Towards Sustainable Computation: Synergizing LLM Heat Recovery and Algorithmic Trading for Energy-Efficient AI Systems*

Ivan Letteri[1,*,†], Pierpaolo Vittorini[1,*,†] and Tamsir Jobe[1,†]

[1]University of L'Aquila, P.le S. Tommasi 1, 67100 Coppito, L'Aquila (Italy)

Abstract

The exponential growth of Large Language Models (LLMs) presents a dual challenge: unparalleled computational capabilities juxtaposed with unsustainable energy consumption and thermal waste. Concurrently, the financial sector's reliance on AI-driven algorithmic trading intensifies demands for computational efficiency. This paper introduces the AITA-LLM framework, a novel paradigm integrating waste heat recovery from LLM operations with advanced algorithmic trading systems. The framework repurposes recovered thermal energy for domestic heating, enhancing energy sustainability, while employing Retrieval-Augmented Generation (RAG)-enhanced financial LLMs (e.g., FinBERT) for real-time market analysis. A digital twin facilitates holistic simulation, optimising energy efficiency and trading performance. By harmonising AI's computational footprint with circular economy principles, this research advances greener computational finance, demonstrating that high-performance AI and environmental stewardship are mutually achievable.

Keywords

Large Language Models, Algorithmic Trading, Sustainable AI, Digital Twin, Retrieval-Augmented Generation

1. Introduction

The climate crisis and global energy conservation mandates necessitate urgent scrutiny of digital technologies' environmental impact [1]. Artificial Intelligence (AI), particularly Large Language Models (LLMs), exemplifies this challenge, with training and inference processes consuming gigawatt-hours of electricity and generating substantial waste heat [2, 3]. Current thermal management strategies prioritise cooling over reuse, squandering opportunities for energy valorisation. Simultaneously, AI-driven algorithmic trading systems—critical to modern finance—exploit LLMs for market analysis, further amplifying energy demands.

This paper addresses the dual imperative of advancing AI capabilities while mitigating environmental harm. We propose the AITA-LLM framework, which synergises LLM waste heat recovery with AI-enhanced trading. Our core hypothesis posits that heat generated during LLM-based financial analysis can be repurposed to offset energy costs, while optimised trading algorithms yield economic gains. Key contributions include: (i) A closed-loop system integrating LLM computation, thermal recovery, and financial decision-making. (ii) A couple of digital twins for simulating energy flows, thermal dynamics, and trading performance (i.e. AITA [4]). (iii) Empirical validation of energy savings and trading efficacy via a prototype leveraging NVIDIA Jetson Orin hardware.

This paper details the conceptualization and methodological approach of the AITA-LLM framework. We outline the state-of-the-art in LLM energy consumption, heat recovery technologies, and AI in algorithmic trading (Section 2). The proposed system architecture integrates LLM operations, thermal management, and an algorithmic trading engine (Section 3). The multi-faceted methodology encompasses computational heat recovery, digital twin implementation for simulation, LLM optimization for

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

* AITA-LLM project, aiming to synergize advancements in AI-driven finance with green computing principles for sustainability.

*Corresponding author: Ivan Letteri, Pierpaolo Vittorini

†These authors contributed equally.

ivan.letteri@univaq.it (I. Letteri); pierpaolo.vittorini@univaq.it (P. Vittorini); jtamsir7@gmail.com (T. Jobe)

0000-0002-3843-386X (I. Letteri); 0000-0002-6975-8958 (P. Vittorini); 0009-0001-6653-8239 (T. Jobe)

finance, and the development of the computational infrastructure (Section 4). We conclude by discussing potential challenges and future research directions that AITA-LLM aims to explore (Section 5). This work seeks to pave the way for a new generation of AI systems that are not only powerful but also environmentally responsible.

2. Related Work

The AITA-LLM project draws upon several distinct but interconnected fields of research: LLM energy efficiency, waste heat recovery, AI in algorithmic trading, and digital twin technology for complex system simulation.

The remarkable capabilities of LLMs like GPT-3 [5] come at a significant environmental cost. Training these models can consume GigaWatt-hours of electricity and result in substantial carbon emissions [3]. Inference, though less intensive per query, can accumulate to high energy usage at scale. This energy is primarily converted into heat by computational hardware (GPUs, TPUs). Efforts to mitigate this include model pruning, quantization, and developing more energy-efficient hardware [6], but the heat byproduct remains a challenge.

The concept of waste heat recovery (WHR) is well-established in industrial processes and, more recently, in data centers [7]. Data centers, which face similar thermal management challenges as LLM clusters, have explored liquid cooling and heat reuse for district heating or other applications. Some pioneering projects even use the heat from cryptocurrency mining operations for heating greenhouses or swimming pools [8]. However, specific research on direct heat recovery from dedicated LLM fine-tuning/inference hardware for localized applications like domestic heating is less common.

Algorithmic trading has evolved from rule-based systems to sophisticated machine learning (ML) models [9]. Deep learning, reinforcement learning, and NLP techniques are increasingly used for market prediction, sentiment analysis, and portfolio optimization [10]. Financial LLMs, such as FinBERT [11] and BloombergGPT [12], are designed to understand the nuances of financial language, offering new capabilities for extracting insights from textual data. The AITA-LLM project specifically considers integrating FinBERT, given its open-source nature and proven efficacy in financial sentiment analysis [13]. The project also references the AITA framework for AI agent orchestration [4].

To keep LLMs updated with current information and reduce hallucinations, Retrieval-Augmented Generation (RAG) has emerged as a powerful technique [14]. RAG combines pre-trained LLMs with external knowledge retrieval systems. In finance, this can involve fetching real-time market data, news articles, or company reports to inform the LLM's analysis and predictions. Frameworks like LangChain [15] and vector databases like ChromaDB facilitate the implementation of RAG systems.

Digital Twins (DTs) are virtual replicas of physical assets, processes, or systems, used for simulation, monitoring, and optimization [16]. In the energy sector, DTs model power grids or renewable energy plants [17]. For complex systems like the proposed AITA-LLM, a DT can help analyze the interplay between computational workloads, heat generation, thermal transfer, energy savings, and trading performance under various scenarios before physical implementation. Tools like EnergyPlus [18] for building energy simulation, SimPy [19] for discrete-event process simulation, and PyBullet [20] for physics simulation are relevant.

3. Theoretical Framework

The AITA-LLM framework is designed as a modular, integrated system that harmonizes computational finance operations with energy recovery and sustainable practices. Figure 1 provides a high-level scenario overview.

The core components of the AITA-LLM architecture are:

1. **LLM Computational Core**: This unit is responsible for the fine-tuning and inference of financial LLMs. It will be built around low-power, high-performance hardware such as the NVIDIA Jetson

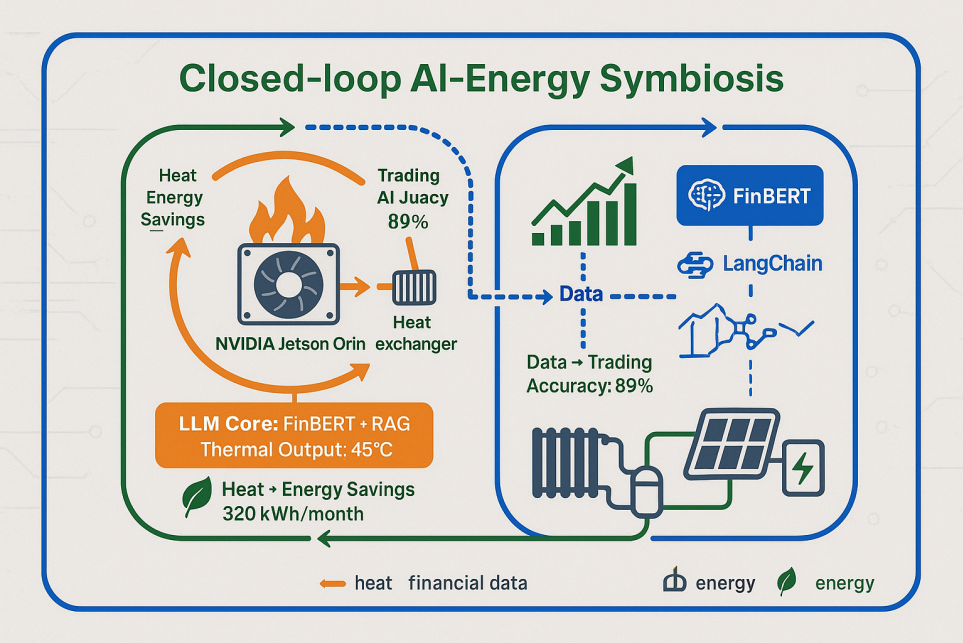

Figure 1: AI–energy symbiosis simulation: heat from an LLM FinBERT (Jetson Orin, 45 °C) warms interiors via a heat exchanger, powers a solar installation and fuels a trading algorithm (89 % accuracy), saving 320 kWh per month.

Orin [21], chosen for its balance of AI processing capabilities and manageable thermal design power. This core executes tasks related to financial data analysis, sentiment extraction, and trade signal generation.

2. **Thermal Management and Heat Recovery System**: This subsystem is engineered to capture the waste heat generated by the LLM Computational Core. It involves: Direct heat capture mechanisms (e.g., specialized heat sinks, liquid cooling loops) integrated with the computational hardware. A heat transfer unit (e.g., heat exchanger, heat pump) to channel the recovered thermal energy to an ambient heating system (e.g., domestic radiators, underfloor heating, hot water storage). Intelligent sensors and actuators to monitor temperatures, flow rates, and energy transfer, enabling dynamic adjustments for optimal efficiency.
3. **Algorithmic Trading Engine**: This software engine implements the financial trading strategies. Its key elements include: **Financial LLM (FinBERT-based)**, a pre-trained model on financial corpora and fine-tuned for specific tasks like market trend prediction from news or social media sentiment analysis. **Retrieval-Augmented Generation (RAG) Module** like LangChain and ChromaDB to provide the LLM with up-to-date financial data, news, and market analyses, enhancing the accuracy and timeliness of its outputs. **AITA-Framework Integration**[4] for creating, training, orchestrating AI agents, and managing trade execution via regulated broker APIs.
4. **Digital Twin Simulation Environment**: A virtual counterpart of the entire AITA-LLM system, used for: Modeling and simulating the energy flow, heat generation, and thermal transfer dynamics. Scenario analysis and "what-if" studies before physical deployment or modifications using tools like SimPy, PyBullet, and EnergyPlus.
5. **Data Ingestion and Processing Layer**: Responsible for collecting and pre-processing diverse data streams: Real-time and historical financial market data (prices, volumes) from broker APIs and financial data providers. Textual data (news articles, analyst reports, social media feeds) for NLP by the LLM.
6. **Energy Supply and Monitoring Unit**: While the primary focus is heat recovery, the system is designed to be modular regarding its primary energy source. The project envisions integration with renewable sources like solar panels (as depicted in Figure 1) to further enhance its green

credentials. This unit monitors overall energy consumption and production.

4. Methodology

4.1. Computational Heat Recovery and Thermal Integration

The initial phase of the methodology is dedicated to the practical aspects of quantifying and capturing the waste heat generated by the chosen LLM hardware, specifically the NVIDIA Jetson Orin, during typical financial workloads such as model fine-tuning and continuous inference. Thermal sensors will meticulously measure the heat output under varying computational loads that are representative of LLM fine-tuning processes and inference tasks pertinent to financial analysis. Simultaneously, power consumption will be closely monitored to establish a baseline for energy usage. Following this characterization, the **Thermal Transfer System Design** will commence. Based on the quantified heat output, a prototype thermal transfer system will be engineered. Design considerations may include custom air-cooling ducts to direct hot air towards a heat exchanger, or alternatively, a small-scale liquid cooling loop designed to transfer heat efficiently to a water circuit. The overarching goal is to channel the recovered thermal energy effectively to a secondary medium that can be used for space heating, such as a small radiator or as a pre-heating stage for a domestic hot water system. A critical step will be the **Energy Balance Calibration**. Techniques for energy optimization will be explored, including intelligent workload scheduling to align computational activity with heating demand, or dynamically adjusting fan and pump speeds to maximize efficiency. Finally, **Sensor Integration** will see the incorporation of intelligent sensors (monitoring temperature, flow rates, and power) and actuators (such as variable speed fans or pumps). These components will be integrated to monitor and dynamically regulate thermal dissipation and recovery processes in real-time, typically controlled via a microcontroller like a Raspberry Pi or ESP32, ensuring responsive and optimized system operation.

4.2. Digital Twin Implementation

To facilitate comprehensive performance analysis and system optimization prior to full-scale physical implementation, a digital twin of the entire AITA-LLM system will be developed. This virtual replica will model and simulate the system's behavior under a wide range of conditions. The development involves several stages of **Component Modeling**. The **Computational Unit Model** will utilize SimPy to simulate the discrete processes of LLM task execution, algorithmic trading decisions, and their associated energy consumption profiles, which will be derived from empirical measurements obtained in the previous phase. Concurrently, a **Thermal Model** will be developed using PyBullet or a similar physics engine. This model will simulate the physical environment, the heat generation dynamics of the Jetson Orin, and the subsequent heat transfer to the recovery system, potentially incorporating basic Computational Fluid Dynamics (CFD) principles or simplified heat transfer equations. Furthermore, a **Building/Environment Model** will be created using EnergyPlus, integrated with Python via the 'eppy' library. This will model the target thermal zone (e.g., a room or small dwelling) to simulate its heating load, assess the impact of the recovered heat on ambient temperature, and calculate the overall energy balance. Once individual models are developed, **Integration and Co-simulation** will be performed. The component models will be interconnected to allow for co-simulation, where data is exchanged between SimPy (handling process logic and AI decisions), PyBullet (managing physics and thermal transfer), and EnergyPlus (simulating building energy performance). This integrated approach will capture the coupled dynamics of the complete system. The completed digital twin will then be utilized for **Scenario Analysis and Optimization**. It will enable the execution of various simulated scenarios, including different trading activity levels, fluctuating ambient temperatures, diverse heating demands, and alternative hardware configurations.

4.3. LLM Optimization for Algorithmic Trading

This stream of the methodology focuses on the development and refinement of the Artificial Intelligence core that drives the trading engine. The process begins with **Base Model Selection and Adaptation**. FinBERT [11], a BERT-based model pre-trained on extensive financial text corpora, will serve as the foundational LLM. This model will undergo further fine-tuning using domain-specific datasets, such as archives of financial news and sentiment-annotated financial texts, to tailor its capabilities for specific tasks like predicting market sentiment or identifying actionable trading signals from textual data. To ensure efficient fine-tuning on the resource-constrained Jetson Orin platform, techniques such as LoRA (Low-Rank Adaptation) or QLoRA [22] will be investigated and implemented. A key enhancement will be the **Retrieval-Augmented Generation (RAG) Integration**. A RAG system will be constructed, likely leveraging tools like LangChain and a vector database such as ChromaDB. This system will be designed to retrieve relevant, up-to-date financial documents—including news articles, company reports, and summaries of price data—in real-time. This contextual information will be fed to the FinBERT model, thereby improving the accuracy, relevance, and timeliness of its analytical outputs and subsequent recommendations. The outputs from the RAG-enhanced LLM, which may include sentiment scores, trend likelihoods, or identified market anomalies, will then be translated into actionable trading signals during the **Signal Generation and Strategy Development** phase. These signals will be integrated into comprehensive trading strategies developed within the AITA-framework [4]. This framework supports agent-based modeling and provides robust tools for strategy backtesting and validation. Finally, rigorous **Performance Evaluation** of the developed trading strategies is paramount. Strategies will be thoroughly backtested on historical market data. Following successful backtesting, they will be paper-traded in a simulated live market environment to evaluate their real-world performance characteristics, including profitability, risk metrics (e.g., Sharpe ratio, drawdown), and robustness, before any consideration of real-capital deployment.

5. Conclusion

The AITA-LLM project addresses the critical intersection of escalating AI energy demands and the need for sustainable financial technologies. By proposing an innovative framework that integrates the recovery of waste heat from Large Language Model operations with an advanced algorithmic trading system, this research charts a course towards more environmentally responsible and economically efficient computational finance. The core contributions lie in the synergistic design: repurposing LLM-generated heat for practical applications like domestic heating, thereby reducing overall energy consumption, while simultaneously leveraging the power of financial LLMs (FinBERT enhanced with RAG) for sophisticated market analysis and trading. The development of comprehensive digital twins will provide an invaluable tool for simulating, analyzing, and optimizing this complex interplay of AI workloads, thermal dynamics, energy flows, and financial performance.

While challenges related to small-scale heat recovery efficiency, edge AI performance, and market volatility are acknowledged, the proposed methodology provides a robust plan to address them. The expected impact is significant, offering a tangible pathway for the financial industry to reduce their carbon footprint, valorize energy byproducts, and advance the capabilities of AI in finance. AITA-LLM aims to demonstrate that high-performance AI and environmental sustainability are not mutually exclusive. By fostering a circular economy approach within computational systems, where outputs of one process become valuable inputs for another, we can move towards a future where technological advancement aligns more closely with planetary well-being.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content

as needed and took full responsibility for the publication's content.

References

[1] Y. Li, X. Yang, Q. Ran, H. Wu, M. Irfan, M. Ahmad, Energy structure, digital economy, and carbon emissions: evidence from China, Environmental Science and Pollution Research 28 (2021) 64606–64629. URL: https://link.springer.com/10.1007/s11356-021-15304-4. doi:10.1007/s11356-021-15304-4.

[2] E. Strubell, A. Ganesh, A. McCallum, Energy and Policy Considerations for Modern Deep Learning Research, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 13693–13696. doi:10.1609/aaai.v34i09.7123.

[3] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, J. Dean, Carbon emissions and large neural network training, arXiv preprint arXiv:2104.10350 (2021).

[4] I. Letteri, AITA: A new framework for trading forward testing with an artificial intelligence engine, in: Proceedings of the Italia Intelligenza Artificiale (Ital IA 2023), Pisa, Italy, May 29-30, 2023, volume 3486 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 506–511. URL: https://ceur-ws.org/Vol-3486/paper-05.pdf.

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

[6] R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni, Green ai, Communications of the ACM 63 (2020) 54–63.

[7] R. K. Shah, J. E. Hesselgreaves, Waste heat recovery technologies and applications, Applied Thermal Engineering 28 (2008) 737–738.

[8] P. Gao, L. Zhang, X. Zhu, Bitcoin's energy consumption: A market-driven approach, Energy Economics 100 (2021) 105357.

[9] E. Frank, Algorithmic & high-frequency trading, EasyChair Preprint 13951, EasyChair, 2024.

[10] T. Fischer, C. Krauss, Deep learning with long short-term memory networks for financial market predictions, European Journal of Operational Research 270 (2018) 654–669.

[11] D. Araci, Finbert: Financial sentiment analysis with pre-trained language models, arXiv preprint arXiv:1908.10063 (2019).

[12] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, G. Mann, BloombergGPT: A large language model for finance, arXiv preprint arXiv:2303.17564 (2023).

[13] L. Yang, I. Lunesu, G. Mulas, M. Marchesi, Finbert: A pre-trained language model for financial communications, arXiv preprint arXiv:2006.08097 (2020).

[14] P. Lewis, E. Pérez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, U. Khandelwal, et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 9459–9474.

[15] H. Chase, Langchain, https://github.com/hwchase17/langchain, 2022.

[16] N. Julien, E. Martin, How to characterize a digital twin: A usage-driven classification, IFAC-PapersOnLine 54 (2021) 894–899. URL: https://www.sciencedirect.com/science/article/pii/S2405896321008557. doi:https://doi.org/10.1016/j.ifacol.2021.08.106, 17th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2021.

[17] A. Rasheed, O. San, T. Kvamsdal, Digital twin: A review on enabling technologies, challenges, and future research directions, Engineering Science and Technology, an International Journal 23 (2020) 1189–1210.

[18] U.S. Department of Energy, EnergyPlus Engineering Reference, https://energyplus.net/, 2023.

[19] T. SimPy, Simpy: Discrete-event simulation for python, 2023.

[20] E. Coumans, Y. Bai, Pybullet, a python module for physics simulation for games, robotics and machine learning, http://pybullet.org, 2016–2021.

[21] NVIDIA, Jetson generative ai supercomputer, https://blogs.nvidia.com/blog/jetson-generative-ai-supercomputer/, 2023.

[22] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: efficient finetuning of quantized llms, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.