

Multi-Turn Intent Modeling for Financial Question Answering with Large Language Models

Quentin Barnett

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV,
USA.

quentin.work@unr.edu

Abstract

Multi-turn financial question answering requires systems to interpret evolving user intent across sequential interactions, often involving implicit references, corrections, and shifting risk constraints. Large language models offer powerful language understanding and generation capabilities, but their deployment in financial settings introduces structural challenges related to state persistence, memory management, retrieval grounding, calibration, governance, and operational cost. This paper presents a system-level examination of multi-turn intent modeling for financial question answering. It describes intent as a dynamic construct shaped by dialogue context, user goals, and regulatory obligations. The discussion evaluates architectural options including modular state trackers, memory stores, retrieval-augmented generation, and policy layers, and it analyzes trade-offs between monolithic and component-based designs. The paper further addresses robustness and fairness concerns, including hallucination risk, distributional bias, and auditability. Deployment considerations such as latency, energy consumption, model distillation, and infrastructure sustainability are explored. The analysis emphasizes that successful financial question answering systems must balance model capability with governance controls, memory privacy, and operational resilience. The paper contributes a conceptual framework for designing multi-turn financial assistants that are not only accurate but also accountable, fair, and sustainable.

Keywords

large language models; financial question answering; multi-turn dialogue; intent modeling; memory systems; governance; deployment.

1. Introduction

Financial question answering has undergone a substantial transformation from static retrieval over structured data to dynamic conversational interaction. Early systems relied heavily on keyword matching, predefined templates, and manually curated taxonomies, but these approaches struggled with the contextual dependencies that naturally arise when users ask follow-up questions, correct prior assumptions, or introduce new constraints. The introduction of pretrained bidirectional encoders enabled more robust language representation [1], and subsequent autoregressive large language models demonstrated remarkable few-shot capabilities across diverse tasks [2]. However, single-turn answer generation remains insufficient for financial users who frequently refine their queries over multiple turns, refer back to earlier statements, and shift between general information seeking and personal financial decision making.

The challenge of multi-turn intent modeling is not merely a language understanding problem. It is also a systems problem involving the stable representation of user goals, the selective retention of dialogue context, and the controlled integration of external evidence. Retrieval-

augmented generation has emerged as an important mechanism for grounding responses in relevant documents and reducing factual errors [3]. Instruction tuning has improved the ability of models to follow user intent across diverse prompts [4]. Chain-of-thought prompting has demonstrated that explicitly structured reasoning can help models handle multi-step financial questions [5]. These capabilities are valuable, but they introduce new architectural demands. A financial assistant must decide how to store, update, and expire contextual intent across turns, how to retrieve supporting information without introducing latency, and how to enforce regulatory and ethical boundaries.

This paper provides a system-level examination of multi-turn intent modeling for financial question answering with large language models. It argues that intent modeling should be treated as an infrastructure problem rather than an isolated model quality problem. The paper examines the characteristics of financial intent, architectural foundations for state management, memory persistence, robustness, fairness, governance, deployment sustainability, and evaluation trade-offs. The goal is to articulate a conceptual framework that supports accountable, resilient, and efficient financial question answering systems.

2. Characterizing Multi-Turn Intent in Financial Question Answering

Intent in financial question answering is more than slot filling or single-sentence classification. A user may begin by asking about the current shape of the yield curve, then ask what that shape implies for a bond portfolio, and then ask how a specific regulatory change might alter that implication. Each turn depends on the prior turn, and the underlying intent evolves as the user acquires information and narrows or broadens their goal. A broad survey of large language models notes that these systems exhibit strong language capabilities but also significant variability in task-specific reliability [6]. In the financial domain, this variability is especially consequential because intent misinterpretation can lead to unsuitable recommendations, missed compliance obligations, or direct financial loss.

The challenge is compounded by the many forms of implicit reference that occur in natural dialogue. Pronoun resolution, ellipsis, and temporal references require the system to maintain a coherent model of what the user has already said and what they are now asking. Users may refer to the previous rate, the earlier fund, or the same period without restating the full context. Financial intents can be informational, comparative, advisory, risk assessment oriented, or regulatory in nature. Each intent type carries different evidentiary standards and different degrees of acceptable inference. Foundation models exhibit emergent conversational behavior [7], but they also carry documented risks related to factual errors, bias, and overconfidence [8]. Analyses of large language model behavior have highlighted important limits in understanding, reasoning, and context retention [9]. These limits are particularly problematic when a user shifts from a general inquiry to a personalized financial decision without explicitly announcing that shift.

Compared with open-domain question answering, financial question answering has a lower tolerance for ambiguity. In customer support settings, a system can often recover from a misunderstanding with a clarifying question or a graceful apology. In financial settings, a misinterpreted intent can persist across multiple turns, accumulate into an incorrect portfolio comparison, or produce a response that appears authoritative while being materially incomplete. The system must therefore treat intent as a dynamic, partially observable state that must be actively managed rather than passively inferred once. This characterization has direct implications for architecture, because it requires mechanisms for state tracking, confidence estimation, disambiguation, and controlled forgetting.

3. Architectural Foundations for Intent State Management

A production financial question answering system should not rely solely on a monolithic language model to handle all aspects of multi-turn dialogue. Instead, system-level decomposition can improve auditability, maintainability, and policy enforcement. A modular architecture typically includes a dialogue encoder, an intent state tracker, a memory manager, a retrieval component, a generator, and a policy layer. Open foundation models enable finer control and internal fine-tuning for specialized financial tasks [10]. The modular design allows each component to be evaluated and updated independently, which is especially important in regulated environments where changes to one subsystem may need to be documented and tested without retraining the entire model.

Prompting methods offer one way to manage dialogue context by packaging conversation history and task instructions into the model input [11]. However, prompting alone has structural limits. Long conversations create large input sequences that increase computational cost, exceed context windows, and may degrade attention quality across distant turns. Prompting also blurs the boundary between state representation and instruction, making it difficult to audit what the system believes about the user at any point in time. Persistent memory systems have therefore been proposed to complement prompting by storing relevant user information outside the immediate context. Recent work on user memory systems for query matching and intent modeling [12] provides evidence that persistent user representations can improve retrieval efficiency and downstream intent interpretation. The attention mechanism remains the foundational computational building block for many of these systems [13], but effective intent state management requires additional structure around attention rather than reliance on attention alone.

The architectural trade-off between monolithic and component-based designs is central to financial deployment. A monolithic model may be simpler to serve and may reduce inter-component integration failures, but it offers limited transparency for regulators, auditors, and risk officers. A modular system can expose the intent state, memory decisions, and retrieval evidence as separate artifacts, which supports explainability and targeted intervention. The cost of modularity is increased latency and the need for careful interface design. In financial question answering, this cost is often justified because the ability to inspect and control intermediate states reduces operational risk.

4. Memory Systems and Contextual Persistence

Memory is a defining element of multi-turn intent modeling. Short-term memory can be represented by a dialogue buffer that holds recent turns, but this buffer alone cannot support cross-session personalization or long-term user understanding. Episodic memory stores records of past interactions, while semantic memory stores generalized user preferences, constraints, and profile attributes. In financial question answering, semantic memory may include risk tolerance, portfolio composition, regulatory jurisdiction, and communication preferences. A memory subsystem must decide what information to store, how long to store it, under what consent it may be used, and how it should influence future intent interpretations.

Memory consolidation and forgetting are critical design issues. Not all past turns should persist indefinitely. Users may change their objectives, revise earlier statements, or explicitly ask the system to disregard previous guidance. Stale intents can corrupt future answers and create legal exposure if outdated preferences are treated as current facts. Systems should implement expiration policies, confidence decay, and user correction mechanisms. The

memory approach described in [12] emphasizes the value of structured user memory for improving query matching and intent modeling over time, and similar principles apply in financial assistants that need to maintain a trustworthy representation of the user without assuming that all historical signals remain valid.

Privacy and regulatory compliance heavily constrain memory design in financial applications. Data retention laws, data minimization principles, and consumer protection rules require memory stores to be encrypted, auditable, and region-segregated. Users may have the right to inspect, correct, or delete inferred attributes. A system that infers a risk profile from dialogue must distinguish between user-provided information and model-inferred information, because inferred attributes may be subject to different legal treatment. Cross-session intent continuity can improve user experience by reducing repetition, but it also creates regulatory risk if preferences are inferred incorrectly or if sensitive financial attributes are retained without proper consent. Robust memory architectures therefore include consent tracking, provenance labels, and interfaces for correction.

5. Robustness, Calibration, and Risk Containment

Financial question answering systems must avoid hallucination and overconfidence. Calibration refers to the alignment between a model’s expressed confidence and its actual correctness. Well-calibrated systems are essential in finance because users may act on generated advice. Model documentation frameworks provide structured reporting about intended use, limitations, and evaluation procedures [14]. Such documentation can help risk officers understand when a system should be trusted and when it should escalate to a human expert. Risk containment also requires source grounding, answer refusal, and explicit uncertainty signaling when the model cannot verify critical information.

Known harms from language models include toxic generation, sensitive information disclosure, and over-reliance by users [15]. In financial question answering, over-reliance on a confident but incorrect projection is a serious failure mode. A system may generate a plausible explanation of a tax treatment or investment strategy while omitting necessary disclaimers or jurisdiction-specific exceptions. Robustness therefore requires more than accuracy on benchmark data. It requires adversarial testing, prompt injection defenses, and monitoring for distributional shifts in user queries. A policy layer can reject out-of-domain requests, enforce answer formats, and log decisions for later review.

Calibration also intersects with retrieval. A grounded answer that cites a specific regulatory document is generally easier to audit than a free-form generation that lacks evidence. Retrieval-augmented systems can reduce hallucination by constraining generation to retrieved passages, but retrieval quality then becomes a critical dependency. In multi-turn settings, the retrieval query must be constructed from both the current utterance and the accumulated intent state. If the intent state is wrong, the retrieval may return irrelevant documents, and the final answer may be confidently irrelevant. This coupling between memory, retrieval, and generation makes robustness a system property rather than a property of the language model alone.

6. Fairness, Governance, and Regulatory Alignment

Fairness in financial question answering requires detecting disparate treatment across demographic groups and user segments. Training data may underrepresent certain financial situations, and models may produce different quality answers depending on vocabulary, dialect, or assumed knowledge. Governance frameworks should include rigorous fairness

testing, impact assessments, and ongoing monitoring. The same model may perform well for mainstream queries but fail for users with nontraditional financial arrangements or limited financial literacy. Intent modeling can amplify these differences if it infers user sophistication or risk tolerance from biased signals.

Regulatory alignment is a central concern. A system that provides personalized financial advice may trigger licensing, fiduciary, or disclosure obligations. The system should differentiate general financial education from personalized recommendation, and it should preserve an audit trail that records the intent state at each turn. Audit trails are important not only for compliance but also for accountability when a user contests a system-generated answer. Model documentation tools such as model cards [14] and harm taxonomies [15] provide operational support for governance, but they must be integrated into deployment workflows rather than treated as static documents.

Human oversight is another essential governance mechanism. Financial assistants should support escalation to human advisors when the intent is ambiguous, when the query involves high-stakes decisions, or when the system detects conflicting user constraints. The governance system should define thresholds for escalation and should track the rate at which users override or correct system interpretations. These signals can be used to improve intent models and to identify areas where the system should be more conservative. Governance, in this sense, becomes a feedback loop that continuously aligns system behavior with regulatory and ethical expectations.

7. Deployment Infrastructure and Sustainability

Deployment infrastructure for financial question answering must balance latency, cost, reliability, and environmental sustainability. Large language models impose significant computational burdens at both training and inference time. Research on energy consumption in deep learning for natural language processing has highlighted the environmental cost of training and serving large models [16]. Carbon emissions associated with neural network training have become a major policy concern [17]. Scaling laws show that model quality tends to improve with parameter count and training compute [18], while compute-optimal training suggests that model size and data volume should be jointly optimized [19]. These findings create pressure to use large models for capability while simultaneously seeking efficiency.

Model distillation offers one avenue for reducing deployment cost. Distilled models can retain much of the functionality of larger teachers while requiring less memory and computation [20]. In financial question answering, distillation can be combined with retrieval and memory components so that the generative model remains relatively compact. Retrieval caching, precomputed embeddings, and tiered serving architectures can further reduce latency. Financial workloads often exhibit burst patterns around market events, news releases, and regulatory deadlines, so deployment systems must support elastic scaling without sacrificing response quality.

Sustainability is not only about energy consumption. It also includes operational maintainability, data pipeline quality, and the ability to update system components without full retraining. Modular architectures support partial updates, which can reduce both financial and environmental costs over the system lifecycle. Monitoring intent drift and memory quality can trigger targeted recalibration rather than expensive retraining. A sustainable

financial question answering system is one that remains accurate, compliant, and efficient under changing market conditions and evolving user needs.

8. Evaluation and System-Level Trade-Offs

Evaluation of multi-turn financial question answering requires a broader set of metrics than single-turn answer accuracy. Intent modeling quality should be measured through dialogue-level task success, including whether the system correctly resolves ambiguous references, maintains consistency across turns, and asks clarifying questions only when necessary. Factual correctness, answer faithfulness, and source grounding are essential but insufficient on their own. User satisfaction and trust are also important, especially in financial contexts where repeated errors may drive users away even if later answers are correct.

System-level trade-offs are unavoidable. Larger models may improve language understanding but increase cost, latency, and carbon footprint. Retrieval augmentation may improve grounding but adds dependency on external data quality and freshness. Memory systems may enhance personalization but increase privacy risk and regulatory burden. Governance controls may reduce harm but can slow system development and limit user flexibility. A mature deployment treats these trade-offs as dynamic constraints that must be revisited as market conditions, regulatory expectations, and user behavior change.

The evaluation framework should therefore include technical metrics, risk metrics, and operational metrics. Technical metrics capture intent accuracy and answer quality. Risk metrics capture hallucination rates, bias indicators, and unsafe generation frequency. Operational metrics capture latency, availability, cost, and energy use. These metrics should be monitored jointly because optimizing one dimension often degrades another. The system-level perspective advocated in this paper is intended to help architects and decision makers recognize these interactions rather than treating model performance as the sole goal.

9. Conclusion

Multi-turn intent modeling for financial question answering with large language models is a system-level challenge that extends far beyond the language model itself. It requires careful management of dialogue state, memory persistence, retrieval grounding, robustness, fairness, governance, and deployment sustainability. Financial question answering operates under strict accountability constraints because users may act on generated answers and regulators may examine system behavior after adverse outcomes. Architectural choices must therefore support transparency, auditability, and controlled intervention.

The conceptual framework presented in this paper emphasizes that intent should be treated as dynamic, partially observable, and subject to governance. Modular architectures, persistent memory systems, and robust policy layers can help meet these demands, but they introduce trade-offs in latency, cost, and system complexity. Evaluation must account for these trade-offs rather than relying on narrow benchmark accuracy. By adopting a system-level view, organizations can design financial question answering systems that are not only capable of understanding multi-turn intent but also resilient, fair, and sustainable under real-world constraints.

References

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186.

2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
4. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
5. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
6. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.
7. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
8. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
9. Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681–694.
10. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
11. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1–35.
12. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
14. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.
15. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Myatt, J., Glaese, A., Raileanu, R., Irving, G., et al. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

16. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650.
17. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.
18. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
19. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. (2022). Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35, 30016–30030.
20. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.