FINMEM: A Performance-Enhanced LLM Trading Agent with Layered Memory and Character Design

Yangyang Yu * **Haohang Li** * **Zhi Chen** * **Yuechen Jiang** * **Yang Li** *
Denghui Zhang **Rong Liu,** **Jordan W. Suchow,** **Khaldoun Khashanah** †
Stevens Institute of Technology, Hoboken, NJ, United States
{yyu44, hli113, zchen100, yjiang52, yli269, dzhang42, rliu20, jws, kkhashan}@stevens.edu

Abstract

With the influx of diverse financial data streams from the web, traders face a deluge of information from various sources. This requires them rapidly to understand, memorize, and filtrate crucial events for investment decisions. However, innate cognitive limitations restrict human traders from processing information within their perception and memory capacity, a span much narrower than the actual volume of available information (Black 1986). Consequently, insufficiently considering or even dismissing critical events affecting trading decisions becomes increasingly concerning as data availability expands. To overcome the physical limitations in the memory systems of human traders, researchers have been consistently working on designing autonomous trading agent systems. These systems are expected to effectively integrate all available information and have a sophisticated backbone algorithm for enhanced trading performance.

The development of autonomous trading systems has progressed from initial rule-based strategies (Edwards, Magee, and Bassetti 2018) to advanced machine-learning algorithms (Huang et al. 2019). Recently, Reinforcement Learning (RL) agents, particularly those using Deep Reinforcement Learning (DRL) as their core algorithms (Millea 2021), have attracted attention in both academia and industry. Leveraging both RL principles and deep learning, DRL agents effectively handle and learn from scalable and diverse financial data, including stock prices, key financial indicators, and market sentiments. Research suggests DRLs can meet the crucial needs of trading agents to process and make informed decisions from large volumes of data. However, certain inherent features of DRL algorithms exhibit notable deficiencies in financial applications. **Firstly, DRL agents lack interpretability in the rationale behind their decision-making** (Balhara et al. 2022). **Secondly, integrating textual data with numerical features, crucial in finance, poses a challenge for DRL agents on convergence** (Gershman and Ölveczky 2020). Thus, a backbone algorithm that offers transparent reasoning and effectively captures investment-related textual insights is essential.

Recent advancements in Large Language Models (LLMs), like Generative Pre-trained Transformers (GPTs) (OpenAI 2023), have opened new avenues for developing trading agents, addressing past limitations. These LLM-based agents can articulate reasons and outcomes from their immediate observations. Their extensive pre-trained knowledge and ability to integrate diverse data sources, including textual and numerical information, allow them to transcend the constraints of isolated environments. This approach, when reinforced with well-designed prompt templates, markedly improves decision-making in various sectors (Wang et al. 2023). Notably, a growing body of research has focused on utilizing LLMs to make informed trading decisions for stocks (Yang, Liu, and Wang 2023; Wu et al. 2023). However, in currently available approaches, LLMs primarily serve as a question-answering (QA) role rather than functioning as autonomous agents. **The potential issue with these approaches is the incapability of fully understanding the varying timeliness associated with different types of financial data.** While these LLM agents outperform traditional trading benchmarks, their uniform processing in QA sequences leads to failure in remembering key messages. Additionally, their recognition of financial data timeliness depends heavily on the uncertain and laborious fine-tuning of LLMs. These shortfalls affect their daily knowledge updating capability, due to a lack of a memory component, thereby diminishing their effectiveness in prioritizing critical events.

To bridge this gap, we introduce FINMEM, a novel LLM-based autonomous trading agent. It excels in processing multi-source financial data through a layered memory module and adapts to market volatility by offering a self-adaptive character setting.

Our concept is initially inspired by the Generative Agents framework by Park et al., aimed at enhancing the efficient retrieval of key events for general-purpose LLM agents. This framework encompasses a unique character design and seed memory, activating the agent upon specific query through prompts. It prioritizes events in a unified memory stream, ranked by a linear combination of recency, relevancy, and importance. The framework (Park et al. 2023) lays the groundwork for LLM agent design, featuring a character profiling module, a memory module for recording and retrieving memories, and an action module for informed actions. This structure is instrumental in guiding the agent toward goal fulfillment in general social contexts. **However, Park et al.'s framework struggles with comprehending financial data with varying timeliness and importance**, like daily news versus quarterly and annual reports. Key challenges involve quantifying the distinct timeliness of data, optimizing information retrieval, and providing detailed reflections to improve the future decisions. To tackle these challenges, we further

*These authors contributed equally.
†Corresponding Author

propose FINMEM with the following improvements.

FINMEM maintains a modular approach similar to Park et al., but features novel design of profiling and memory modules. The profiling module of FINMEM provides a trading-task-specific professional background and includes a self-adaptive risk inclination feature, augmenting its robustness against market fluctuations. FINMEM's memory module innovatively incorporates working memory and layered long-term memory components, ideal for stratified information processing. Its working memory acts as a dynamic "workspace," enabling operations like summarization, observation, and reflection on information to facilitate trading decisions. Its long-term memory, composed of shallow, intermediate, and deep layers (Craik and Lockhart 1972), manages varied decay rates to retain different types of information across various timescales, aligning with their distinct timeliness. For instance, daily news, with its immediate effects on stock markets, is channeled into the shallow processing layer. Meanwhile, annual company reports, exerting a more prolonged impact, are processed in the deep layer by FINMEM. Each layer in FINMEM prioritizes memory events based on the assemble of recency, relevancy, and importance close to Park et al.'s method. However, it introduces new measurements for recency and importance, tailored to better allocate and prioritize data in the appropriate long-term memory layer. FINMEM's memory mechanism can also transit significantly impactful investment memory events to deeper processing layers, ensuring their retention for extended periods. FINMEM's memory module can mirror the human cognitive system (Sweller 2012) and facilitate agile, real-time decisions (Sun 2004). It enables continuous evolution in professional knowledge through structured summarizing, retrospecting experiences, and reacting to new trading scenarios. Additionally, FINMEM includes a decision-making module that synergizes top memories with current market conditions to derive investment decisions.

FINMEM provides three key contributions:

FINMEM presents a state-of-the-art LLM-based trading agent with a human-aligned memory mechanism and character design, particularly crafted to capture investment insights from the financial market. In its agent memory module design, FINMEM innovatively emulates human working and layered long-term memory mechanisms. This approach effectively harnesses the time-sensitive aspects of financial data, capturing crucial investment insights and thereby boosting trading performance. FINMEM's profiling module features dynamic character settings, enabling continuous updates in domain knowledge and risk preference to adapt to volatile trading environments. This enhances FINMEM's capabilities. Our experiments show that FINMEM evolves its knowledge base from past trades and ongoing market interactions, maintaining robustness in complex market conditions.

FINMEM can utilize its distinctive features to expand the agent's perceptual range beyond the human limitation to make well-informed trading decisions. Cognitive research suggests that human working memory is limited to recalling five to nine events at once (Miller 1956). While this avoids information overload, it may lead to insufficient insight for accurate decision-making. In contrast, FINMEM's memory module transcends this constraint. It allows adjusting cognitive load by selecting a flexible number of top-ranked events from each layer of its long-term memory, allowing FINMEM to deliver superior trading decisions in data-rich contexts.

FINMEM achieves impressive trading performance using training data that is limited in volume and spans a short time period. Experiments show that training FINMEM with a timeframe much shorter than that required by comparative models. This efficiency stems from optimally utilizing multi-source data and capturing critical trading signals. Notably, FINMEM is effective even on smaller datasets and with general-purpose LLMs, with its performance expected to enhance further with larger, higher-quality financial datasets and LLMs fine-tuned for financial applications.

References

Balhara, S.; Gupta, N.; Alkhayyat, A.; Bharti, I.; Malik, R. Q.; Mahmood, S. N.; and Abedi, F. 2022. A survey on deep reinforcement learning architectures, applications and emerging trends. *IET Communications*.

Black, F. 1986. Noise. *The journal of finance*, 41(3): 528–543.

Craik, F. I.; and Lockhart, R. S. 1972. Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, 11(6): 671–684.

Edwards, R. D.; Magee, J.; and Bassetti, W. C. 2018. *Technical analysis of stock trends*. CRC press.

Gershman, S. J.; and Ölveczky, B. P. 2020. The neurobiology of deep reinforcement learning. *Current Biology*, 30(11): R629–R632.

Huang, B.; Huan, Y.; Xu, L. D.; Zheng, L.; and Zou, Z. 2019. Automated trading systems statistical and machine learning methods and hardware implementation: a survey. *Enterprise Information Systems*, 13(1): 132–144.

Millea, A. 2021. Deep reinforcement learning for trading—A critical survey. *Data*, 6(11): 119.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2): 81.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701320.

Sun, R. 2004. Desiderata for cognitive architectures. *Philosophical psychology*, 17(3): 341–373.

Sweller, J. 2012. Human cognitive architecture: Why some instructional procedures work and others do not.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2023. A survey on large language model-based autonomous agents. *arXiv preprint arXiv:2308.11432*.

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

Yang, H.; Liu, X.-Y.; and Wang, C. D. 2023. FinGPT: Open-Source Financial Large Language Models. *arXiv preprint arXiv:2306.06031*.