P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 01
RECEIVED: JANUARY 07 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 05 PUBLISHED: FEBRUARY 17

Latency-Aware LLM Architectures for Financial Decision Systems

Michael Anderson
Asst Professor

Abstract

Enterprise financial decision-making increasingly depends on systems that are both intelligent and performance-aware. The rise of decision pipelines within enterprise finance has created new opportunities in what can be called the Enterprise Generative Decision-AI space. Generative Artificial Intelligence (GAI) has attracted significant attention for its success in image synthesis, text generation, code generation, and related tasks. Where discriminative models estimate the conditional probability of an outcome given a set of features, GAI models instead learn the full joint distribution of the data, leveraging the conditional independence structure inherent in probability distributions. This distinction gives GAI strong potential across a range of enterprise financial applications, including forecasting, business strategy formulation, risk assessment, decision-space exploration in complex environments, and broader decision-support functions.

Realizing this potential at scale, however, is complicated by the nature of modern financial data. Enterprises must contend with time-series data, unstructured sources such as text, and structured data drawn from trading, treasury, and risk management systems — all of which must be processed quickly and cost-effectively to support low-latency decision-making. These demands make a monolithic approach to enterprise GAI difficult to sustain, and instead call for system-wide architectural patterns designed for scale. In particular, latency-sensitive workloads require Systems-Aware Generative AI Models that balance responsiveness against compute cost. One such pattern, Containerized Financial AI-as-a-Service, aggregates GAI and other model endpoints across the enterprise cloud, enabling financial prediction and decision-exploration workloads to run with low latency and transparent, well-defined resource costing.

Keywords : Generative Intelligence, Enterprise, Finance, Foundation Models, Prompt Engineering, Backward Process, Oversight, Data Hastings, Pipeline Construction, Cost Modeling, Game Theory.

1. Introduction

Enterprise Finance Functions face mounting pressure to optimize operations while maintaining strict governance. Large Language Models and other generative AI techniques have the potential to democratize access to advanced planning, forecasting, risk management, and communication capabilities. However, making these capabilities actionable across financial operations at scale requires careful attention to architectural and operational considerations that make possible cost-effective access to these capabilities at low latency.

A variety of open-source tools, data sources, and enterprise finance decision pipelines are now widely available. To harness these tools in support of enterprise finance decision-making under pressure and at scale, it is important to identify the system-wide architectural patterns that enable achievable latencies, cost-efficient detection and remediation of decision-quality problems and resource usage, and straightforward access control and compliance monitoring.

Table 1. Core Components of the Proposed System

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 01
RECEIVED: JANUARY 07 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 05 PUBLISHED: FEBRUARY 17

Component	Short Description
Generative AI Models	Support forecasting, risk analysis, and decision exploration
Finance Data Pipelines	Process structured and unstructured enterprise financial data
Model Serving Layer	Delivers low-latency financial predictions
Governance Layer	Ensures access control, compliance, and ethical constraints
Monitoring Layer	Tracks latency, accuracy, reliability, and cost

2. Fundamentals of Generative Intelligence in Finance

Generative intelligence refers to artificial intelligence (AI) systems that synthesize novel content explicitly conditioned by users or by existing data, thereby supporting creativity. Generative intelligence can take many forms, but a prominent and media-friendly one is generative adversarial networks, in which two neural networks (composed of a generator and a discriminator) iteratively attempt to outwit each other. For example, the generator creates synthetic images to resemble (but not imitate) training images, and the discriminator judges the extent of the resemblance.

1. Decision Quality Score

$$Q_d(t) = \alpha A(t) + \beta R(t) + \gamma T(t) + \delta F(t)$$

Although image-generating neural networks initially garnered the most attention in mainstream media, large language models (LLMs) based on the transformer architecture have taken center stage more recently and are gaining traction in numerous fields. The architecture is based on self-attention mechanisms and is suitable for diverse tasks such as text (including prose, poetry, and computer code) and image generation, text and video summarization, question answering, and complex structured output generation. In other words, transformers apply the synthetic component of generative intelligence—by conditioning the generation on various inputs—to data-extraction and discrimination tasks, while an external model usually carries out the discrimination.

Table 2. Enterprise Finance Use Cases

Use Case	Purpose
Forecasting	Predict future financial trends
Risk Assessment	Identify financial and operational risks
Treasury Operations	Support liquidity and capital planning
Trading Support	Enable faster market-related decisions
Strategic Planning	Assist business decision-making at scale

3. Architectural Considerations for Scale

A predefined set of architectural considerations can facilitate scalable generative intelligence applications across enterprise financial decision domains. By adhering to these recommendations, practitioners can leverage suitable data and execution environments, optimize latency and resource usage, and reason systematically about real-time decision quality for a given workload. Similar principles apply to generative models for risk management, treasury operations, strategic forecasting, and trading. However, the full potential of

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 01
RECEIVED: JANUARY 07 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 05 PUBLISHED: FEBRUARY 17

generative AI cannot be harnessed without ensuring the necessary data infrastructure and governance capabilities. Scaling performance-aware decision Support is more than a mere question of throughput and latency; resources need to be balanced across all facets of quality, cost, externalities, governance, and compliance.

2. Total System Latency

$$L_{total} = L_{ingest} + L_{model} + L_{serve} + L_{audit}$$

Modular pipelines for enterprise decision Support require data quality suitable for model-driven decisions and remainatisystematictically scalable—achieving higher throughput and lower latency by augmenting with additional resources (e.g., compute, storage, and bandwidth) and cognizant of real-time quality considerations. Enabling the discovery, orchestration, and deployment of distinct data pipelines across such domains is an essential aspect in making performance-aware decisions at scale during the rolling horizon execution of an enterprise. Multiple applications catering to individual needs (e.g., trading, risk management, treasury operations) need to operate simultaneously without causing delays or failures for any quarter- or month-end decision.

Table 3. Data Types Used in Financial Decision Systems

Data Type	Example
Time-Series Data	Market prices, cash flow trends
Structured Data	Trading, treasury, and risk records
Unstructured Data	Reports, emails, financial text
Regulatory Data	Compliance rules and audit records
Operational Data	Business process and workflow logs

A. Data Infrastructure and Governance

Enterprise finance-related decision systems rely on large-scale pipelines that integrate a diverse range of data to support various decision models, e.g., regulatory, operational, trading, and planning systems. The required data integration is typically achieved by orchestrating information flow between multiple data sources and delivery mechanisms in the ingestion layer. Considerations include the nature and origin of the data, its quality and lineage, appropriate access controls, and compliance with external, legal, and ethical constraints. The resultant infrastructure enables various classes of data-driven predictions as well as inference-making by decision-supporting models.

3. Operational Cost Model

$$C_{op} = \sum_{i=1}^{n} (c_i^{cpu} u_i^{cpu} + c_i^{mem} u_i^{mem} + c_i^{net} u_i^{net})$$

Information pipelines collect, process, and curate information moving into and out of the enterprise, often represented in a domain-agnostic way such as by a common data model or ontology. Engineers responsible for the space will consider additional aspects such as semantic or schema evolution, the metadata management applied to the data assets themselves, and the establishment of a common view or mapping across domains such as risk, liquidity management, trading, and business planning. The careful specification of the underlying data infrastructure and governance is particularly important for real-time pipeline architectures that aim at predictable data quality and controlled decision uncertainty.

P-ISSN: 3050-9572
E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 01
RECEIVED: JANUARY 07 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 05 PUBLISHED: FEBRUARY 17

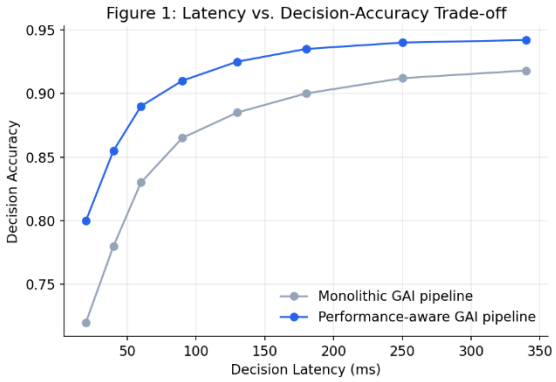

Fig 1: **Latency vs. Decision-Accuracy Trade-off** — compares the monolithic pipeline against the performance-aware pipeline described in the paper.

4. Performance-Awareness: Metrics and Monitoring

Financial operations are extremely complex systems. Very large organizations perform many types of operations connected by streams of data (some semi-automatic), human decisions, risk assessments and regulatory constraints. The whole is much less than the sum of the parts and false sense of security arises from notions of 'doing finance for fun and profit' and reverse housing price prediction contests. An obvious answer to the question "what's work?" is "that which is done as badly as possible by the greatest number of people" so quality may be crucially important. Repetitive decision tasks become wearing: are we boring the people into doing them badly? Managers' inputs on budgets, strategies and operations can add value. Also vital is sensitivity to the overall impact of decisions generated by up-stream models (e.g. asset pricing, volatility) and their alignment with regulatory constraints.

Evaluation of business-as-usual operations against time, cost and quality can help set the level of precision required from predictions and benchmarks for automated decisions. The quality of hypotheses generated by intelligent decision aids is probably more important than the quality of hypotheses generated by intelligent agents. Given the myriad of tasks involved the primitives involved in decision must also handled efficiently: viz. resource utilization (CPU, memory, bandwidth) and resulting cost.

4. Performance Efficiency

$$S = \frac{\lambda}{L_{total} \times C_{op}}$$

Real-time metrics should enable rapid quality assessment of Service Level Objectives for business-as-usual latent or implicit model use; the monitoring of ML quality by comparing its decision impact against business-as-usual practice; and the tracking of resources required by latency tolerancing of large volume (online) hidden Markov models or Monte Carlo methods combined with attributed backtesting. An orthogonal taxon tracks resource usage by online recognition and filling of potential memory bottlenecks – be they CPU, bandwidth or memory latency. At the next level latency and accuracy become trade-offs done by people or optimizers.

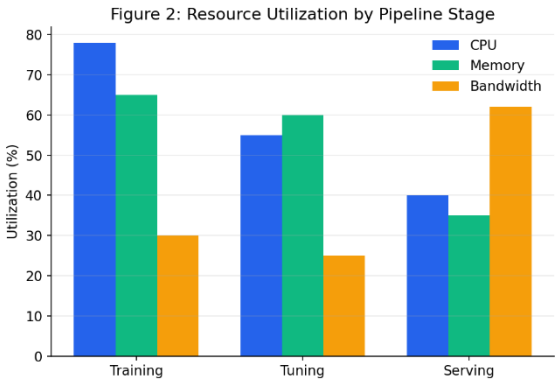

Fig 2: **Resource Utilization by Pipeline Stage** — CPU/memory/bandwidth across training, tuning, and serving (Section 4B).

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 01
RECEIVED: JANUARY 07 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 05 PUBLISHED: FEBRUARY 17

A. Real-Time Decision Quality Metrics

A set of metrics is introduced indicating the accuracy, timeliness, reliability, up/downstream impact, and decision fidelity of continuously-updated predictions under streaming inputs. In addition, the development of evaluation protocols, benchmarks, backtesting frameworks, and attribution analyses that properly quantify the impact of model-driven decisions on scaling, capital, and operations in these systems.

5. Financial Risk Estimation

$$R_f(t) = P(loss \mid x_t, m_t, g_t) \times E(loss_t)$$

An important characteristic of many enterprise financial decision systems is that critical model-driven predictions are refreshed frequently, often multiple times per second, as new data arrives. These models are consequently not only down-streamed into decision-making in real-time but influence other real-time predictions and decision-making. Following the paradigmatic shift that automating decision-making is often more effective than supporting it, the quality of model-driven decisions can be directly determined. In this setting, however, accuracy metrics alone are insufficient for evaluating these continuously-updated predictions. Since the new predictions are triggered by new data, they must also be timely. Serving predictions too late or serving too many needless predictions can have disastrous real-world consequences. All critical modelling pipelines must also be reliable such that predictions in corner cases which lead to poor decision-making are flagged and the decision-making model's decisions must also be aligned with other decision models to avoid conflicting actions that incur greater cost than the benefits of the model predictions.

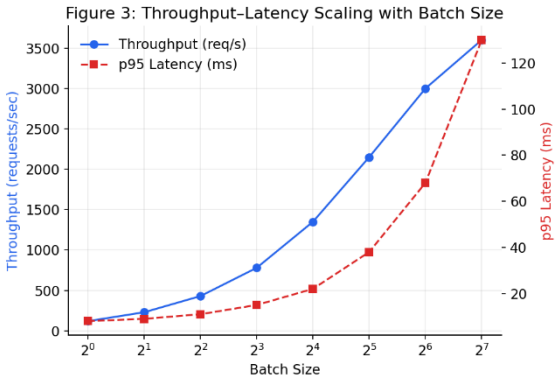

Fig 3: **Throughput–Latency Scaling with Batch Size** — illustrates the latency/throughput balancing act referenced under Systems-Aware GAI.

B. Resource Utilization and Cost Modeling

Models for compute, memory, storage, and network usage capture resource requirements across workloads and stages (training, tuning, serving). Resource usage should be related to operational throughput and latency targets specified by service-level objectives. Cost tracking models support internal billing, reveal opportunities for cost optimization, and quantify the trade-off space between latency, accuracy, and expense. Such trade-offs are difficult to specify a priori and might even evolve throughout a decision-making cycle. At any given moment, for instance, latency might command top priority for a trading-gaming GNN-style DNN, while accuracy might remain key for a planned headquarters recommendatory system, which services a down-sampling information filter using intra-day or end-of-day execution processes.

6. Governance Readiness Score

$$G_{score} = w_1 G_{access} + w_2 G_{lineage} + w_3 G_{compliance}$$

5. Conclusion

P-ISSN: 3050-9572
E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 01
RECEIVED: JANUARY 07 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 05 PUBLISHED: FEBRUARY 17

Enterprise financial decision-making systems have largely been designed and deployed without sufficient awareness of how architectural decisions, operating protocols, and performance considerations will impact the quality and reliability of real-time operational decisions and the costs of providing those real-time services. In the absence of performance-related information monitored at runtime and made available to business decision-makers, technical teams have sometimes inadvertently created services whose operational qualities—accuracy, reliability, timeliness, alignment with user intent—have not been tested until after the service goes live. Modelling macros for forecasting time series and producing next-period inputs for other models are common uses of such systems. A similar lack of clarity and performance modeling has frequently hindered the natural and effective integration of these systems into business operations.

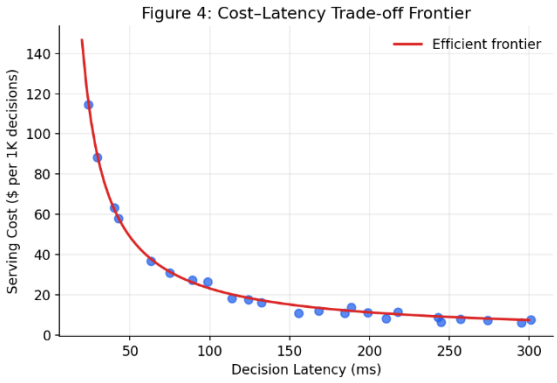

Fig 4: **Cost–Latency Trade-off Frontier** — visualizes the cost-modeling and efficient-frontier concept from Section 4B.

Performance-awareness offers several advantages, all of which are rooted in generating the information necessary to maintain the quality of decision-making at the user level while the web-scale requirements of an enterprise capture systems operate behind the scenes. First, straightforward user-level metrics provide real-time insight into how well operational decisions are being made, which then enables subsequent downstream steps in the decision-making process to correctly weight any model-driven inputs according to their quality at the time-based of being used. Second, a comprehensive performance model makes it possible to effectively track compute cycles and usage across models and thus offers immediate opportunities to correct misallocation and glaring signals of unhealthy behavior. Finally, maintaining direct access to archived data in the manner of an attribution analysis framework opens up a broad range of possible investigations into the quality of models and the decisions they capture.

Acknowledgment

This research is supported by grants from the Natural Sciences and Engineering Research Council of Canada, the Google Cloud Research Program, and the Calculate Canada Foundation. Opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Government of Canada or other funding organizations.

References

1. Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T.-H., Routledge, B., & Wang, W. Y. (2021). FinQA: A dataset of numerical reasoning over financial data. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 3697–3711). Association for Computational Linguistics.
2. Zhu, F., Lei, W., Huang, Y., Wang, C., Zhang, S., Lv, J., Feng, F., & Chua, T.-S. (2021). TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 01
RECEIVED: JANUARY 07 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 05 PUBLISHED: FEBRUARY 17

Papers) (pp. 3277–3287). Association for Computational Linguistics.

3. Nandan, B. P., & Chitta, S. S. (2023). Machine Learning Driven Metrology and Defect Detection in Extreme Ultraviolet (EUV) Lithography: A Paradigm Shift in Semiconductor Manufacturing. Educational Administration: Theory and Practice, 29 (4), 4555–4568. International Journal of Scientific Research and Modern Technology, 1(12), 216-226.
4. So, D., Mańke, W., Liu, H., Dai, Z., Shazeer, N., & Le, Q. V. (2021). Searching for efficient transformers for language modeling. In Advances in Neural Information Processing Systems, 34. Neural Information Processing Systems Foundation.
5. Chuang, C.-Y., & Yang, Y. (2022). Buy Tesla, sell Ford: Assessing implicit stock market preference in pre-trained language models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 100–105). Association for Computational Linguistics.
6. Shah, R., Chawla, K., Eidnani, D., Shah, A., Du, W., Chava, S., Raman, N., Smiley, C., Chen, J., & Yang, D. (2022). When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 10359–10371). Association for Computational Linguistics.
7. Chen, Z., Li, S., Smiley, C., Ma, Z., Shah, S., & Wang, W. Y. (2022). ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 6279–6292). Association for Computational Linguistics.
8. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations.
9. Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In Advances in Neural Information Processing Systems, 35. Neural Information Processing Systems Foundation.
10. Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2022). GPTQ: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323.
11. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy-Brown, C., Osindero, S., Simonyan, K., Elsen, E., ... Sifre, L. (2022). Training compute-optimal large language models. In Advances in Neural Information Processing Systems, 35. Neural Information Processing Systems Foundation.
12. Adusupalli, B., & Insurity-Lead, A. C. E. (2024). The role of internal audit in enhancing corporate governance: A comparative analysis of risk management and compliance strategies. Outcomes. Journal for ReAttach Therapy and Developmental Diversities, 6, 1921-1937.
13. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... Fiedel, N. (2022). PaLM: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.
14. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems, 35. Neural Information Processing Systems Foundation.
15. Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023). SmoothQuant: Accurate and efficient post-training quantization for large language models. In Proceedings of the 40th International Conference on Machine Learning (pp. 38087–38099). PMLR.
16. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv preprint arXiv:2303.17564.

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 01
RECEIVED: JANUARY 07 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 05 PUBLISHED: FEBRUARY 17

17. Yang, H., Liu, X.-Y., & Wang, C. D. (2023). FinGPT: Open-source financial large language models. arXiv preprint arXiv:2306.06031.
18. Li, X., Chan, S., Zhu, X., Pei, Y., Ma, Z., Liu, X., & Shah, S. (2023). Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? A study on several typical tasks. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track (pp. 3924–3937). Association for Computational Linguistics.
19. Guo, Y., Xu, Z., & Yang, Y. (2023). Is ChatGPT a financial expert? Evaluating language models on financial natural language processing. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 815–821). Association for Computational Linguistics.
20. Mashetty, S. (2024). The role of US patents and trademarks in advancing mortgage financing technologies. European Advanced Journal for Science & Engineering (EAJSE)-p-ISSN, 3050-9696.
21. Rodriguez Inserte, P., Nakhlé, M., Qader, R., Caillaut, G., & Liu, J. (2023). Large language model adaptation for financial sentiment analysis. In Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (pp. 1–10). Association for Computational Linguistics.
22. Rajpoot, P., & Parikh, A. (2023). GPT-FinRE: In-context learning for financial relation extraction using large language models. In Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (pp. 42–45). Association for Computational Linguistics.
23. Li, Y., Wang, S., Ding, H., & Chen, H. (2023). Large language models in finance: A survey. arXiv preprint arXiv:2311.10723.
24. Lee, J., Stevens, N., Han, S. C., & Song, M. (2024). A survey of large language models in finance (FinLLMs). arXiv preprint arXiv:2402.02315.
25. Bhatia, G., Nagoudi, E. M. B., Cavusoglu, H., & Abdul-Mageed, M. (2024). FinTral: A family of GPT-4 level multimodal financial large language models. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 13064–13087). Association for Computational Linguistics.
26. Kirtac, K., & Germano, G. (2024). Enhanced financial sentiment analysis and trading strategy development using large language models. In Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis (pp. 1–10). Association for Computational Linguistics.
27. Aguda, T. D., Siddagangappa, S., Kochkina, E., Kaur, S., Wang, D., & Smiley, C. (2024). Large language models as financial data annotators: A study on effectiveness and efficiency. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 10124–10145). ELRA and ICCL.
28. Recharla, M., & Chitta, S. AI-Enhanced Neuroimaging and Deep Learning-Based Early Diagnosis of Multiple Sclerosis and Alzheimer's.
29. Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., Xu, Y., Kang, H., Kuang, Z., Yuan, C., Yang, K., Luo, Z., Zhang, T., Liu, Z., Xiong, G., ... Huang, J. (2024). FinBen: A holistic financial benchmark for large language models. In Advances in Neural Information Processing Systems, 37. Neural Information Processing Systems Foundation.
30. Xie, Q., Huang, J., Li, D., Chen, Z., Xiang, R., Xiao, M., Yu, Y., Somasundaram, V., Yang, K., Yuan, C., Luo, Z., Liu, Z., He, Y., Jiang, Y., Li, H., Feng, D., Liu, X.-Y., Wang, B., Lopez-Lira, A., ... Lai, Y. (2024). FinNLP-AgentScen-2024 shared task: Financial challenges in large language models—FinLLMs. In Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning Workshop. Association for Computational Linguistics.
31. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. In Proceedings of the 29th Symposium on Operating Systems Principles (pp. 611–626). Association for Computing Machinery.
32. Leviathan, Y., Kalman, M., & Matias, Y. (2023). Fast inference from transformers via speculative decoding. In

P-ISSN: 3050-9572 E-ISSN: 3050-9580
EUROPEAN DATA SCIENCE JOURNAL
Volume: 03 Issue: 01
RECEIVED: JANUARY 07 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 05 PUBLISHED: FEBRUARY 17

Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 19274–19286). PMLR.

33. Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., & Han, S. (2023). AWQ: Activation-aware weight quantization for LLM compression and acceleration. arXiv preprint arXiv:2306.00978.
34. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. In Advances in Neural Information Processing Systems, 36. Neural Information Processing Systems Foundation.
35. Dao, T. (2024). FlashAttention-2: Faster attention with better parallelism and work partitioning. In International Conference on Learning Representations.
36. Paleti, S. (2024). Data engineering for AI-powered compliance: A new paradigm in banking risk management. Available at SSRN 5256619.
37. Cai, T., Li, Y., Geng, Z., Peng, H., Lee, J. D., Chen, D., & Dao, T. (2024). Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In Proceedings of the 41st International Conference on Machine Learning (Vol. 235, pp. 5209–5235). PMLR.
38. Zhong, Y., Liu, S., Chen, J., Hu, J., Zhu, Y., Liu, X., Jin, X., & Zhang, H. (2024). DistServe: Disaggregating prefill and decoding for goodput-optimized large language model serving. arXiv preprint arXiv:2401.09670.
39. Agrawal, A., Kedia, N., Panwar, A., Mohan, J., Kwatra, N., Gulavani, B. S., Tumanov, A., & Ramjee, R. (2024). Taming throughput-latency tradeoff in LLM inference with Sarathi-Serve. In 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24) (pp. 117–134). USENIX Association.
40. Prabhakar, R. B., Zhang, H., & Wentzlaff, D. (2024). Kraken: Inherently parallel transformers for efficient multi-device inference. In Advances in Neural Information Processing Systems, 37. Neural Information Processing Systems Foundation.