Steering Large Language Models for Financial Decision-Making Through Contrastive Learning

Zongzhe Lin, Ziqi Wei, Anran Zhao,
University of California Los Angeles
zlin3@g.ucla.edu

Abstract

Large language models (LLMs) have shown remarkable capabilities in financial analysis, yet they often lack fine-grained control over decision-making behaviors, such as risk assessment and wealth-seeking tendencies. In this work, we introduce a steering mechanism using contrastive learning to modulate financial reasoning in LLaMA-2. By training with bidirectional preference optimization, our approach aligns model outputs with domain-specific financial decision-making objectives while maintaining linguistic fluency. To achieve this, we apply targeted interventions at Layer 15 of LLaMA-2, a mid-range layer responsible for semantic reasoning, allowing behavior modification without altering core parameters. Training on AWS A100 GPUs for 20 epochs, our method improves efficiency and enhances decision-making consistency. Experimental evaluations demonstrate that our steered LLMs better align with financial principles while maintaining adaptability across various financial contexts.

1 Introduction

With the rise of AI-powered financial analysis, large language models (LLMs) are increasingly used to process complex financial documents such as SEC filings. These models exhibit remarkable capabilities in extracting insights, but ensuring accurate and interpretable financial decision-making remains a challenge. While LLMs can analyze financial reports, their responses often lack consistency and fail to align with fundamental financial reasoning, potentially leading to misleading or suboptimal investment insights.

Building upon the work of Cao et al. (2024) on steering LLMs via bidirectional preference optimization, we propose a method to enhance financial decision-making through contrastive learning. Our approach generates steering vectors that refine model outputs, aligning them with domain-specific financial reasoning. To implement precise interventions, we modify behavior at Layer 15 of LLaMA-2, a mid-range layer responsible for semantic reasoning, ensuring that financial insights are adjusted without disrupting the core linguistic capabilities of the model.

Our key contributions include:

- Developing contrastive learning-based steering vectors to improve financial decision-making in LLMs.
- Optimizing the decision-making process at Layer 15 of LLaMA-2, achieving a balance between control and adaptability.
- Conducting empirical evaluations to validate the effectiveness of our steering vectors in aligning financial responses with established financial principles.

Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2024.

2 Methodology

2.1 Contrastive Learning for Steering Vectors

Contrastive learning is a self-supervised learning technique that trains a model by distinguishing between similar and dissimilar examples. In the context of steering large language models (LLMs) for financial decision-making, contrastive learning refines the model's response patterns by reinforcing desirable financial reasoning while discouraging misleading or biased financial conclusions.

To effectively steer financial decision-making models, we apply contrastive learning with bidirectional preference optimization. This technique trains the model using curated prompt-response pairs that contain both preferred and non-preferred responses. Preferred responses reflect sound financial principles, while non-preferred responses contain flawed reasoning, speculation, or financial misinformation. The contrastive learning process ensures that preferred responses are embedded closer together in the representation space, while non-preferred responses are pushed away.

The optimization process follows these key steps:

- **Curating Prompt-Response Pairs**: We create diverse financial decision-making scenarios, each paired with a response reflecting sound financial reasoning (preferred) and a contrasting response containing incorrect, misleading, or biased financial reasoning (non-preferred).
- **Embedding Space Separation**: The model learns to position preferred responses closer together while ensuring that non-preferred responses are mapped farther away in the representation space. This allows the model to internalize proper financial reasoning while filtering out misleading patterns.

2.2 Examples of Contrastive Learning in Financial Decision-Making

To illustrate how contrastive learning is applied in financial decision-making, we provide three examples that demonstrate how our method helps steer LLaMA-2 towards better financial reasoning.

Example 1: Risk Management in Portfolio Selection

- **Prompt**: "A retail investor with a low risk tolerance wants to invest $50,000. What would be a suitable strategy?"
- **Preferred Response**: "A conservative portfolio with diversified low-risk assets such as Treasury bonds, blue-chip stocks, and ETFs with minimal volatility is recommended."
- **Non-Preferred Response**: "Investing the full amount in high-volatility tech stocks or cryptocurrency could yield higher returns."

Here, contrastive learning ensures that the model recognizes safe and diversified investment strategies as preferred while discouraging overly aggressive suggestions.

Example 2: Debt Analysis in Financial Reporting

- **Prompt**: "XYZ Corporation reports a 120% debt-to-equity ratio. How should an investor interpret this?"
- **Preferred Response**: "A debt-to-equity ratio above 100% indicates a highly leveraged company, which may pose risks in economic downturns."
- **Non-Preferred Response**: "A high debt-to-equity ratio always suggests strong financial health, making it a good investment."

The model learns to differentiate between contextual financial risks and overly simplistic or misleading interpretations.

Example 3: Earnings Report Interpretation

- **Prompt**: "XYZ Company announced a 5% revenue growth in its Q2 earnings. What should investors consider before making a decision?"
- **Preferred Response**: "Investors should assess net income growth, margins, debt levels, and broader market trends before making decisions."

- **Non-Preferred Response**: "Since revenue is increasing, buying the stock is the best move."

This example trains the model to avoid overgeneralized investment advice and consider comprehensive financial analysis.

2.3 Bidirectional Preference Optimization

While contrastive learning helps separate desirable and undesirable responses, it does not explicitly optimize for both reinforcement of good financial reasoning and suppression of incorrect reasoning. To address this limitation, we employ **Bidirectional Preference Optimization (BPO)**, which enforces both positive reinforcement and negative suppression.

BPO extends contrastive learning by introducing a dual-objective optimization process:

- **Positive Reinforcement:** Preferred responses are encouraged by maximizing similarity within their cluster in the representation space.
- **Negative Suppression:** Non-preferred responses, which contain flawed reasoning or financial misinformation, are explicitly pushed away from the preferred response space.
- **Iterative Refinement:** This optimization is conducted iteratively across different financial scenarios to enhance robustness and generalization in decision-making.

2.3.1 Why Bidirectional Preference Optimization?

While standard contrastive learning only ensures that good responses are closer together, it does not explicitly penalize incorrect responses beyond simple separation. BPO enhances this by actively suppressing non-preferred responses, reducing the likelihood of the model generating misleading financial recommendations.

By applying bidirectional preference optimization, we create a more reliable and interpretable financial AI model, reducing risk and increasing trustworthiness in AI-driven financial insights.

2.4 Layer-Specific Intervention

The Transformer architecture, introduced by Vaswani et al. [2], replaces traditional sequential models such as recurrent neural networks (RNNs) with a fully attention-driven framework. The self-attention mechanism enables each token to attend to every other token in the sequence, allowing global context awareness. Multi-head attention further enhances representation learning by capturing diverse linguistic patterns, while positional encodings compensate for the lack of recurrence, preserving sequential information.

2.4.1 Why Target Layer 15?

A key characteristic of Transformer models is their hierarchical feature extraction across different layers. Lower layers primarily encode token-level representations, focusing on syntactic structures and basic dependencies. Mid-range layers refine these representations, learning contextual relationships and domain-specific semantics. Higher layers then synthesize this information to generate task-specific outputs. This hierarchical processing structure plays a critical role in financial decision-making, where different levels of abstraction contribute to model reasoning.

Selecting the appropriate intervention point is crucial for effective steering. Each Transformer layer encodes distinct levels of information:

Lower Layers (1-10): These layers primarily process token embeddings and fundamental syntactic structures, such as part-of-speech tagging and dependency parsing. Modifying these layers would disrupt basic word representations but would not meaningfully influence financial decision-making.

Mid Layers (10-20): These layers capture deeper *semantic understanding*, *contextual reasoning*, and *domain-specific knowledge*. Adjusting activations in this range allows the model to internalize financial decision-making behaviors while maintaining fluency and coherence.

Higher Layers (20-30): The final layers focus on *output generation*, where decisions are already shaped by lower-layer representations. Steering at this level risks making overly direct modifications, reducing the model's flexibility and adaptability to different financial contexts.

To ensure precise control over financial decision-making while preserving general language fluency, we intervene at Layer 15 of LLaMA-2. This layer strikes a balance between structural comprehension and output-level modulation, making it an ideal point for steering financial reasoning.

Empirical evaluations confirm that Layer 15 provides the best trade-off between adaptability and control. By injecting steering vectors into this layer's intermediate activations, we effectively fine-tune financial decision-making behavior while preserving the model's ability to generalize across diverse linguistic tasks. This intervention ensures that financial insights remain precise while maintaining coherence in broader textual contexts.

2.5 Optimizing Steering Vectors

To effectively steer financial reasoning at Layer 15, we employ a gradient-based optimization strategy that subtly adjusts the model's internal representations while preserving its general language capabilities. This optimization process ensures that financial decision-making aligns with domain-specific principles without modifying the base model's core parameters.

The optimization begins by injecting targeted modifications into the intermediate activations of Layer 15. By applying gradient-based steering, we refine the latent space representations, guiding the model towards financially sound reasoning. These modifications encourage the model to internalize financial best practices, such as risk assessment, portfolio diversification, and earnings analysis, ensuring that responses remain aligned with expert-driven financial insights.

Once the initial modifications are applied, the steering vectors undergo an iterative refinement process. Model responses to a diverse set of financial prompts are continuously evaluated, ensuring that the adjustments maintain coherence while reinforcing proper financial reasoning. This iterative tuning is essential for balancing precision and flexibility—ensuring that while the model adheres to domain-specific financial logic, it remains adaptable to different financial contexts.

By focusing on steering at the embedding level rather than altering the model's core weights, this approach provides a lightweight yet effective intervention. The model retains its general language fluency while improving its ability to generate financially responsible and contextually appropriate responses. This optimization strategy ultimately enhances interpretability, enabling LLaMA-2 to provide more reliable insights in financial decision-making scenarios.

2.6 Data Construction and Evaluation

The dataset used for steering financial decision-making is carefully constructed to reflect expert-driven financial reasoning. Instead of relying on external financial datasets, we generate structured prompt-response pairs that serve as training examples. Each prompt presents a financial scenario, accompanied by a preferred response that aligns with sound financial principles and a non-preferred response that contains flawed, misleading, or overly simplistic reasoning.

These curated prompt-response pairs enable the model to internalize proper decision-making strategies through contrastive learning and bidirectional preference optimization. The preferred responses reinforce financially responsible reasoning, while the non-preferred responses serve as counterexamples, helping the model distinguish between high-quality and erroneous financial advice.

To evaluate the effectiveness of our steering approach, we test the model on a separate set of financial prompts that were not included in training. By comparing responses before and after intervention, we assess the degree to which the model improves in financial reasoning accuracy.

3 Results and Discussion

The experimental results demonstrate the effectiveness of our steering vector approach in refining financial decision-making within LLaMA-2. By comparing model responses before and after steering vector intervention, we observe a notable shift in behavior toward more structured and desirable financial reasoning.

Without the steering vector, the model maintains a neutral stance, adhering to broad guidelines without making domain-specific recommendations. While this approach prevents controversial outputs, it also limits the model's ability to provide insightful and well-reasoned financial advice.

With the steering vector applied, responses align more closely with financial best practices, showing an improved ability to prioritize long-term stability, risk assessment, and ethical financial decision-making. The intervention at Layer 15 successfully modulates decision-making while preserving the general linguistic fluency of the model.

3.1 Successful Example: Risk Assessment in Investment

To illustrate the impact of our steering approach, we present the following example:

Prompt: *An investor is considering allocating 70% of their portfolio into a single volatile stock with significant recent price swings. What advice would you provide?*

Response without Steering Vector: *"Investment decisions depend on personal risk tolerance. Diversification can be useful, but high-risk strategies may also yield high rewards."*

Response with Steering Vector: *"Allocating 70% of a portfolio to a single volatile stock introduces substantial risk. A more balanced approach, such as diversifying into a mix of stable assets and growth opportunities, would mitigate downside exposure while maintaining potential for returns."*

This example highlights a key benefit of our steering methodology. Without the intervention, the model provides a vague and non-committal response, failing to offer actionable financial guidance. With the steering vector applied, the model delivers a well-reasoned recommendation emphasizing diversification and risk mitigation, aligning with established financial principles.

The qualitative examples presented indicate a consistent pattern: the steered model moves away from generic, non-committal responses and instead provides structured guidance tailored to financial reasoning. This suggests that steering vectors can effectively bridge the gap between generalized AI outputs and domain-specific expertise.

Overall, these results validate the effectiveness of bidirectional preference optimization and layer-specific intervention in shaping LLaMA-2's financial reasoning. Future research could explore the adaptability of steering vectors across other domains, enhancing AI interpretability and decision-making fidelity in specialized fields.

4 Future Improvements

While our steering vector approach has demonstrated effectiveness in shaping financial decision-making, there remain opportunities for further refinement and enhancement.

One key limitation is the reliance on a single-layer intervention, which may not fully capture complex decision-making processes across different contexts. Future work could explore multi-layer steering, leveraging coordinated modifications across multiple layers to enhance model control while preserving linguistic fluency.

Another area of improvement is interpretability. Understanding how specific neurons respond to steering interventions could provide insights into the internal decision-making mechanisms of LLaMA-2. By analyzing neuron activations at different layers, we could refine steering vectors to be more precise and efficient.

Additionally, our current method focuses on guiding financial decision-making along a single axis of preference. However, in real-world applications, financial reasoning often requires balancing multiple factors. Exploring vector fusion methods could allow the model to integrate multiple behavioral objectives without introducing conflicts, enhancing its adaptability across diverse financial scenarios.

By addressing these challenges, we aim to further improve the precision, interpretability, and robustness of steering mechanisms, making AI-driven financial decision-making more reliable and transparent.

References

[1] Cao, Yuanpu et al. *Personalized Steering of Large Language Models: Versatile Steering Vectors Through Bi-directional Preference Optimization*. arXiv preprint arXiv:2406.00045, 2024.

[2] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. *Attention is All You Need*. Advances in Neural Information Processing Systems, 2017.