference (WSC)*, pp. 273–284, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:211243558

[82] S. A. Suryanarayana, D. Sarne, and S. Kraus, "Explainability in mechanism design: recent advances and the road ahead," in *European Conference on Multi-Agent Systems*. Springer, 2022, pp. 364–382.

[83] S. Rodriguez, J. Thangarajah, and A. Davey, "Design patterns for explainable agents (xag)," in *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 2024, pp. 1621–1629.

[84] E. Yan, S. Burattini, J. F. Hübner, and A. Ricci, "A multi-level explainability framework for engineering and understanding bdi agents," *Autonomous Agents and Multi-Agent Systems*, vol. 39, no. 1, p. 9, 2025.

[85] Y. Zhang, Z. Li, and Y. Zhang, "Validation and calibration of an agent-based model: A surrogate approach," *Discrete Dynamics in Nature and Society*, vol. 2020, no. 1, p. 6946370, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2020/6946370

[86] A. Quera-Bofarull, A. Chopra, A. Calinescu, M. Wooldridge, and J. Dyer, "Bayesian calibration of differentiable agent-based models," 2023. [Online]. Available: https://arxiv.org/abs/2305.15340

[87] A. Quera-Bofarull, A. Chopra, J. Aylett-Bullock, C. Cuesta-Lazaro, A. Calinescu, R. Raskar, and M. Wooldridge, "Don't simulate twice: One-shot sensitivity analyses via automatic differentiation," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS '23. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2023, p. 1867–1876.

[88] H. Van Dyke Parunak, R. Savit, and R. L. Riolo, "Agent-based modeling vs. equation-based modeling: A case study and users' guide," in *Multi-Agent Systems and Agent-Based Simulation*, J. S. Sichman, R. Conte, and N. Gilbert, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 10–25.

[89] R. Bagni, R. Berchi, and P. Cariello, "A comparison of simulation models applied to epidemics," *Journal of Artificial Societies and Social Simulation*, vol. 5, 06 2002.

[90] Y. Mansour, Y. Mansour, Y. Mansour, B. Moldovanu, N. Nisan, and B. Vöcking, "Electronic markets and auctions (dagstuhl seminar 13461)," *Dagstuhl Reports*, 2014.

[91] M. Babaioff, B. Lucier, N. Nisan, and R. P. Leme, "On the efficiency of the walrasian mechanism," *CoRR*, vol. abs/1311.0924, 2013. [Online]. Available: http://arxiv.org/abs/1311.0924

[92] Y. A. Gonczarowski, “Manipulation of stable matchings using minimal blacklists,” *CoRR*, vol. abs/1307.7477, 2013. [Online]. Available: http://arxiv.org/abs/1307.7477

[93] S. Dobzinski, N. Nisan, and S. Oren, “Economic efficiency requires interaction,” in *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, ser. STOC ’14. New York, NY, USA: Association for Computing Machinery, 2014, p. 233–242. [Online]. Available: https://doi.org/10.1145/2591796.2591815

[94] G. Lee, “Personalized recommendations on a matching platform,” *Available at SSRN: https://ssrn.com/abstract=3486632*, 2019.

[95] A. D. Procaccia and M. Tennenholtz, “Approximate mechanism design without money,” *ACM Trans. Econ. Comput.*, vol. 1, no. 4, Dec. 2013. [Online]. Available: https://doi.org/10.1145/2542174.2542175

[96] J. D. Donna, P. Schenone, and G. F. Veramendi, “Networks, frictions, and price dispersion,” *Games and Economic Behavior*, vol. 124, pp. 406–431, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0899825620301329

[97] A. Polanski and F. Vega-Redondo, “Bargaining frictions in trading networks,” *The B.E. Journal of Theoretical Economics*, vol. 18, no. 1, p. 20160089, 2018. [Online]. Available: https://doi.org/10.1515/bejte-2016-0089

[98] Y. Cai and G. Velegkas, “How to sell information optimally: an algorithmic study,” *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2011.14570

[99] A. Kirilenko, A. Kyle, , M. Samadi, and T. Tuzun, “Automation, intermediation and the flash crash,” *Journal of Investment Management*, vol. 16, 2018.

[100] B. LeBaron, “Microconsistency in Simple Empirical Agent-Based Financial Models,” *Computational Economics*, vol. 58, no. 1, pp. 83–101, 2021. [Online]. Available: https://doi.org/10.1007/s10614-019-09917-8

[101] N. T. Chan, B. Lebaron, A. W. Lo, and T. Poggio, "Agent-Based Models of Financial Markets: A Comparison with Experimental Markets," *Computing in Economics and Finance Conference Available at SSRN: https://ssrn.com/abstract=290140*, 1999.

[102] A. Todd, P. A. Beling, W. T. Scherer, and S. Y. Yang, "Agent-based financial markets: A review of the methodology and domain," *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2016.

[103] K. Braun-Munzinger, Z. Liu, and A. E. T. and, "An agent-based model of corporate bond trading," *Quantitative Finance*, vol. 18, no. 4, pp. 591–608, 2018. [Online]. Available: https://doi.org/10.1080/14697688.2017.1380310

[104] D. Cockburn, Z. Kobti, and T. Kohler, "A reinforcement learning model for economic agent specialization." *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, FLAIRS-23*, January 2010.

[105] C. Hui, M. Goldberg, M. Magdon-Ismail, and W. A. Wallace, "Simulating the diffusion of information: An agent-based modeling approach," *Int. J. Agent Technol. Syst.*, vol. 2, no. 3, p. 31–46, Jul. 2010. [Online]. Available: https://doi.org/10.4018/jats.2010070103

[106] B. Qian and H. H. Cheng, "A mobile agent-based coalition formation system for multi-robot systems," *2016 12th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)*, 2016.

[107] S. J. Leal, M. Napoletano, A. Roventini, and G. Fagiolo, "Rock around the clock: An agent-based model of low- and high-frequency trading," *Journal of Evolutionary Economics*, 2016.

[108] Y. Lin, J. Xue, and L. Raschid, "Predicting the behavior of dealers in over-the-counter corporate bond markets," in *Proceedings of the First ACM International Conference on AI in Finance*, ser. ICAIF '20. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3383455.3422542

[109] C. Collver, "An application of agent-based modeling to market structure policy: the case of the u.s. tick size pilot program and market maker profitability," *U.S.Securities and Exchange Commission*, 2017.

[110] M. E. Paddrik, R. Hayes, A. Todd, S. Y. Yang, P. A. Beling, and W. T. Scherer, "An agent based model of the e-mini s&p 500 applied to flash crash analysis," *IEEE Conference on Computational Intelligence for Financial Engineering and Economics*, 2012.

[111] E. G. Peterson and L. S. Weber, *The stock market flash crash of 2010.* Nova Science Publishing Incorporated, 2011.

[112] A. Vidler and T. Walsh, "Decoding OTC government bond market liquidity: An abm model for market dynamics," 2024. [Online]. Available: https://arxiv.org/abs/2501.16331

[113] N. Capodieci, G. Cabri, G. A. Pagani, and M. Aiello, "Adaptive game-based agent negotiation in deregulated energy markets," *2012 International Conference on Collaboration Technologies and Systems (CTS)*, 2012.

[114] L. Mitridati, J. Kazempour, and P. Pinson, "Design and game-theoretic analysis of community-based market mechanisms in heat and electricity systems," *Omega*, vol. 99, p. 102177, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0305048319305523

[115] M. A. C. L. Gunarathna, R. J. Yang, and A. Song, "Diverse distributed renewable energy trading paradigms: a business model review," *Journal of Environmental Planning and Management*, vol. 65, no. 1, pp. 1–36, 2022. [Online]. Available: https://doi.org/10.1080/09640568.2021.1877640

[116] E. Kanellou, O. Mastakas, D. Askounis, D. Askounis, and O. Mastakas, "A game theory approach towards real-time pricing in future electricity markets," *International Journal of Decision Support Systems*, vol. 4, 2021.

[117] Y. Wang, X. Wang, Y. Kuang, Q. Peng, H. Zhao, and Z. Wang, "Network-constrained energy consumption game for dynamic pricing markets," *CSEE Journal of Power and Energy Systems*, vol. 8, no. 2, pp. 548–558, 2022.

[118] J. S. Hong and M. Kim, "Game-theory-based approach for energy routing in a smart grid network," *Journal of Computer Networks and Communications*, vol. 2016, no. 1, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2016/4761720

[119] Sarjiya, R. F. S. Budi, and S. P. Hadi, "Game theory for multi-objective and multi-period framework generation expansion planning in deregulated markets," *Energy*, vol. 174, pp. 323–330, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544219303044

[120] R. Zhao, J. Han, S. Zhong, and Y. Huang, "Interaction between enterprises and consumers in a market of carbon-labeled products: a game theoretical analysis," *Environmental Science and Pollution Research*, 2018.

[121] R. Meir, H. Ma, and V. Robu, "Contract design for energy demand response," 2017. [Online]. Available: https://arxiv.org/abs/1705.07300

[122] W. A. Brock and C. H. Hommes, "Heterogeneous beliefs and routes to chaos in a simple asset pricing model," *Journal of Economic Dynamics and Control*, vol. 22, no. 8, pp. 1235–1274, 1998.

[123] S. R. of the Joint CFTC-SEC Advisory Committee on Emerging Regulatory Issues, "Report of the staffs of the cftc and sec to the joint advisory committee on emerging regulatory issues. the stock market flash crash of 2010," *Securities and Exchange Commision, USA*, 2010.

[124] Y. Wu, J. Liang, and B. Hu, "A free boundary problem for defaultable corporate bond with credit rating migration risk and its asymptotic behavior," *Discrete and Continuous Dynamical Systems - B*, vol. 25, no. 3, pp. 1043–1058, 2020. [Online]. Available: https://www.aimsciences.org/article/id/e8b510ad-bab0-415a-84a1-ea3fbe79decf

[125] F. A. LONGSTAFF and E. S. SCHWARTZ, "A simple approach to valuing risky fixed and floating rate debt," *The Journal of Finance*, vol. 50, no. 3, pp. 789–819, 1995. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1995.tb04037.x

[126] Z. Huang, T. Jiang, and Z. Wang, "On a multiple credit rating migration model with stochastic interest rate," *Mathematical Methods in the Applied Sciences*, vol. 43, no. 12, pp. 7106–7134, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mma.6435

[127] D. Duffie and K. J. Singleton, "Modeling term structures of defaultable bonds," *The Review of Financial Studies*, vol. 12, no. 4, pp. 687–720, 06 2015. [Online]. Available: https://doi.org/10.1093/rfs/12.4.687

[128] F. Black and J. C. Cox, "Valuing corporate securities: Some effects of bond indenture provisions," *The Journal of Finance*, vol. 31, no. 2, pp. 351–367, 1976. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1976.tb01891.x

[129] R. C. Merton, "On the pricing of corporate debt: The risk structure of interest rates," *The Journal of Finance*, vol. 29, no. 2, pp. 449–470, 1974. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1974.tb03058.x

[130] J. Liang, H.-M. Yin, X. Chen, and Y. Wu, "On a corporate bond pricing model with credit rating migration risksand stochastic interest rate," *Quantitative Finance and Economics*, vol. 1, no. 3, pp. 300–319, 2017. [Online]. Available: https://www.aimspress.com/article/doi/10.3934/QFE.2017.3.300

[131] C. Albanese and A. Vidler, "Dynamic conditioning and credit correlation baskets," University Library of Munich, Germany, MPRA Paper, 2008. [Online]. Available: https://EconPapers.repec.org/RePEc:pra:mprapa:8368

[132] C. Albanese, O. Chen, A. Dalessandro, , and A. Vidler, "Dynamic credit correlation modeling," *Imperial College - Working paper, Available*, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:181740

[133] C. Albanese and A. Vidler, "A structural model for credit-equity derivatives and bespoke cdos. willmott magazine," *Wilmott Magazine, June 2007, Available at SSRN: https://ssrn.com/abstract=985371*, 2007.

[134] J. Wang, T. Sun, B. Liu, Y. Cao, and D. Wang, "Financial markets prediction with deep learning," *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2018. [Online]. Available: http://dx.doi.org/10.1109/ICMLA.2018.00022

[135] E. Kremer and B. Tinel, "Contingent convertible bonds and macroeconomic stability in a stock-flow consistent model," *Metroeconomica*, vol. 73, no. 4, pp. 1112–1154, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/meca.12392

[136] S. Holm, O. Thees, R. Lemm, R. Olschewski, and L. M. Hilty, "An agent-based model of wood markets: Scenario analysis," *Forest Policy and Economics*, vol. 95, pp. 26–36, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S138993411830234X

[137] J. C. Hull, *Options, Futures, and Other Derivatives*, 11th ed. London: Pearson, 2022.

[138] Australian Securities and Investments Commission, "Conduct risk in wholesale fixed income markets: Report 741," October 2022, accessed: [2025-03-12]. [Online]. Available: https://download.asic.gov.au/media/5aqldfff/rep741-published-25-october-2022.pdf

[139] Bank of England, "Official Operations in the Gilt Market An Operational Notice," 2023. [Online]. Available: https://www.dmo.gov.uk/media/wlcjiyi3/opnot280324.pdf

[140] Y. Cheung, "Trading european sovereign bonds: The microstructure of the MTS trading platforms," *European Central Bank*, vol. 432, 02 2005.

[141] A. Chaboud, E. Correia-Golay, C. Cox, M. J. Fleming, Y. Huh, F. M. Keane, K. Lee, K. Schwarz, C. Vega, and C. Windover, "All-to-All Trading in the U.S. Treasury Market," 2022. [Online]. Available: https://www.newyorkfed.org/research/staff_reports/sr1036

[142] International Capital Markets Association, "The microstructure of the mts trading platforms,"," *International Capital Markets Association*, 2023. [Online]. Available: https://www.icmagroup.org/assets/SMPC-European-Secondary-Bond-Market-Data-H1-2023-270923.pdf

[143] G. Pinter, "An anatomy of the 2022 gilt market crisis," *Bank of England*, vol. Staff Working Paper 1019, 2023.

[144] Australian Financial Review, "Bond market players call time on $6 trillion aussie exchange," *Australian Financial Review*, 2023, accessed: [2023-04-23]. [Online]. Available: https://www.afr.com/chanticleer/bond-market-players-call-time-on-6-trillion-aussie-exchange-20230525-p5db5n

[145] TradeWeb, "Bloomberg, marketaxess and tradeweb sign joint venture agreement," 2023, accessed: [2023-04-23]. [Online]. Available: https://www.tradeweb.com/newsroom/media-center/news-releases/bloomberg-marketaxess-and-tradeweb-sign-joint-venture-agreement

[146] D. Duffie, "Still the World's Safe Haven?" *Hutchins Working Paper*, vol. 62, 2020. [Online]. Available: https://www.brookings.edu/wp-content/uploads/2020/05/WP62_Duffie_updated.pdf

[147] Federal Reserve Bank of New York, "All-to-all trading in the u.s. treasury market," 2024. [Online]. Available: https://www.newyorkfed.org/research/staff_reports/sr1036

[148] P. Asquith, T. R. Covert, and P. A. Pathak, "The effects of mandatory transparency in financial market design: Evidence from the corporate bond market," *National Bureau of Economic Research*, 2013.

[149] A. Owens and C. Weldon, "SEC's fixed income market structure advisory committee discusses bond market liquidity in its inaugural meeting," *Journal of Investment Compliance*, vol. 19, 2018.

[150] H. Blommestein and H. J. Blommestein, "Impact of regulatory changes on government bond market liquidity," *Journal of Financial Regulation and Compliance*, vol. 25, 2017.

[151] The U.S Treasury Department and Board, "Recent Disruptions and potential reforms in the U.S Treasury market: A STAFF PROGRESS REPORT," 2021. [Online]. Available: https://www.sec.gov/about/reports-publications/iawg-treasury-report

[152] L. Inglada-perez, "A comprehensive framework for uncovering non- linearity and chaos in financial markets: Empirical evidence for four major stock market indices," *Entropy*, vol. 22, no. 12, pp. 1–26, 2020.

[153] L. Harris, *Trading and Exchanges: Market Microstructure for Practitioners*. Oxford University Press, New York, 2003.

[154] Reserve Bank of Australia, "Government bond markets in advanced economies during the pandemic," *Reserave Bank of Australia*, pp. 52 – 60, 2021. [Online]. Available: www.rba.gov.au/publications/bulletin/2021/sep/government-bond-markets-in-advanced-economies-during-the-pandemic.html

[155] World Bank Group, "Financial stability," accessed: [2024-09-25]. [Online]. Available: https://www.worldbank.org/en/publication/gfdr/gfdr-2016/background/financial-stability

[156] Bank of England, "Financial stability in focus: The fpc's approach to assessing risks in market-based finance," 2023, accessed: [25 Sep 2024]. [Online]. Available: https://www.bankofengland.co.uk/financial-stability-in-focus/2023/october-2023

[157] Australian Government, "Australian Federal Budget 2022," 2022, accessed: [2025-03-25]. [Online]. Available: https://www.dfat.gov.au/about-us/corporate/portfolio-budget-statements/budget-highlights-2022-23

[158] S. M. Frank J. Fabozzi, *The Handbook of Fixed Income Securities, Seventh Edition.* New York: McGraw-Hill, 2005.

[159] G. Pinter, C. Wang, and J. Zou, "Size Discount and Size Penalty: Trading Costs in Bond Markets," *Review of Financial Studies*, vol. 37, no. 7, pp. 2156–2190, 2024.

[160] J. Cheshire, "Market Making in Bond Markets," *RBA Bulletin March Quater 2015*, vol. March 2015, no. March, pp. 63–74, 2015. [Online]. Available: https://rba.gov.au/publications/bulletin/2015/mar/7.html

[161] Bank for International Settlements, "Market-making and proprietary trading: industry trends, drivers and policy implications," *Bank for International Settlements*, vol. 52, December 2014. [Online]. Available: https://www.bis.org/publ/cgfs52.htm

[162] Bank of International Settlements, "BIS Quarterly Review, December 2023," 2023. [Online]. Available: https://www.bis.org/publ/qtrpdf/r_qt2312.htm

[163] D. Baumeister, D. Neugebauer, J. Rothe, and H. Schadrack, "Verification in incomplete argumentation frameworks," *Artificial Intelligence*, vol. 264, pp. 1–26, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370218304417

[164] W. Dvořák and S. Woltran, "Complexity of semi-stable and stage semantics in argumentation frameworks," *Information Processing Letters*, vol. 110, no. 11, pp. 425–430, 2010.

[165] J.-G. Mailly, "Yes, no, maybe, I don't know: Complexity and application of abstract argumentation with incomplete knowledge," *Argument & Computation*, vol. 13, no. 3, pp. 291–324, 2022.

[166] P. E. Dunne and M. Wooldridge, "Complexity of abstract argumentation," *Argumentation in artificial intelligence*, pp. 85–104, 2009.

[167] D. Baumeister, M. Järvisalo, D. Neugebauer, A. Niskanen, and J. Rothe, "Acceptance in incomplete argumentation frameworks," *Artificial Intelligence*, vol. 295, p. 103470, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370221000217

[168] D. Baumeister, D. Neugebauer, and J. Rothe, "Credulous and skeptical acceptance in incomplete argumentation frameworks," in *Computational Models of Argument*. IOS Press, 2018, pp. 181–192.

[169] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, p. 6000–6010, 2017.

[170] S. Bubeck, V. Chandrasekaran, R. Eldan, J. A. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y.-F. Li, S. M. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, "Sparks of artificial general intelligence: Early experiments with gpt-4," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257663729

[171] H. Luo, Q. Sun, C. Xu, P. Zhao, J. Lou, C. Tao, X. Geng, Q. Lin, S. Chen, and D. Zhang, "WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct," 2023. [Online]. Available: http://arxiv.org/abs/2308.09583

[172] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI*, 2018.

[173] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Proceedings of Workshop at ICLR*, vol. 2013, 01 2013.

[174] A. Chopra, S. Kumar, N. Giray-Kuru, R. Raskar, and A. Quera-Bofarull, "On the limits of agency in agent-based models," *arXiv*, 2024. [Online]. Available: http://arxiv.org/abs/2409.10568

[175] J. Park, J. O'Brien, C. Cai, M. R. Morris, P. Liang, and M. Bernstein, "Generative agents: Interactive simulacra of human behavior," 2023. [Online]. Available: https://arxiv.org/abs/2304.03442

[176] S. Abdelnabi, A. Gomaa, S. Sivaprasad, L. Schönherr, and M. Fritz, "Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation," 2024. [Online]. Available: https://arxiv.org/abs/2309.17234

[177] Y. Zhang, S. Mao, T. Ge, X. Wang, A. Wynter, Y. Xia, and W. Wu, "Llm as a mastermind: A survey of strategic reasoning with large language models," 2024. [Online]. Available: https://arxiv.org/abs/2404.01230

[178] T. Liu, J. Yang, and Y. Yin, "Llm-abm for transportation: Assessing the potential of llm agents in system analysis," 2025. [Online]. Available: https://arxiv.org/abs/2503.22718

[179] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. E. Zhu, L. Jiang, X. Zhang, S. Zhang, A. Awadallah, R. W. White, D. Burger, and C. Wang, "Autogen: Enabling next-gen llm applications via multi-agent conversation," in *COLM 2024*, August 2024. [Online]. Available: https://www.microsoft.com/en-us/research/publication/autogen-enabling-next-gen-llm-applications-via-multi-agent-conversation-framework/

[180] S. Swarup, "Agency in the age of ai," 2025. [Online]. Available: https://arxiv.org/abs/2502.00648

[181] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2307.03109

[182] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, and A. F. et al, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," 2023. [Online]. Available: https://arxiv.org/abs/2206.04615

[183] A. Vidler and T. Walsh, "TraderTalk: An LLM Behavioural ABM applied to Simulating Human Bilateral Trading Interactions," in *2024 IEEE International Conference on Agents (ICA).* IEEE Computer Society, Dec. 2024, pp. 164–167. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICA63002.2024.00042

[184] A. Vidler and T. Walsh, "Evaluating binary decision biases in large language models: Implications for fair agent-based financial simulations," 2025. [Online]. Available: https://arxiv.org/abs/2501.16356

[185] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, and W. Yin, "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.* St. Julian's, Malta: Association for Computational Linguistics, 2024, pp. 225–237.

[186] S. Imani and L. Du, "MathPrompter: Mathematical Reasoning using Large Language Models," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 5, no. 1, pp. 37–42, 2023.

[187] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo *et al.*, "Solving quantitative reasoning problems with language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3843–3857, 2022.

[188] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[189] L. G. de Faria, “An Agent-Based Model With Realistic Financial Time Series: A Method for Agent-Based Models Validation,” arXiv.org, Papers 2206.09772, Jun. 2022. [Online]. Available: https://ideas.repec.org/p/arx/papers/2206.09772.html

[190] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, and Q. Yang, “Why johnny can’t prompt: How non-ai experts try (and fail) to design llm prompts,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544548.3581388

[191] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Comput. Surv.*, vol. 55, no. 12, Mar. 2023. [Online]. Available: https://doi.org/10.1145/3571730

[192] X. Zhang, A. Bosselut, M. Yasunaga, H. Ren, P. Liang, C. D. Manning, and J. Leskovec, “Greaselm: Graph Reasoning Enhanced Language Models for Question Answering,” pp. 1–16, 2022.

[193] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, “Emergent abilities of large language models,” 2022. [Online]. Available: https://arxiv.org/abs/2206.07682

[194] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On faithfulness and factuality in abstractive summarization,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 1906–1919. [Online]. Available: https://aclanthology.org/2020.acl-main.173/

[195] A. Creswell and M. Shanahan, “Faithful reasoning using large language models,” pp. 1–48, 2022. [Online]. Available: http://arxiv.org/abs/2208.14271

[196] Y. Fu, H. Peng, L. Ou, A. Sabharwal, and T. Khot, "Specializing Smaller Language Models towards Multi-Step Reasoning," *Proceedings of Machine Learning Research*, vol. 202, pp. 10 421–10 430, 2023.

[197] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.12948

[198] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Nee-

lakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[199] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[200] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825

[201] X. Guo and S. Vosoughi, "Serial Position Effects of Large Language Models," 2024. [Online]. Available: http://arxiv.org/abs/2406.15981

[202] X. Xu, C. Tao, T. Shen, C. Xu, H. Xu, G. Long, and J. guang Lou, "Re-reading improves reasoning in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2309.06275

[203] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances in Neural Information Processing Systems*, vol. 35, no. NeurIPS, pp. 1–43, 2022.

[204] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large language model based multi-agents: A survey of progress and challenges," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, K. Larson, Ed. International Joint Conferences

on Artificial Intelligence Organization, 8 2024, pp. 8048–8057, survey Track. [Online]. Available: https://doi.org/10.24963/ijcai.2024/890

[205] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "BloombergGPT: A Large Language Model for Finance," pp. 1–76, 2023. [Online]. Available: http://arxiv.org/abs/2303.17564

[206] Y. Yu, H. Li, Z. Chen, Y. Jiang, Y. Li, D. Zhang, R. Liu, J. W. Suchow, and K. Khashanah, "Finmem: A performance-enhanced llm trading agent with layered memory and character design." in *AAAI Spring Symposia*, R. P. A. Petrick and C. W. Geib, Eds. AAAI PRess: AAAI Press, 2024, pp. 595–597. [Online]. Available: http://dblp.uni-trier.de/db/conf/aaaiss/aaaiss2024.html#YuLCJLZLSK24

[207] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior, 60th Anniversary Commemorative Edition.* Princeton: Princeton University Press, 2004. [Online]. Available: https://doi.org/10.1515/9781400829460

[208] R. J. Aumann, *Repeated Games with Incomplete Information*, ser. MIT Press Books. The MIT Press, December 1995, vol. 1. [Online]. Available: https://ideas.repec.org/b/mtp/titles/0262011476.html

[209] J. Nash, "Non-cooperative games," *Annals of mathematics*, vol. 54, no. 2, pp. 286–295, 1951.

[210] J. Harsanyi, "Games with incomplete information played by "bayesian" players," *Management Science*, vol. 50, pp. 1804–1817, 12 1967.

[211] N. Agmon, S. Kraus, G. A. Kaminka, and V. Sadov, "Adversarial uncertainty in multi-robot patrol," in *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, ser. IJCAI'09. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2009, p. 1811–1817.

[212] T. Basar and G. Zaccour, *Handbook of Dynamic Game Theory.* Germany: Springer, Aug. 2018, springer International Publishing AG, part of Springer Nature.

[213] G. A. Kaminka, "Swarms can be rational," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 383, no. 2289: Theme issue 'The road forward with swarm systems' compiled and edited by Hussein A. Abbass and Sanaz Mostaghim, 2025.

[214] A. Vidler and T. Walsh, "Playing games with large language models: Randomness and strategy," 2025. [Online]. Available: https://arxiv.org/abs/2503.02582

[215] ——, "Non cooperative liquidity games and their application to bond market trading," 2024. [Online]. Available: https://arxiv.org/abs/2405.02865

[216] D. Zhou, A. Uddin, X. Tao, Z. Shang, and D. Yu, "Temporal bipartite graph neural networks for bond prediction," in *Proceedings of the Third ACM International Conference on AI in Finance*, ser. ICAIF '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 308–316. [Online]. Available: https://doi.org/10.1145/3533271.3561751

[217] A. Turky, M. S. Rahaman, W. Shao, F. D. Salim, D. Bradbrook, and A. Song, "Deep learning assisted memetic algorithm for shortest route problems," in *Computational Science – ICCS 2020*, V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, and J. Teixeira, Eds. Springer International Publishing, 2020, pp. 109–121.

[218] Bank of England, "Gemm guidebook: A guide to the roles of the dmo and primary dealers in the uk government bond market," December 2015. [Online]. Available: https://dmo.gov.uk/media/ownmub2i/guidebook20150401.pdf

[219] TP ICAP, "Best execution," May 2021, accessed: [2024-01-02]. [Online]. Available: https://tpicap.comBestExecution.pdf

[220] J. Branke, "Artificial societies," in *Encyclopedia of Machine Learning.* Boston, MA: Springer US, 2010, pp. 44–48.

[221] Z. Mehrab, L. Stundal, S. Swarup, S. Venaktramanan, B. Lewis, H. Mortveit, C. Barrett, A. Pandey, C. Wells, A. Galvani, B. Singer, D. Leblang, R. Colwell,

and M. Marathe, “Network agency: An agent-based model of forced migration from ukraine,” in *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS ’24. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2024, p. 1372–1380.

[222] R. Czech and G. Pinter, “Informed Trading and the Dynamics of Client-Dealer Connections in Corporate Bond Markets,” *Bank of England*, vol. Staff Working Paper, no. 895, 2022. [Online]. Available: https://www.bankofengland.co.uk/working-paper/2020

[223] Tucker Balch and team, “Abides (agent-based interactive discrete event simulation),” 2020, a high-performance, agent-based market simulation platform. [Online]. Available: https://github.com/abides-sim/abides

[224] J. Kazil, D. Masad, and A. Crooks, “Utilizing python for agent-based modeling: The mesa framework,” in *Social, Cultural, and Behavioral Modeling*, R. Thomson, H. Bisgin, C. Dancy, A. Hyder, and M. Hussain, Eds. Cham: Springer International Publishing, 2020, pp. 308–317.

[225] J. Kehoe, “The specification of sugarscape,” *CoRR*, vol. abs/1505.06012, 2015. [Online]. Available: http://arxiv.org/abs/1505.06012

[226] C. D. RICH, “The rationale of the use of the geometric average as an investment index,” *Journal of the Institute of Actuaries (1886-1994)*, vol. 74, no. 2, pp. 338–339, 1948. [Online]. Available: http://www.jstor.org/stable/41138034

[227] C. Song and R. Kawai, “Monte carlo and variance reduction methods for structural reliability analysis: A comprehensive review,” *Probabilistic Engineering Mechanics*, vol. 73, p. 103479, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0266892023000681

[228] J. Dyer, P. Cannon, J. D. Farmer, and S. M. Schmon, “Black-box bayesian inference for agent-based models,” *Journal of Economic Dynamics and Control*, vol. 161, p. 104827, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165188924000198

[229] M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert, “Adaptive approximate bayesian computation,” *Biometrika*, vol. 96, no. 4, pp. 983–990, 10 2009. [Online]. Available: https://doi.org/10.1093/biomet/asp052

[230] P. Glasserman and J. Li, “Importance Sampling for Portfolio Credit Risk,” *Management Science*, vol. 51, no. 11, pp. 1643–1656, November 2005. [Online]. Available: https://ideas.repec.org/a/inm/ormnsc/v51y2005i11p1643-1656.html

[231] C. L. Ilut and M. Schneider, “Modeling uncertainty as ambiguity: a review,” National Bureau of Economic Research, Working Paper 29915, April 2022. [Online]. Available: http://www.nber.org/papers/w29915

[232] P. Bossaerts, P. Ghirardato, S. Guarnaschelli, and W. R. Zame, “Ambiguity in Asset Markets: Theory and Experiment,” *The Review of Financial Studies*, vol. 23, no. 4, pp. 1325–1359, 01 2010. [Online]. Available: https://doi.org/10.1093/rfs/hhp106

[233] L. Tjuatja, V. Chen, T. Wu, A. Talwalkwar, and G. Neubig, “Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design,” *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 1011–1026, 2024.

[234] A. Kim, M. Muhn, and V. Nikolaev, “Financial statement analysis with large language models,” 2024. [Online]. Available: https://arxiv.org/abs/2407.17866

[235] E. F. Fama, “Random walks in stock market prices,” *Financial Analysts Journal*, vol. 21, pp. 55–59, 1965. [Online]. Available: https://api.semanticscholar.org/CorpusID:8478792

[236] P. A. Warren, U. Gostoli, G. D. Farmer, W. El-Deredy, and U. Hahn, “A re-examination of "Bias" in human randomness perception,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 44, no. 5, pp. 663–680, 2018.

[237] N. Ginsburg and P. Karpiuk, “Random generation: Analysis of the responses,” *Perceptual and Motor Skills*, vol. 79, no. 3, pp. 1059–1067, 1994. [Online]. Available: https://doi.org/10.2466/pms.1994.79.3.1059

[238] J. N. Towse and D. Neil, “Analyzing human random generation behavior: A review of methods used and a computer program for describing performance,” *Behavior Research Methods, Instruments, and Computers*, vol. 30, no. 4, pp. 583–591, 1998.

[239] T. Angelike and J. Musch, “A comparative evaluation of measures to assess randomness in human-generated sequences,” *Behavior Research Methods*, vol. 56, no. 7, pp. 7831–7848, 2024. [Online]. Available: https://doi.org/10.3758/s13428-024-02456-7

[240] A. Wong, G. Merholz, and U. Maoz, “Characterizing human random-sequence generation in competitive and non-competitive environments using Lempel–Ziv complexity,” *Scientific Reports*, vol. 11, no. 1, 2021.

[241] S. Baena-Mirabete, R. García-Viedma, S. Fernández-Guinea, and P. Puig, “A model to analyze age-related differences in human-generated head-tail sequences,” *bioRxiv*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260777307

[242] R. M. Harrison, “A Comparison of Large Language Model and Human Performance on Random Number Generation Tasks,” in *Proceedings of The First Workshop on AI Behavioral Science (KDD24)*, vol. 1. Association for Computing Machinery, 2024.

[243] M. Peeperkorn, T. Kouwenhoven, D. Brown, and A. Jordanous, “Is Temperature the Creativity Parameter of Large Language Models?” 2024. [Online]. Available: http://arxiv.org/abs/2405.00492

[244] M. Figurska, M. Stańczyk, and K. Kulesza, “Humans cannot consciously generate random numbers sequences: Polemic study,” *Medical Hypotheses*, vol. 70, no. 1, pp. 182–185, 2008.

[245] J. S. Bridle, “Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters,” Cambridge, MA, USA, p. 211–217, 1989.

[246] M. Renze, "The effect of sampling temperature on problem solving in large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024, p. 7346–7356. [Online]. Available: http://dx.doi.org/10.18653/v1/2024.findings-emnlp.432

[247] Y. Zhu, J. Li, G. Li, Y. Zhao, J. Li, Z. Jin, and H. Mei, "Hot or cold? adaptive temperature sampling for code generation with large language models," in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. [Online]. Available: https://doi.org/10.1609/aaai.v38i1.27798

[248] S. Ouyang, Z. Zhang, B. Yan, X. Liu, Y. Choi, J. Han, and L. Qin, "Structured chemistry reasoning with large language models," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24. JMLR.org, 2024.

[249] L. Hickman, P. D. Dunlop, and J. L. Wolf, "The performance of large language models on quantitative and verbal ability tests: Initial evidence and implications for unproctored high-stakes testing," *International Journal of Selection and Assessment*, vol. 32, no. 4, pp. 499–511, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ijsa.12479

[250] Y. Wang, Z. Zhang, H. Chen, and H. Shen, "Reasoning with large language models on graph tasks: The influence of temperature," in *2024 5th International Conference on Computer Engineering and Application (ICCEA)*, 2024, pp. 630–634.

[251] K. V. Koevering and J. Kleinberg, "How random is random? evaluating the randomness and humaness of llms' coin flips," 2024. [Online]. Available: https://arxiv.org/abs/2406.00092

[252] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," 2017. [Online]. Available: https://arxiv.org/abs/1606.04080

[253] Common Crawl, "Common crawl: An open repository of web data," 2024, accessed: [2024-11-23]. [Online]. Available: https://commoncrawl.org

[254] S. Baack, "A critical analysis of the largest source for generative ai training data: Common crawl," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 2199–2208. [Online]. Available: https://doi.org/10.1145/3630106.3659033

[255] B. M. Tessema, A. Kedia, and T.-S. Chung, "Unifiedcrawl: Aggregated common crawl for affordable adaptation of llms on low-resource languages," 2024. [Online]. Available: https://arxiv.org/abs/2411.14343

[256] U.S. Securities and Exchange Commission, "Market structure: Exchange trade volume," 2024, accessed: [2024-09-10]. [Online]. Available: https://www.sec.gov/marketstructure/datavis

[257] ——, "Trade and order volume ratios," Staff working paper, Accessed: [2024-09-10], Tech. Rep., 2013. [Online]. Available: https://www.sec.gov/about/trade-order-volume-ratios

[258] J. D. Farmer and S. Skouras, "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.

[259] V. Dalko and M. H. Wang, "How Effective are the Order-to-Trade Ratio and Resting Time Regulations?" *Journal of Financial Regulation*, vol. 4, no. 2, pp. 321–325, 08 2018. [Online]. Available: https://doi.org/10.1093/jfr/fjy007

[260] CFA Institute, "Gender diversity in investment management: New research for practitioners on how to close the gender gap," *CFA Institute Research Foundation*, 2016, available at: https://www.cfainstitute.org/en/research/survey-reports/gender-diversity-report.

[261] C. Crowe, B. Middleweek, L. Ryan, A. Vidler, and B. Whiting, "The role of gender in promotion rates in the australian finance industry," 2024. [Online]. Available: https://arxiv.org/abs/2409.14384

[262] H. Kotek, R. Dockum, and D. Sun, "Gender bias and stereotypes in large language models," in *Proceedings of The ACM Collective Intelligence Conference*, ser. CI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 12–24. [Online]. Available: https://doi.org/10.1145/3582269.3615599

[263] G. Charness and U. Gneezy, "Strong evidence for gender differences in risk taking," *Journal of Economic Behavior & Organization*, vol. 83, no. 1, pp. 50–58, 2012, gender Differences in Risk Aversion and Competition. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167268111001521

[264] H. Chase *et al.*, "Langchain," https://github.com/langchain-ai/langchain, 2022, accessed: [2024-12-30].

[265] O. Topsakal and T. C. Akinci, "Creating large language model applications utilizing langchain: A primer on developing llm apps fast," in *International Conference on Applied Engineering and Natural Sciences*, vol. 1, 2023, pp. 1050–1056.

[266] Australian Prudential Regulation Authority, "Annual fund-level superannuation statistics," https://www.apra.gov.au/annual-fund-level-superannuation-statistics, 2003, accessed: [2024-08-05].

[267] K. Braun-Munzinger, Z. Liu, and A. Turrell, "Staff Working Paper No. 592 An agent-based model of dynamics in corporate bond trading," 2016. [Online]. Available: www.bankofengland.co.uk/research/Pages/workingpapers/default.aspx

[268] J. Varghese and M. Edward, "Relationship between job orientation and performance of sales people: A financial services industry perspective," *IIM Kozhikode Society & Management Review*, vol. 7, no. 1, pp. 88–96, 2018.

[269] R. Cont, M. Cucuringu, V. Glukhov, and F. Prenzel, “Analysis and modeling of client order flow in limit order markets,” *Quantitative Finance*, vol. 23, no. 2, pp. 187–205, 2023. [Online]. Available: https://doi.org/10.1080/14697688.2022.2150282

[270] M. Massa and A. Simonov, “Reputation and interdealer trading: a microstructure analysis of the treasury bond market,” *Journal of Financial Markets*, vol. 6, no. 2, pp. 99–141, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386418102000459

[271] J.-D. Fermanian, O. Guéant, and J. Pu, “The Behavior of Dealers and Clients on the European Corporate Bond Market: The Case of Multi-Dealer-to-Client Platforms,” *Market Microstructure and Liquidity*, vol. 02, no. 03n04, 2016.

[272] P. Knops and N. Yorke-Smith, “Influence of reputation on resistance against bad agents in the iterated prisoner’s dilemma,” 2021, available online at: https://repository.tudelft.nl/record/uuid:c3580c40-d103-4519-bc0e-c77a6d03f3de.

[273] M. Toman and N. Yorke-Smith, “Localised reputation in the prisoner’s dilemma,” in *BNAIC/BeneLearn 2021*, 2021, pp. 761–763, 33rd Benelux Conference on Artificial Intelligence and 30th Belgian-Dutch Conference on Machine Learning, BNAIC/BeneLearn 2021; Conference date: 10-11-2021 Through 12-11-2021.

[274] L. Epstein and M. Schneider, “Ambiguity and asset markets,” 2010. [Online]. Available: https://EconPapers.repec.org/RePEc:nbr:nberwo:16181

[275] B. Hirchoua, B. Ouhbi, and B. Frikh, “Deep reinforcement learning based trading agents: Risk curiosity driven learning for financial rules-based policy,” *Expert Systems with Applications*, vol. 170, p. 114553, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420311970

[276] J. Rüdiger and A. Vigier, “Who acquires information in dealer markets?†,” *The American Economic Review*, vol. 110, no. 4, pp. pp. 1145–1176, 2020. [Online]. Available: https://www.jstor.org/stable/26921601

[277] A. A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzun, "Automation, Intermediation and the Flash Crash," *Journal of Investment Management*, vol. 16, no. 4, pp. 17–28, 2018.