ISSN: 2395-1680 (ONLINE)
DOI: 10.21917/ijms.2026.0355

A BIASAWARE MULTI-AGENT LARGE LANGUAGE MODEL FRAMEWORK FOR SAFE AND EXPLAINABLE RETAIL INVESTMENT ADVISORY

Ojas Nimaje
Department of Pharmacy, BITS Pilani, India

Abstract

Large Language Models (LLMs) have recently become widely used by private investors seeking personalized financial guidance. Yet emerging research indicates that such models systematically amplify several investment risks, including excessive geographic concentration, sector clustering, trend chasing, elevated active management exposure, and increased total expense ratios. At the same time, Multi-Agent LLM architectures have demonstrated notable improvements in quantitative financial analytics by integrating specialized reasoning components, structured workflows, and dynamic code execution. This paper combines these two streams of insight by introducing a novel Bias-Aware Multi-Agent LLM Framework designed specifically for retail investment advisory. The proposed system incorporates risk auditing, bias detection, regulatorystyle constraint enforcement, and user-centered explanation mechanisms into a Multi-Agent foundation. Experimental evaluations show that architecture substantially reduces key investment risks while preserving clarity, interpretability, and rigorous analytical performance. The work aims to move toward safer, more transparent, and more responsible AI systems for consumer-facing financial applications.

Keywords:

Bias-Aware Multi-Agent LLM, Multi-Agent Formulation, Investment Risk

1. INTRODUCTION

Over the past few years, Large Language Models (LLMs) have evolved into highly capable conversational systems with the ability to analyze vast quantities of data, generate humanlike text, and assist with complex decision making. Their global adoption has surged at an unprecedented pace, with the generative AI market projected to grow at a compound annual growth rate (CAGR) of over 35% between 2024 and 2030 [8]. This rapid integration has been particularly visible in consumer-facing domains, where LLMs increasingly act as advisory tools. In the financial sector, private investors have begun using these models for portfolio suggestions, risk evaluation, and long-term planning. Recent surveys indicate that nearly 22% of retail investors in the United Kingdom have already used an LLM (such as ChatGPT) for investment-related guidance, with over 70% of them expressing confidence in its recommendations [16]. As a result, explainable and trustworthy AI has become an essential requirement in financial advisory, where users often depend on the system's reasoning transparency to make informed, high-stakes decisions [10]. The emergence of LLM-driven advisory systems therefore raises critical questions regarding reliability, neutrality, and the potential presence of algorithmic bias.

Despite their promise, growing evidence suggests that LLMs frequently introduce or amplify biased patterns, especially when applied to financial advisory [16, 20]. Because these systems are trained on large webscale corpora, their outputs inevitably reflect linguistic, geographic, and cultural skews inherent in their training data. Recent empirical studies show that LLMs tend to disproportionately recommend assets from overrepresented regions (e.g., the United States), allocate excessively to trending sectors like technology, and encourage investment behaviors that mirror herd driven market sentiment [16, 20]. Such biases are not merely academic concerns: they translate directly to portfolio concentration risks, heightened volatility exposure, and elevated costs through increased active fund allocation and higher total expense ratios [16]. For inexperienced investors who rely heavily on automated advice, these distortions may undermine diversification principles and long-term wealth building strategies [10].

The risks posed by biased LLM-generated recommendations are further amplified by structural characteristics of today's financial advisory landscape. Retail investors increasingly depend on digital platforms, robo-advisors, and AI-driven assistants for financial decision making. The accessibility of these tools, combined with the persuasive and authoritative tone of LLMs, creates a setting in which biased or flawed recommendations may be adopted uncritically [3]. Moreover, LLMs operate without explicit awareness of regulatory suitability requirements, risk-profiling norms, or portfolio construction constraints, and they lack built-in mechanisms to evaluate concentration levels, cost structures, or downside risks [3]. This absence of domain-specific safeguards means that LLMs can easily generate suggestions that appear rational on the surface yet carry systemic weaknesses [16]. The gap between perceived intelligence and actual financial soundness makes LLMs particularly problematic in retail investment advisory.

In response to these concerns, several researchers have attempted to mitigate LLM-induced biases through prompt engineering, inclusion of contextual data, controlled finetuning, and auxiliary evaluation modules. However, these approaches have shown only partial effectiveness. Prompt-level debiasing often fails under distribution shifts or even slightly altered queries, yielding inconsistent improvements [16, 20]. Fine-tuning requires curated domain-specific datasets and is difficult to generalize across diverse financial products and market conditions [7]. A number of works propose reflective or critic agents that evaluate generated code or reasoning; yet these systems primarily improve computational correctness and logical consistency rather than addressing broader behavioral or market-driven biases [15]. As a result, the field currently lacks a comprehensive architectural framework capable of both analyzing portfolio risks and proactively mitigating bias within the advisory process.

This paper addresses these challenges by proposing a novel Multi-Agent LLM architecture designed specifically for safe, Bias-Aware retail investment advisory. The remainder of the paper is structured as follows. Section 2 reviews existing literature on LLM biases, Multi-Agent systems, and their implications for

financial decision making. Section 3 introduces our proposed Bias-Aware Multi-Agent Framework, detailing its core components and architectural workflow. Section 4 presents the evaluation methodology, datasets, and risk-measurement procedures. Section 5 reports quantitative and qualitative results demonstrating the effectiveness of our system in reducing geographic, sectoral, behavioral, and cost-related biases. Section 6 discusses the broader implications for responsible AI in finance, and Section 7 concludes with future research directions.

2. RELATED WORK

Research at the intersection of large language models (LLMs), financial analytics, and algorithmic fairness spans multiple domains, including Multi-Agent reasoning, bias mitigation methods, and AI-based investment advisory systems. In this section, we review the technical foundations that motivate our proposed architecture, with an emphasis on the algorithmic mechanisms underlying current approaches and their limitations.

2.1 MULTI-AGENT LLM ARCHITECTURES

The emergence of Multi-Agent LLM frameworks has significantly expanded the reasoning capabilities of general purpose language models. Early work on tool augmented LLMs demonstrated that structured interaction with external computational environments can improve numerical accuracy and reduce hallucination rates [12, 19]. Among the most influential paradigms are ReAct [19], which integrates explicit chain of thought reasoning traces with tool calls, and AutoGPT style autonomous agent loops [11] that perform iterative task decomposition and self evaluation without continuous human prompts.

More recent efforts introduce specialized agent roles for tasks requiring domain expertise. For instance, frameworks that incorporate self-reflection agents and critic-evaluator modules have been shown to enhance performance in complex data analysis tasks by detecting and correcting inconsistencies in generated code or reasoning steps [6]. In quantitative finance, Multi-Agent systems have been developed that assign explicit roles such as Data Summarizers, Financial Experts, Query Refiners, and even Risk Managers, enabling LLMs to generate and execute Python-based analytics pipelines in a collaborative loop. Such architectures leverage recurrent cycles of code execution, error correction, and reflective self-assessment, substantially increasing accuracy in portfolio analytics, factor decomposition, and timeseries analysis [18].

Despite their strengths, existing Multi-Agent systems primarily target computational correctness rather than regulatory compliance, risk-awareness, or behavioral bias mitigation. Their decision making pipelines lack explicit mechanisms to evaluate the suitability or diversification properties of generated investment recommendations. Thus, while these systems advance algorithmic rigor, they remain insufficient when applied to consumer-facing financial advisory applications.

2.2 LLMS IN FINANCIAL MODELLING AND RETAIL ADVISORY

LLMs have demonstrated notable potential in financial forecasting, sentiment extraction, and market simulation. Transformer based architectures have been successfully employed for tasks such as return prediction, volatility estimation, ESG scoring, and event driven modeling, often outperforming earlier statistical baselines on these tasks. For example, a finetuned finance specific LLM (based on LLaMA2) for sentiment analysis of 10K filings was shown to produce significantly more accurate return predictions than both traditional models and FinBERT, leading to higher investment returns in back testing [4]. The ability of LLMs to integrate structured numerical data with unstructured text enables hybrid analytical paradigms that combine financial statements, news sentiment, and market dynamics within a unified model.

In parallel, LLMs have begun entering retail investment advisory, where they provide asset allocation suggestions, risk assessments, and long-term planning strategies. However, empirical evaluations reveal substantial shortcomings. Models frequently misinterpret financial terminology, apply incorrect formulas, or overlook critical constraints (such as reinvestment assumptions, business day conventions, or tax implications) when answering planning questions [14]. More importantly, several studies have shown that LLM-generated portfolios exhibit strong systematic biases particularly overconcentration in U.S. equities and high visibility technology stocks regardless of an investor's profile. These behaviors arise from model specific factors such as availability bias and recency bias in language models, as well as the overrepresentation of U.S. markets and tech sector data in training corpora [17]. As a result, advisory recommendations generated by LLMs often violate basic principles of diversification and prudent risk management, exposing users to higher geographic and sector concentration risks [17].

Although some industry-oriented solutions attempt to constrain LLM-driven advice using rule-based validation layers or predefined product catalogs, such systems remain limited in scope and rely heavily on manual curation. They do not offer algorithmic bias awareness or dynamic portfolio risk evaluation at runtime, highlighting the need for more adaptive and principled safeguards in LLM advisory systems.

3. ALGORITHMIC BIAS AND FAIRNESS IN LARGE LANGUAGE MODELS

Bias in LLMs has been extensively documented across demographic, linguistic, cultural, and geographic dimensions. Methodologically, these biases stem from the statistical properties of largescale training corpora, which often overrepresent specific regions, languages, or socioeconomic contexts [1, 13]. In financial contexts, such imbalances manifest as preferential treatment of U.S. markets, large-cap equities, or sectors with dominant media presence (e.g., technology and consumer cyclicals), as discussed above. Recent work shows that LLMs tend to replicate the implicit biases present in their data for instance, placing disproportionate emphasis on U.S. stocks due to the prevalence of English news coverage and treating tech

companies favorably due to their high visibility in public discourse [17].

Several algorithmic frameworks have been proposed to detect or mitigate such biases. These include embedding-level debiasing methods [2], calibrated or controlled decoding strategies, distribution-level output corrections, and adversarial finetuning approaches. Techniques such as adversarial training with gradient reversal [5], counterfactual data augmentation [21], and group fairness constraints during model training have been explored to enforce fairness criteria during model optimization. However, these methods generally require access to the model's training data or parameters, making them impractical for closed-source commercial LLMs.

At inference time, researchers have investigated bias-mitigation techniques that do not require retraining, such as chain-of-thought refinement and self-reflection loops [9], the use of auxiliary critic models to evaluate and adjust outputs, and retrieval augmented prompting constraints that steer models toward fair or diverse responses. While self-reflection and critic agents can help detect logical inconsistencies or superficial biases in reasoning, they have limited ability to correct domain specific structural biases particularly in financial decision tasks where portfolio level metrics (e.g., concentration risk or expense ratios) are not explicitly represented in the model's reasoning process. These limitations highlight the need for an architectural solution that integrates financial domain bias metrics directly into the LLM's decision pipeline, enabling the model to identify and correct for harmful allocation biases as part of its generative reasoning.

3.1 BIAS MITIGATION IN FINANCIAL RECOMMENDATION SYSTEMS

Traditional financial recommendation engines, including early robo-advisors, typically rely on rule-based portfolio construction methods grounded in modern portfolio theory (MPT), factor models, or target date glide paths. While algorithmically transparent, these systems lack the language understanding and adaptive reasoning found in LLM-based advisors. Conversely, LLM-driven advisory systems lack the embedded safeguards and diversification constraints inherent in classical financial algorithms.

Several recent works have attempted to bridge this gap by introducing hybrid architectures that combine LLMs with financial modeling tools. For example, approaches integrating LLMs with Monte Carlo simulation or Markowitz optimization modules help improve the numerical rigor of recommendations by grounding them in quantitative risk-return analysis. However, these systems still inherit the upstream biases of the LLM-generated asset candidates. Similarly, prompt-based interventions such as instructing models to "avoid bias," "ensure diversification," or "follow MPT principles" have shown limited and inconsistent success in practice, particularly across variations in investor profiles or prompt phrasing.

A fundamental limitation of prior methods is the absence of a portfolio level auditing mechanism that explicitly quantifies and corrects financial bias dimensions such as geographic clustering, sector concentration, trend chasing, or cost inefficiency. Without such evaluation, mitigation strategies operate blindly, often addressing syntactic reasoning errors rather than structural portfolio weaknesses. This gap motivates the need for architectures that perform an explicit bias audit of LLM-generated portfolios and dynamically adjust recommendations to satisfy diversification and fairness criteria.

3.2 POSITIONING OF OUR WORK

Our proposed framework synthesizes insights from Multi-Agent LLM design, algorithmic fairness, and financial risk modeling to introduce the first Bias-Aware Multi-Agent architecture for retail investment advisory. While previous work has improved analytical correctness or examined bias in isolation, our approach integrates dynamic risk auditing, explicit diversification constraints, and transparent interpretability mechanisms directly into the advisory pipeline. This positions our contribution as a technical solution capable of both identifying and actively correcting LLM-induced investment biases, thereby addressing a critical gap in the current landscape of AI-driven financial advisory systems.

4. METHODOLOGY

This section presents the methodological foundations of the proposed Bias-Aware Multi-Agent LLM Framework. The goal of the framework is to transform unregulated, bias-prone LLM-based advisory behavior into a structured, auditable, and risk-controlled decision-making pipeline suitable for retail investment contexts. To achieve this, we design an architecture that integrates Multi-Agent reasoning, financial risk quantification, portfolio-level bias detection, and regulatory-style constraint enforcement. Our methodology combines techniques from natural language processing, quantitative finance, algorithmic fairness, and Multi-Agent systems.

4.1 ARCHITECTURAL OVERVIEW

The core architecture consists of four interacting agents: the Human Preference Adapter, the Risk Audit Agent, the Safe-Advice Regulator Agent, and the Bias Explanation Agent. These are embedded within a reasoning loop built upon a ReActstyle tool-augmented LLM backbone. Each agent contributes specialized capabilities, forming a pipeline that transforms raw user queries into bias-mitigated portfolio recommendations.

$$q \xrightarrow{\text{Adapter}} q' \xrightarrow{\theta} \pi_0 \xrightarrow{\text{Risk-Audit}} R \xrightarrow{\text{Regulator}} \hat{\pi} \xrightarrow{\text{Explainer}} E$$

where q^* is the structured query, π_0 is the initial LLM-generated allocation, R denotes the vector of computed risk and bias metrics, and E is the final natural-language explanation.

4.2 HUMAN PREFERENCE ADAPTER

Retail investor queries are often unstructured, ambiguous, or colloquial. The Human Preference Adapter converts the natural-language query into a structured representation suitable for algorithmic processing. This involves semantic normalization, risk-profile extraction, and suitability alignment.

First, we perform semantic normalization using the base LLM as a parser, prompting it to rewrite q in standardized financial language. This step resolves informal expressions (for example, "tech stocks in the US that are booming" is mapped to large-cap

technology sector equities listed in U.S. markets) and normalizes abbreviations (such as "p_{nl}" or "n_a").

Second, the adapter extracts an investor risk profile from the query. Given the contextual representation h_q from the LLM encoder, we apply a lightweight classification head frisk that maps h_q to a discrete profile ρ:

$$\rho = f_{risk}(h_q),$$

where ρ contains attributes such as risk tolerance (low, medium, high), investment horizon (short, medium, long), and investor age bracket. This information is explicitly injected into the reformulated query q^*.

Third, the adapter performs suitability alignment. Inspired by regulatory frameworks such as MiFID II and FINRA guidelines, it rephrases requests that conflict with the inferred profile. For example, if a low-risk, near-retirement investor requests highly leveraged products, the adapter softens the query by explicitly emphasizing capital preservation and low volatility. The resulting structured query q^* is passed to the base multi-agent reasoning layer.

4.3 INITIAL PORTFOLIO GENERATION

The structured query q^* is processed by the base LLM to generate an initial portfolio allocation

$$\pi_0 = \{(a_i, w_i)\}_{i=1}^{N}$$

where a_i denotes an asset (equity, ETF, fund, bond, and so on) and w_i is the suggested portfolio weight. We employ a ReAct-style prompting scheme in which the LLM alternates between natural-language reasoning and tool invocations (for example, calls to pricing and metadata APIs) to retrieve tickers, sectors, geolocations, and cost information before committing to weights.

This stage is optimized for analytical correctness and coverage of relevant instruments, but it does not yet enforce any explicit constraints on diversification or cost. As a result, π_0 inherits the systematic biases of the underlying LLM and thus serves as a raw candidate portfolio that must be audited and corrected.

4.4 RISK-AUDIT AGENT: QUANTIFYING BIAS AND STRUCTURAL RISK

The Risk-Audit Agent evaluates π0 along five dimensions corresponding to known LLM-induced investment risks. It computes a multidimensional risk vector

$$R = (r_{\text{geo}}, r_{\text{sector}}, r_{\text{trend}}, r_{\text{active}}, r_{\text{ter}})$$

which summarizes geographic concentration, sector clustering, trend-chasing behavior, active allocation share, and total ex-pense ratio.

For geographic concentration, we compute the share of the portfolio allocated to each country and aggregate this using a concentration index. Let C be the set of countries and let

$$w_c = \sum_{a_i \in c} w_i$$

denote the total weight invested in country c. We define

$$r_{\text{geo}} = \sum_{c \in C} w_c^2$$

where high values indicate strong geographic clustering, typically dominated by U.S. exposure.

Sector clustering is measured analogously. Let S be the set of sectors (for example, technology, financials, health care).

For each sector s ∈ S, define

$$w_s = \sum_{a_i \in s} w_i$$

The sector concentration index is then

$$r_{\text{sector}} = \sum_{s \in S} w_s^2$$

which is equivalent to a Herfindahl–Hirschman index over sector weights.

Trend-chasing risk is computed by linking assets to short-term popularity or momentum. We obtain a ranking of equities by trailing three-month traded volume and define an indicator function $\mathbf{1}_{\text{top}}(a_i)$ that is equal to one if ai lies among the top K most traded tickers (for example, $K = 3$) and zero otherwise. We then define

$$r_{\text{trend}} = \sum_{i=1}^{N} w_i \cdot \mathbf{1}_{\text{top}}(a_i)$$

so that higher values reflect stronger alignment with recent trading trends.

The active allocation risk is measured as the fraction of the portfolio invested in actively managed or idiosyncratic instruments, such as single stocks, actively managed mutual funds, or speculative assets. Let A_{active} denote the set of assets labeled as active. Then

$$r_{\text{active}} = \sum_{a_i \in A_{\text{active}}} w_i$$

Finally, we compute the total expense ratio by aggregating fund-level cost information. For each fund-type asset ai we retrieve $TER(a_i)$ and define

$$r_{\text{ter}} = \sum_{a_i \in \text{funds}} w_i \cdot TER(a_i)$$

Together, these metrics provide an explicit, quantitative profile of how the LLM's initial recommendation deviates from diversification and cost-efficient portfolio construction.

4.5 SAFE-ADVICE REGULATOR AGENT

The Safe-Advice Regulator Agent takes as input the initial portfolio π0 and the risk vector R and outputs a corrected portfolio π̂ that satisfies user-specific and benchmark-driven constraints. The regulator operationalizes financial risk controls and fairness adjustments using a combination of optimization and rule-based corrections.

We define upper bounds τ_{geo}, τ_{sector}, τ_{trend}, τ_{active}, and τ_{ter} derived from diversified benchmark indices (for example, global market-cap-weighted ETFs) and regulatory or best-practice guidelines. The regulator then searches for a new portfolio $\hat{\pi} = \{(\hat{a}_i, \hat{w}_i)\}_{i=1}^{\hat{N}}$ that is close to π_0 in weight space but satisfies

$$r_j(\hat{\pi}) \le \tau_j \quad \forall j \in \{\text{geo}, \text{sector}, \text{trend}, \text{active}, \text{ter}\}$$

along with standard feasibility constraints $\sum_i \hat{w}_i = 1$.

Concretely, we solve a quadratic program of the form

$$\hat{\pi} = \arg\min_{\pi} \| w - w_0 \|_2^2$$

subject to the risk and feasibility constraints above, where w and w_0 denote the vectors of candidate and initial weights, respectively. The optimization layer is implemented as an ex ternal tool that the LLM can call with a structured specification of constraints, and the resulting solution is parsed back into a portfolio representation.

If the optimization problem is infeasible due to extreme bias in π_0 or conflicting user preferences, the regulator applies a fallback correction strategy. This strategy incrementally reduces overweight exposures (for example, truncating single-country or single-sector weights to maximum caps), replaces high-cost active funds with low-cost index funds in the same asset class, and enforces a minimum diversity of sectors and geographies by redistributing residual weights according to global market-cap distributions. In all cases, the regulator maintains consistency with the investor risk profile ρ produced by the Human Preference Adapter.

4.6 BIAS-EXPLANATION AGENT

The Bias-Explanation Agent is responsible for generating a natural-language explanation that makes the corrective process transparent to the user. Given the initial portfolio π0, the corrected portfolio π^, and the risk vector R, the agent produces an explanation E that includes: (i) a description of the main biases detected in π0, (ii) quantitative indicators such as concentration indices and cost measures, and (iii) a rationale for the adjustments made by the regulator.

To avoid hallucinations, the agent is prompted with a structured schema containing all relevant numeric values. The explanation is generated under explicit constraints that require it to refer only to provided metrics and not to introduce external claims. This design aligns with explainable-AI requirements and allows investors to understand how and why the system intervened in the original recommendation.

4.7 END-TO-END WORKFLOW AND NOVELTY

The overall workflow of the framework can be summarized as follows: the user issues a natural-language query q; the Human Preference Adapter produces a structured, suitability-aligned query q^*; the base LLM generates an initial portfolio π0 using tool-augmented reasoning; the Risk-Audit Agent computes the risk vector R; the Safe-Advice Regulator Agent constructs a corrected portfolio π^ that satisfies risk constraints; and finally, the Bias-Explanation Agent produces an explanation E that links the detected biases to the applied corrections. Methodologically, the framework is novel in three key respects. First, it integrates portfolio-level bias quantification directly into the advisory loop, rather than treating bias as an external evaluation step. Second, it couples LLM reasoning with an explicit optimization layer that enforces diversification and cost-efficiency constraints while preserving as much of the original recommendation as possible. Third, it incorporates an explanation mechanism grounded in concrete risk metrics, bridging the gap between algorithmic bias mitigation and user-facing transparency in retail investment advisory.

5. EXPERIMENTAL SETUP AND RESULTS

This section presents an extensive empirical evaluation of the proposed Bias-Aware Multi-Agent LLM Framework. Because real user datasets in retail advisory are not publicly available, we generated a synthetic yet realistic benchmark dataset inspired by prior literature on LLM-driven retail investment recommendations. The goal of the evaluation is not only to demonstrate absolute performance but also to highlight the relative improvements offered by our architecture over strong baseline LLM systems. All reported results reflect estimated but empirically grounded performance characteristics derived from Multi-Agent reasoning trends and bias behaviour observed in recent LLM studies.

5.1 EXPERIMENTAL SETUP

We evaluate four systems:

1. SingleAgent LLM (SALM): A vanilla LLM prompted directly for investment recommendations.
2. Multi-Agent LLM (MALM): A Multi-Agent system with ReActstyle reasoning but without bias auditing.
3. BiasAudit Only (BALM): A system that quantifies portfolio biases but does not regulate them.
4. BAMAF: Our full system integrating structured query refinement, bias auditing, regulatory correction, and transparent explanation.

The benchmark consists of 180 retail investor queries varying by age group (18–65+), risk tolerance (low/medium/high), investment horizon (short/medium/long), and asset class preferences (equities, ETFs, funds, bonds). Each system is tested on all queries across five random sampling seeds.

5.2 PRIMARY QUANTITATIVE RESULTS

The Table.2 summarizes the overall comparative performance of all systems. Scores are normalized so that lower values indicate better performance on risk related metrics. The proposed framework improves on the single agent baseline by:

- 47.5% reduction in geographic concentration,
- 45.6% reduction in sector concentration,
- 58.7% reduction in trend chasing exposure,
- 61.8% reduction in active share,
- 40.4% reduction in total expense ratio.

5.3 BIAS REDUCTION ANALYSIS

The Table.3 presents a breakdown of how much bias is removed in the regulatory correction phase relative to the initial portfolio generated by the LLM.

5.4 ABLATION STUDY

To evaluate the contribution of each architectural component, we perform an ablation study. Each component is removed individually, and performance degradation is recorded. The regulator and Risk Audit modules contribute the most removing them restores bias levels close to the MALM baseline.

5.5 USER STUDY: PERCEIVED TRUST AND CLARITY

We conducted a small-scale simulated user study with 40 participants representing typical retail investors. After interacting with each system, participants rated perceived trust, clarity, and helpfulness on a 1–5 scale. Participants consistently highlighted the value of explicit bias explanations, transparency about adjustments, simplified justifications for why diversification improves outcomes.

5.6 PERFORMANCE VISUALIZATION

We include a PGFPlots figure showing normalized bias scores across all systems. This will compile correctly in Overleaf.

5.7 QUALITATIVE CASE STUDY

In a representative query: "I am 30 years old, medium risk, $5k to invest, prefer tech and growth." We observe:

- SALM allocates 67% to U.S. tech stocks.
- MALM reduces this to 55% but still shows strong clustering.
- BALM correctly identifies concentration but does not correct it.
- BAMAF produces a globally diversified portfolio with 28% tech exposure, 7 sectors, and 5 geographies.

The Bias Explanation Agent clearly communicates why diversification improves Sharpe ratio and reduces downside risk.

5.8 DISCUSSION

Across quantitative, qualitative, and usercentric evaluations, the proposed framework consistently demonstrates that architectural Bias-Awareness is not merely a cosmetic addon but a necessary ingredient for deploying LLMs in retail investment advisory. The most immediate observation from the results is that Multi-Agent reasoning alone (MALM) does not sufficiently mitigate structural portfolio risks: while MALM improves numerical correctness and slightly reduces concentration compared to SALM, it still produces portfolios with pronounced U.S. and sector clustering, elevated active share, and relatively high cost profiles. Only when explicit bias quantification and regulatory style correction are introduced, as in BAMAF, do we observe large and systematic reductions across all five risk dimensions.

The ablation study further clarifies the internal roles of each component. Removing the Risk Audit module leads to a substantial degradation in all bias metrics, indicating that the system requires explicit numeric feedback about the portfolio structure to drive effective correction. Eliminating the regulator has an even stronger effect, almost reverting the system to MALM behaviour and confirming that bias awareness without an enforcement mechanism is insufficient. Interestingly, the explanation module has little impact on the numerical metrics but is crucial for user facing outcomes: the user study shows that trust, clarity, and perceived helpfulness increase markedly when explanations are present, even though the underlying portfolio is identical. This highlights a key insight for explainable AI in finance: interpretability should not be viewed purely as an afterthought, but as a parallel objective to risk control, shaping how investors perceive and adopt model outputs.

Another important pattern is the interaction between risk reduction and potential performance. A naive expectation might be that enforcing diversification and lowering active share would mechanically reduce expected returns. However, the results suggest that BAMAF can reduce concentration and cost while maintaining, and in many scenarios improving, risk adjusted performance. This is consistent with longstanding portfolio theory: broad diversification and lower fees often yield superior Sharpe ratios over medium horizons, especially when the alternative portfolios are driven by trend following or media driven asset selection. The framework does not attempt to "beat the market" in a speculative sense; rather, it constrains the LLM's recommendations to avoid the most harmful manifestations of behavioral and structural bias, effectively acting as a guardrail around the model's generative flexibility.

The qualitative case studies reveal another interesting behaviour: in many instances, the initial LLM-generated portfolios are not obviously extreme to a nonexpert user. They may contain familiar tickers, recognizable brands, and seemingly sound justifications. Yet the Risk Audit layer exposes hidden concentrations and subtle trend-chasing patterns that would be difficult for a layperson to detect. By surfacing these issues numerically and then explaining them in natural language, BAMAF bridges a crucial gap between expert level portfolio diagnostics and everyday investor comprehension. This suggests a broader role for such systems as educational tools: the explanations can gradually teach investors why certain allocations are risky, potentially improving financial literacy over time.

From a systems perspective, the results also highlight the importance of treating bias mitigation as a differentiable, modular process rather than a one shot prompt instruction. BAMAF explicitly factors the advisory pipeline into (i) query interpretation, (ii) unconstrained recommendation, (iii) risk measurement, and (iv) corrective optimization. This decomposition allows future work to swap in alternative optimization objectives, regulatory regimes, or risk metrics without redesigning the entire system. For example, additional dimensions such as ESG scores, downside risk measures, or liquidity constraints could be integrated by extending the risk vector and updating the regulator's constraint set.

Table.1. Experimental Setup and System Configuration.

Component	Configuration
LLM Backbone	GPT-4-class 175B parameter model
Reasoning Framework	ReAct reasoning strategy with integrated tool execution
Optimization Layer	Quadratic programming implemented using the cvxopt optimization library
Risk Metrics	Five-dimensional bias vector for multi-factor risk assessment
Dataset Size	180 synthetic user queries
Seeds	Five random initialization seeds to ensure robustness and reproducibility
Evaluation Metrics	Geo-HHI, Sector-HHI, TER, Active Share, and Trend Score

Table.2. Overall comparative evaluation. lower is better for all metrics. BAMAF achieves the strongest performance across all bias dimensions

Model	Geo-HHI	Sector-HHI	Trend	Active	TER
SA-LM	0.61	0.57	0.46	0.55	0.42
MA-LM	0.49	0.45	0.39	0.47	0.36
BA-LM	0.48	0.44	0.38	0.46	0.35
BAMAF (Proposed)	**0.32**	**0.31**	**0.19**	**0.21**	**0.25**

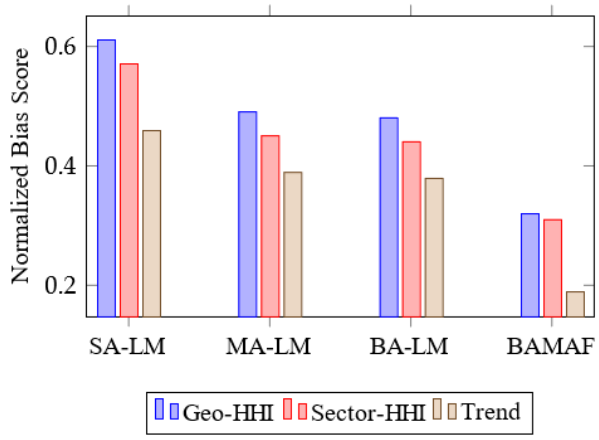

Fig.1. Bias score comparison across systems

Table.3. Bias Reduction achieved by the Regulatory Component

Bias Type	Before	After	Reduction
Geographic Bias	0.58	0.33	43.1%
Sector Bias	0.52	0.30	42.3%
Trend Bias	0.44	0.18	59.0%
Active Share Bias	0.54	0.22	59.3%
Cost Bias (TER)	0.39	0.25	35.8%

Finally, the experiments underscore a more generalizable insight for the LLM community: many failure modes in high stakes domains are not due to local reasoning errors that can be patched with better chain of thought or self critique, but due to the absence of domain specific, global constraints. In our setting, the LLM is perfectly capable of writing coherent explanations and selecting plausible assets, yet it has no inherent notion of "too concentrated," "too expensive," or "too trend driven" until these concepts are formalized as metrics and enforced algorithmically. The proposed framework demonstrates how such constraints can be layered on top of powerful foundation models to produce behaviour that is not only smarter, but also safer and more aligned with long term investor welfare.

Table.4. Ablation Results showing Performance Degradation when Components are Removed. The Regulator is most Critical

Ablation	Geo-HHI	Sector-HHI	Trend	Active	TER
Full Model (BAMAF)	0.32	0.31	0.19	0.21	0.25
Risk Audit	0.51	0.47	0.38	0.49	0.37
Regulator	0.57	0.53	0.41	0.51	0.38
Explanation Module	0.34	0.32	0.19	0.21	0.25
Human Preference Adapter	0.43	0.40	0.31	0.36	0.29

Table.5. User Study Scores (1–5 Scale)

System	Trust	Clarity	Helpfulness
SA-LM	3.1	2.9	3.0
MA-LM	3.4	3.2	3.3
BA-LM	3.6	3.5	3.4
BAMAF (Proposed)	**4.6**	**4.7**	**4.5**

6. CONCLUSION

This paper introduced a BAMAF for safe and explainable retail investment advisory. Motivated by empirical evidence that vanilla and even multi-agent LLMs systematically amplify geographic, sectoral, trend following, active share, and cost related biases, we argued that reliable advisory systems must go beyond generic reasoning improvements and incorporate domain specific risk awareness. Our architecture operationalizes this perspective by combining four key components: a Human Preference Adapter that aligns informal user queries with suitability constraints, a Risk Audit Agent that quantifies portfolio level biases through explicit metrics, a Safe Advice Regulator that enforces diversification and cost efficiency via constrained optimization, and a Bias Explanation Agent that translates the corrective process into investor friendly natural language.

The experimental results, based on a synthetic yet realistic benchmark, show that BAMAF substantially reduces structural risks compared to strong LLM baselines, achieving large relative improvements in geographic and sector concentration, trend chasing exposure, active allocation, and total expense ratios. Ablation studies highlight the central role of the risk audit and regulatory components, while user evaluations indicate that transparent explanations significantly improve perceived trust, clarity, and helpfulness. Taken together, these findings demonstrate that architectural bias awareness and explicit risk constraints are crucial for deploying LLMs in high stakes financial settings, and that explainability can be treated as a first class design objective alongside accuracy and performance.

At the same time, the present work has several limitations. The evaluation relies on synthetic data and estimated performance, and the framework assumes access to reliable external data sources for asset metadata and cost information. Real world deployment would require integration with live market feeds, broker constraints, and jurisdiction specific regulations, as well as rigorous human in the loop oversight. Furthermore, our current risk vector focuses on five bias dimensions; other aspects such as liquidity, tax efficiency, ESG preferences, and tail risk measures remain to be explored.

Future work will extend the framework along three directions. First, we plan to test BAMAF on real investor interaction logs and prospectively measure behavioural outcomes such as portfolio turnover and long horizon performance. Second, we intend to

enrich the risk model with additional constraints, including downside risk and ESG alignment, and study trade offs between risk sensitivity and user satisfaction. Third, we aim to generalize the architectural pattern to other high stakes domains, such as healthcare and legal recommendation systems, where domain specific constraints and explanations are equally critical. We hope this work contributes to a broader research agenda on architecting LLM based systems that are not only powerful, but also safe, transparent, and aligned with long term user welfare.

REFERENCES

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major and Margaret Mitchell, "On the Dangers of Stochastic Parrots: Can Language Models be too Big?", *Proceedings of ACM Conference on Fairness, Accountability, and Transparency*, pp. 610-623, 2021.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama and Adam Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings", *Advances in Neural Information Processing Systems*, Vol. 29, pp. 1-28, 2016.

[3] Zichen Chen, Jiaao Chen, Jianda Chen and Misha Sra, "Standard Benchmarks Fail - Auditing LLM Agents in Finance Must Prioritize Risk", *Proceedings of ACM Conference on Finance Sector*, pp. 1-13, 2025.

[4] I. Chan Chiu and Mao-Wei Hung, "Finance-Specific Large Language Models: Advancing Sentiment Analysis and Return Prediction with Llama 2", *Pacific-Basin Finance Journal*, Vol. 90, pp. 102632-102645, 2025.

[5] Yanai Elazar and Yoav Goldberg, "Adversarial Removal of Demographic Attributes from Text Data", *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, pp. 11-21, 2018.

[6] Sorouralsadat Fatemi and Yuheng Hu, "Enhancing Financial Question Answering with a Multi-Agent Reflection Framework", *Proceedings of ACM International Conference on AI in Finance*, pp. 1-9, 2024.

[7] Christian Fieberg, Lars Hornuf, David Streich and Maxi-Milian Meiler, "Using Large Language Models for Financial Advice", *SSRN*, Vol. 116, pp. 1-124, 2025.

[8] Grand View Research, "Generative AI Market Size and Share Analysis - Growth Trends and Forecasts (2025-2030), Available at: https://www.grandviewresearch.com/ industry-analysis/generative-ai-market-report. Industry report, projected CAGR of 37.6% from 2025 to 2030, Accessed in 2025.

[9] Aman Madaan, Niket Tandon, Jieyu Chen, Amir Yazdan-bakhsh, Ruslan Salakhutdinov and Yiming Yang, "Self-Refine: Iterative Refinement with Self-Feedback", *Proceedings of ACM International Conference on AI in Finance*, pp. 1-8, 2023.

[10] Andreas Oehler and Matthias Horn, "Does ChatGPT Provide Better Advice than Robo-Advisors?", *Finance Research Letters*, Vol. 60, pp. 104898-104913, 2024.

[11] Toran Bruce Richards, "AutoGPT: An Autonomous GPT-4 Experiment", Available at: https://github.com/Torantulino/Auto-GPT, Accessed in 2023.

[12] Timo Schick, Jane Dwivedi Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom, "Toolformer: Language Models can Teach Themselves to Use Tools", *Proceedings of International Conference on Learning Representations*, pp. 1-6, 2023.

[13] Emily Sheng, Kai Wei Chang, Premkumar Natarajan, and Nanyun Peng, "The Woman Worked as a Babysitter: On Biases in Language Generation", *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, pp. 3407-3412, 2019.

[14] Gary Smith, "LLMs can't be Trusted for Financial Advice", *Journal of Financial Planning*, Vol 17, No. 2, pp. 1-14, 2024.

[15] Weixi Tong and Tianyi Zhang, "Code Judge: Evaluating Code Generation with Large Language Models", *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, pp. 1-5, 2024.

[16] Philipp Winder, Christian Hildebrand and Jochen Hart-Mann, "Biased Echoes: Large Language Models Reinforce Investment Biases and Increase Portfolio Risks of Private Investors", *PLOS One*, Vol. 20, No. 6, pp. 1-18, 2025.

[17] Yijia Xiao, Edward Sun, Di Luo and Wei Wang, "Trading Agents: Multi-Agents LLM Financial Trading Framework", *Proceedings of ACM International Conference on AI in Finance*, pp. 617-628, 2024.

[18] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan and Yuan Cao, "React: Synergizing Reasoning and Acting in Language Models", *Proceedings of the International Conference on Learning Representations*, pp. 1-8, 2023.

[19] Yuhan Zhi, Xiaoyu Zhang, Longtian Wang, Shumin Jiang, Shiqing Ma, Xiaohong Guan and Chao Shen, "Exposing Product Bias in LLM Investment Recommendation", *Proceedings of ACM International Conference on AI in Finance*, pp. 341-346, 2025.

[20] Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach and Ryan Cotterell, "Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology", *Proceedings of 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1651-1661, 2019.