AI-AUGMENTED PORTFOLIO OPTIMIZATION:

Integrating LLM-Generated Views within the Black-Litterman Framework

by

MD AMIR KHAN

A THESIS

Submitted to the Faculty of the Stevens Institute of Technology
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
in
FINANCIAL ENGINEERING

____
MD Amir Khan, Candidate

ADVISORY COMMITTEE

____
Steve Yang, Chair Date

____
Papa Ndiaye Date

STEVENS INSTITUTE OF TECHNOLOGY
Castle Point on Hudson
Hoboken, NJ 07030

November 2025

Abstract

Portfolio optimization remains challenging due to the sensitivity of traditional mean–variance models to return estimation errors. The Black–Litterman (BL) model mitigates this limitation by combining equilibrium returns with investor views, yet generating such views systematically remains a key difficulty. This study proposes an enhanced BL framework that integrates predictive views derived from Large Language Models (LLMs) trained on financial data. The approach leverages LLMs to infer expected stock returns from historical price movements, firm fundamentals, and sector-level attributes, while accounting for uncertainty through the variance of model predictions. These probabilistic views are incorporated into the BL Bayesian update, allowing dynamic adjustment of portfolio weights based on confidence levels. The framework is empirically validated on the top 50 constituents of the S&P 500 index from June 2024 to February 2025, using biweekly rebalancing and two-week rolling input windows. Performance is benchmarked against the S&P 500 index, an equal-weight portfolio, and a traditional mean–variance optimized portfolio. Results show that LLM-augmented portfolios achieve improved risk-adjusted returns and greater stability across market regimes, highlighting the potential of language models as scalable, data-driven view generators in modern portfolio management.

Code Availability:

All code, data preprocessing scripts, and model prompts used in this study are available at: https://github.com/Mkhan2317/Master-s_Thesis

Dedication

This thesis is lovingly dedicated to my mother — whose endless love, unwavering faith, and quiet strength have been the guiding light of every step I have taken.

Her sacrifices, resilience, and boundless encouragement have shaped who I am, reminding me that no dream is too distant when it is pursued with sincerity and courage.

Every page of this work carries a reflection of her hope, and every accomplishment stands as a humble tribute to her enduring belief in me.

Acknowledgment

First and foremost, I am profoundly grateful to **Almighty Allah** for granting me the strength, patience, and perseverance for this research. Without His countless blessings and guidance, none of this would have been possible.

I would like to extend my sincere appreciation to **Professor Steve Yang** for his exceptional mentorship, insightful feedback, and steadfast support throughout my thesis journey. His deep knowledge in quantitative finance and portfolio optimization, combined with his encouragement to pursue innovative ideas, has been instrumental in shaping the direction and quality of this work.

My heartfelt thanks also go to **Associate Professor Papa Ndiaye**, who graciously served as the reader of this thesis. His thoughtful comments, valuable suggestions, and encouragement provided clarity and depth to my research, strengthening its academic rigor and relevance.

I am also deeply thankful to **Yang Li**, Ph.D. candidate at Stevens Institute of Technology, whose thoughtful advice, technical guidance, and generous support greatly enriched this project. His willingness to assist and collaborate was both inspiring and invaluable.

To all of them, I owe a debt of gratitude for their time, trust, and intellectual generosity. This thesis stands as a testament to their influence and the collaborative spirit of research.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background: Portfolio Optimization Challenges

Portfolio optimization remains a cornerstone of financial decision-making, yet it continues to pose significant practical challenges. Since the seminal work of Markowitz (1952), the mean-variance optimization (MVO) framework has served as the foundation for asset allocation by seeking to balance expected return against risk. However, this model is highly sensitive to input estimation errors—particularly in expected returns and the covariance matrix of asset returns. Even minor perturbations in these parameters can lead to disproportionately large shifts in optimal weights, often resulting in extreme and unstable portfolios that lack robustness in real-world applications Michaud (1989).

To address these shortcomings, the Black-Litterman model Black and Litterman (1992) was introduced as a Bayesian extension that combines market equilibrium returns with subjective investor views. This formulation improves stability by introducing a prior and blending it with user-defined expectations. Yet, its effectiveness is fundamentally constrained by the challenge of generating consistent, high-quality views—especially under volatile market conditions. Traditional methods for forming these views rely either on analyst expertise or quantitative forecasting models, both of which demand substantial domain knowledge and are susceptible to behavioral biases, overfitting, or lack of scalability Idzorek (2007).

Moreover, as the dimensionality of asset universes increases, so does the complexity of the optimization problem. Estimating and inverting large covariance matrices becomes computationally intensive and error-prone. Compounding the issue, most conventional frameworks fail to utilize rich unstructured data—such as earnings call transcripts, financial news, and social media commentary—that often contain valuable insights into asset-level fundamentals and macroeconomic sentiment Zhao et al. (2024). The exclusion of these alternative data sources limits the model's responsiveness to real-time market dynamics.

It is against this backdrop that this thesis seeks to make a methodological contribution. The core challenge addressed here is the systematic generation and integration of forward-looking investor views—using a framework that is data-driven, scalable, and adaptable to fast-changing financial environments. Recent advancements in natural language processing, particularly large language models (LLMs), offer a novel solution to this problem. By leveraging the text understanding capabilities of LLMs, it becomes possible to automate the generation of return forecasts derived from diverse financial text corpora. These model-generated views can then be integrated into the Black-Litterman framework through a Bayesian updating mechanism that accounts for uncertainty in LLM predictions.

This research aims to design such an integrated framework, wherein LLMs act as the mechanism for generating structured, interpretable views based on textual and numerical data, and these views are embedded into the Black-Litterman model using robust optimization techniques. The ultimate goal is to enhance the reliability and adaptability of portfolio construction by combining financial theory with modern AI-based information extraction—laying the foundation for more dynamic, transparent, and information-aware investment strategies.

1.2 Role of AI in Financial Decision-Making

Artificial Intelligence (AI) is transforming financial decision-making by enabling systems that can synthesize vast amounts of structured and unstructured data, uncover latent patterns, and adapt to changing market environments. Traditional financial models, such as mean-variance optimization, depend on rigid assumptions and static inputs, which often render them brittle in the face of market complexity and noise. In contrast, AI methods, particularly those based on deep learning and large language models (LLMs), allow for a more dynamic, data-driven approach to asset allocation and risk assessment Zhao et al. (2024); Hwang et al. (2025).

A notable advancement in this field is the integration of LLMs into the Black-Litterman portfolio optimization framework. The Black-Litterman model has long been recognized for its ability to incorporate subjective investor views into market equilibrium returns. However, one persistent limitation has been the difficulty of generating those views in a scalable and consistent manner. LLMs offer a solution by leveraging natural language understanding and numerical reasoning to generate forward-looking return estimates based on structured data such as price histories and metadata, as well as unstructured information like news and corporate disclosures Lee et al. (2025). These models can be systematically prompted to simulate analyst-like behavior, generating predictions along with confidence intervals, which are then embedded into the Black-Litterman model through the construction of a view vector $\mathbf{q}$, a picking matrix $\mathbf{P}$, and a confidence matrix $\boldsymbol{\Omega}$.

Recent empirical studies underscore the effectiveness of this AI-enhanced approach. For instance, Lee et al. (2025) evaluated portfolios optimized using views generated by various LLMs, including LLaMA-3, Qwen-2, and GPT-4o-mini, and compared them to traditional mean-variance and equal-weighted portfolios. The study demonstrated that models like LLaMA-3 consistently produced more stable and discriminative forecasts, leading to superior cumulative returns and Sharpe ratios over an eight-month test period Lee et al. (2025). These findings suggest that LLMs not only replicate human-style reasoning but may outperform traditional heuristics due to their scalability and contextual awareness.

Moreover, the role of AI extends beyond view generation. Reinforcement learning frameworks are being used for dynamic rebalancing, while graph neural networks (GNNs) are applied to model cross-asset dependencies and systemic risk Feng et al. (2023). There is also growing interest in multimodal AI systems that fuse text, time series, and macroeconomic indicators into unified decision architectures Hwang et al. (2025); Zhao et al. (2024).

Despite their promise, AI-driven systems pose new challenges related to transparency, bias, and robustness. Issues such as hallucination in LLMs or adversarial sensitivity in neural networks necessitate rigorous validation and governance protocols. However, with the emergence of explainable AI (XAI) and uncertainty-aware modeling, many of these concerns are being addressed, allowing AI to act as a powerful augmentation layer to human expertise in financial contexts.

In sum, AI—particularly LLMs—represents a major leap forward in financial decision-making, enabling more responsive, scalable, and integrated portfolio management strategies. When embedded into frameworks like the Black-Litterman model, these tools shift the paradigm from subjective opinion-based allocation to a robust, data-driven optimization process that adapts fluidly to the evolving landscape of global markets.

1.3 Research Objectives, Hypothesis, and Contributions

The primary objective of this thesis is to develop a next-generation portfolio optimization framework that unites the interpretability and theoretical rigor of classical finance with the adaptive reasoning and contextual understanding of modern Large Language Models (LLMs). The research centers on extending the Black-Litterman (BL) model by embedding predictive views generated by LLMs that draw insights from both structured market data and unstructured textual information.

In traditional portfolio construction, the estimation of expected returns is highly uncertain, often resulting in unstable or inefficient allocations. The Black-Litterman model mitigates this problem by integrating market equilibrium returns with investor views within a Bayesian framework. However, the practical implementation of the model remains constrained by the challenge of formulating consistent, data-driven, and scalable views that accurately reflect evolving market conditions.

This thesis advances the hypothesis that fine-tuned LLMs—when guided by domain-specific prompts and financially relevant signals—can systematically generate informed, context-aware, and forward-looking views. Drawing upon existing research in financial natural language processing, Bayesian portfolio theory, and robust optimization, this study posits that integrating LLM-generated views—weighted by their confidence and adjusted for volatility-based uncertainty—into the Black-Litterman model can lead to superior portfolio outcomes. The anticipated benefits include higher risk-adjusted returns, enhanced robustness across diverse market regimes, and improved interpretability of investment decisions.

Three core propositions support this hypothesis. **First**, LLMs can be trained and prompted to infer directional asset return expectations by synthesizing information from earnings calls, macroeconomic reports, analyst commentaries, and sentiment indicators. **Second**, the uncertainty inherent in these forecasts can be quantified using ensemble variance or inter-prompt disagreement, allowing each view to be weighted according to its estimated reliability in the posterior distribution. **Third**, incorporating a robust optimization layer can further suppress the influence of low-confidence or noisy views, reducing overfitting and ensuring greater portfolio stability during volatile market periods.

Within this proposed structure, the relationships among LLM-generated views, confidence measures, posterior return estimates, and final portfolio weights are designed such that high-confidence and contextually justified views exert a proportionally greater influence on the optimization process. This aligns with empirical evidence in quantitative finance suggesting that reliability-weighted information enhances both performance and stability in portfolio construction.

The contributions of this research are both theoretical and practical. Conceptually, it introduces a novel, data-driven approach for generating investor views using LLMs, integrated within a transparent and modular portfolio optimization pipeline. Methodologically, it extends the Black-Litterman model to incorporate non-human, machine-generated perspectives while maintaining theoretical coherence and Bayesian consistency. Finally, the proposed framework is designed for adaptability—capable of evolving alongside advances in language model architectures, shifts in financial narratives, and the availability of multimodal data sources that include numerical, textual, and sentiment-based information.

1.4 Research Structure & Outline

This thesis is organized to provide a coherent progression from theoretical foundations to empirical validation, ensuring a systematic exploration of large language model (LLM)-enhanced portfolio optimization.

Chapter 1 introduces the research background, outlining the motivation, objectives, and contributions of this work. It frames the limitations of traditional portfolio optimization—such as parameter instability, myopic rebalancing, and narrow asset universes—and motivates the integration of LLMs into the Black–Litterman framework to address these gaps.

Chapter 2 provides a comprehensive review of the literature, beginning with classical approaches like the Mean–Variance Portfolio (MVP), Equal-Weighted Portfolio (EWP), and Risk Parity (RP). It then examines modern hierarchical and graph-based methods, including Hierarchical Risk Parity (HRP) and Hierarchical Equal Risk Contribution (HERC). The chapter concludes with an in-depth discussion of the Black–Litterman model, its robust extensions, and recent advances incorporating machine learning and LLM-based forecasting.

Chapter 3 details the methodological framework. It first describes the dataset construction process, including the collection, cleaning, and structuring of both historical market and textual data. It then explains the LLM-based view generation framework, encompassing prompt design, temporal anchoring, and uncertainty quantification. The chapter further elaborates on the enhanced Black–Litterman integration, Bayesian updating procedures, and optimization design—culminating in the complete pipeline for portfolio construction, calibration, and evaluation.

Chapter 4 reports the empirical findings. It analyzes portfolio performance across different LLMs (e.g., LLaMA, Gemma, Qwen, GPT) and contrasts them with traditional optimization strategies. The results are discussed in terms of return efficiency, risk control, and interpretability, highlighting the advantages and limitations of integrating LLM-based predictive signals into the Black–Litterman framework.

Chapter 5 concludes the thesis by summarizing key contributions and theoretical implications. It discusses the limitations of the current research, proposes future directions—including multi-asset extensions, real-time portfolio rebalancing, and reinforcement learning integration—and reflects on the broader implications of AI-augmented decision systems in financial portfolio management.

Chapter 2

Related Work

2.1 Mean-Variance Portfolio (MVP)

The rational investor faces an eternal dilemma: how to allocate capital among risky assets to maximize returns while minimizing exposure to risk. This trade-off is formalized in the mean-variance framework introduced by Harry Markowitz in 1952, a model that laid the groundwork for modern portfolio theory Markowitz (1952). The mean-variance approach is built on the premise that investors are risk-averse and that the desirability of a portfolio depends not only on the expected return but also on the risk associated with that return. The goal is not merely to maximize returns, but to achieve the best possible returns per unit of risk.

2.1.1 Risk Return Tradeoff

In the mean-variance framework, the return of a portfolio is defined as the expected value of its component assets' returns, and risk is defined as the standard deviation (or variance) of those returns. Let μ_i represent the expected return of asset i, and let Σ be the covariance matrix of asset returns. The expected return of a portfolio with weights $w = (w_1, w_2, ..., w_n)^T$ is given by:

$$R_p = w^T \mu \tag{1}$$

where $\mu = (\mu_1, \mu_2, ..., \mu_n)^T$ is the vector of expected returns.

The risk, or variance of the portfolio return, is:

$$\sigma_p^2 = w^T \Sigma w \tag{2}$$

The trade-off between these two quantities leads to the concept of the efficient frontier—the set of portfolios that achieve the maximum expected return for a given level of risk, or equivalently, the minimum risk for a given level of return Elton et al. (2014). Every point on the efficient frontier represents an optimal portfolio under the mean-variance criterion.

2.1.2 Mathematical Formulation

The general optimization problem in mean-variance analysis is formulated as:

$$\min_{w} \; w^T \Sigma w \quad \text{subject to} \quad w^T \mu = \mu_p, \quad \sum_{i=1}^{n} w_i = 1 \tag{3}$$

This objective function minimizes the portfolio variance (risk) while satisfying constraints on the target return μ_p and ensuring the portfolio is fully invested. This is a quadratic programming problem, and its solution gives the optimal weights w^* for a portfolio on the efficient frontier Boyd and Vandenberghe (2004).

A specific case is the Global Minimum Variance Portfolio (GMVP), which minimizes variance regardless of return:

$$\min_{w} \; w^T \Sigma w \quad \text{subject to} \quad \sum_{i=1}^{n} w_i = 1 \tag{4}$$

This formulation leads to a closed-form solution:

$$w^* = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^T \Sigma^{-1} \mathbf{1}} \tag{5}$$

where $\mathbf{1}$ is a vector of ones. This portfolio provides the lowest possible variance among all portfolios of risky assets Jagannathan and Ma (2003).

2.1.3 Limitations

While the mean-variance framework is foundational, it suffers from significant sensitivity to input estimates, particularly the expected return vector μ. Small errors in estimation can lead to extreme and unstable portfolio weights Michaud (1989). This issue has led to the development of improved approaches such as shrinkage estimators Ledoit and Wolf (2003) and robust optimization techniques Ben-Tal and Nemirovski (2000).

One particularly influential extension is the **Black-Litterman model** Black and Litterman (1992), which blends market equilibrium returns with subjective or model-based views using a Bayesian framework. By integrating prior beliefs with quantitative forecasts, Black-Litterman produces more stable and intuitive portfolios, especially when incorporating multiple views with varying confidence levels.

In recent work, machine learning and LLM-based forecasts have been integrated into these frameworks to improve return estimation and enhance robustness Gu et al. (2020); Lee et al. (2025).

2.2 Equal-Weighted Portfolio (EWP)

In the evolving landscape of portfolio theory, where mathematical elegance often dictates strategic decisions, the Equal-Weighted Portfolio (EWP) stands out as a counterpoint to complexity. While the mean-variance framework of Markowitz seeks to optimally balance risk and return using estimates of expected returns and covariances, the equal-weighted portfolio represents an important special case: a portfolio where all assets are weighted equally due to a complete absence of forecasts. In fact, the EWP can be interpreted as the limiting case of the mean-variance portfolio when estimation error in returns is so severe that it is rational to assign zero confidence to all inputs.

As such, the equal-weighted portfolio emerges not in contradiction to optimization but as a response to its limitations. Its simplicity belies its long-standing presence in both academic studies and real-world asset management. The idea is disarmingly straightforward: allocate capital equally across all available assets. Yet, this simplicity has prompted extensive academic debate, empirical scrutiny, and practical adoption, especially when contrasted with more elaborate frameworks such as the mean-variance optimization of Markowitz.

2.2.1 Appeal of Simplicity and Diversification

The equal-weighted strategy assumes a uniform distribution of belief across all assets, with no attempt to forecast returns, volatilities, or correlations. Each asset receives the same weight:

$$w_i = \frac{1}{n} \quad \text{for all } i = 1, ..., n \tag{6}$$

where n is the total number of assets in the portfolio. This uniform allocation implicitly promotes diversification and is free from the estimation risk that plagues optimized portfolios. Studies such as DeMiguel et al. (2009) have shown that in many empirical contexts, the 1/n rule often outperforms more sophisticated models out-of-sample, particularly when those models rely heavily on noisy estimates.

The EWP also naturally rebalances over time. As asset prices fluctuate, maintaining equal weights requires periodic buying and selling of outperforming and underperforming assets, which introduces an implicit contrarian mechanism. This mechanical discipline can help capture mean-reversion effects in asset returns, contributing to the EWP's surprisingly robust performance across multiple market cycles Benartzi and Thaler (2001).

2.2.2 Mathematical Formulation and Risk Properties

To examine the EWP from a formal quantitative perspective, consider a portfolio of n risky assets. Let $r = (r_1, r_2, ..., r_n)^T$ be the random vector of asset returns and $\mu = \mathbb{E}[r]$ the vector of expected returns. The equal-weighted portfolio assigns weights:

$$w = \frac{1}{n}\mathbf{1} \tag{7}$$

where $\mathbf{1} \in \mathbb{R}^n$ is a vector of ones. The expected return of the EWP is then:

$$\mathbb{E}[R_p] = w^T\mu = \frac{1}{n}\mathbf{1}^T\mu \tag{8}$$

This expression shows that the EWP return is simply the average of the individual expected returns. The portfolio variance is given by:

$$\text{Var}(R_p) = w^T\Sigma w = \frac{1}{n^2}\mathbf{1}^T\Sigma\mathbf{1} \tag{9}$$

where Σ is the $n \times n$ covariance matrix of asset returns. This indicates that the risk of the equal-weighted portfolio declines with the square of the number of assets, assuming some degree of diversification.

To provide further insight, define the average variance and average pairwise covariance as:

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 \tag{10}$$

$$\bar{\rho} = \frac{2}{n(n-1)}\sum_{i<j}\rho_{ij} \tag{11}$$

Then, under equal weighting, the portfolio variance can also be decomposed as:

$$\sigma_p^2 = \frac{1}{n}\bar{\sigma}^2 + \frac{n-1}{n}\bar{\rho}\bar{\sigma}^2 \tag{12}$$

This shows that even in an equal-weighted setup, the covariance structure matters. As n increases, the influence of individual variances declines, but average correlation becomes a dominant component. This formulation also explains why the EWP performs well in diversified markets but may underperform when correlations are high.

2.2.3 Drawbacks

While the EWP's simplicity is its strength, it is also its most notable limitation. By ignoring differences in expected returns, volatilities, and correlations, the strategy may underperform optimized portfolios in environments where accurate forecasts are available. Moreover, equal weighting can lead to implicit overexposure to small-cap or volatile assets, which may be riskier or less liquid Clarke et al. (2006).

Another drawback is the portfolio's lack of adaptability. Because it applies the same rule irrespective of macroeconomic conditions or regime shifts, the EWP is less responsive to structural changes in markets. Nonetheless, its robustness and resistance to estimation error make it a valuable benchmark. It often serves as a point of comparison in studies evaluating new models or alternative weighting schemes.

The Equal-Weighted Portfolio offers a powerful narrative within portfolio management: a strategy grounded in simplicity that often delivers results competitive with, and sometimes superior to, more complex alternatives. While it eschews optimization and predictive modeling, it compensates with robustness, transparency, and reliability—qualities that are increasingly prized in a world awash with uncertainty.

2.3 Risk Parity

The development of the Risk Parity Portfolio arises as a response to the limitations observed in both the mean-variance optimized (MVO) portfolio and the equal-weighted portfolio (EWP). While the MVO framework is elegant and theoretically sound, it suffers from estimation sensitivity—especially to expected returns—which often leads to unstable and unintuitive allocations. On the other hand, the EWP offers robustness and simplicity, but completely ignores information about asset volatilities and correlations. These shortcomings motivated the design of risk-based allocation methods that distribute risk more evenly across portfolio components.

The Risk Parity (RP) framework was introduced to bridge this gap: it avoids the estimation of expected returns but retains sensitivity to volatility and correlation. The guiding principle is simple yet powerful: construct a portfolio in which each asset contributes equally to overall portfolio risk. This approach ensures that no single asset or asset class dominates the portfolio's risk, thereby enhancing diversification not by capital allocation, but by risk contribution Maillard et al. (2010).

2.3.1 Rationale and concept of Equal Risk Contribution

Traditional allocation schemes like MVO focus on return maximization, and EWP on capital symmetry. In contrast, Risk Parity is based on risk symmetry. For a portfolio with weights $w = (w_1, ..., w_n)^T$ and covariance matrix Σ, the total portfolio variance is:

$$\sigma_p^2 = w^T \Sigma w \tag{13}$$

The marginal contribution to risk of asset i is given by:

$$\partial_{w_i} \sigma_p = \frac{(\Sigma w)_i}{\sigma_p} \tag{14}$$

The risk contribution (RC) of asset i is then:

$$RC_i = w_i \cdot \partial_{w_i} \sigma_p = \frac{w_i (\Sigma w)_i}{\sigma_p} \tag{15}$$

In a Risk Parity portfolio, the objective is to ensure that all RC_i are equal:

$$RC_1 = RC_2 = ... = RC_n = \frac{1}{n} \sigma_p \tag{16}$$

This equality condition leads to a system of nonlinear equations, and finding the exact solution often requires numerical optimization. Nonetheless, the resulting portfolios are more diversified in terms of volatility exposure and tend to avoid over-concentration in low-volatility assets, a flaw common in MVO.

2.3.2 Mathematical Formulation and Implementation

The formal risk parity optimization problem can be expressed as a minimization of the squared deviations of risk contributions:

$$\min_w \sum_{i=1}^{n} \sum_{j=1}^{n} (RC_i - RC_j)^2 \quad \text{subject to} \quad \sum_{i=1}^{n} w_i = 1, \quad w_i \geq 0 \tag{17}$$

Alternatively, the optimization may be approximated using log-barrier or inverse-volatility methods. A commonly used heuristic is the inverse volatility portfolio:

$$w_i^{\text{inv-vol}} = \frac{\frac{1}{\sigma_i}}{\sum_{j=1}^{n} \frac{1}{\sigma_j}} \tag{18}$$

This ignores correlations but gives a rough approximation of risk balance when off-diagonal covariances are small.

More advanced implementations incorporate full covariance matrices, and iterative numerical methods such as cyclical coordinate descent or convex optimization solvers are used to satisfy the equal-risk condition. These portfolios are less sensitive to expected return misestimation and maintain robustness across different market regimes DeMiguel et al. (2020).

2.3.3 Trade-offs, Advantages, and Criticisms

Risk parity portfolios offer several compelling advantages. First, they are model-agnostic with respect to expected returns, relying only on observable volatility and correlation. This reduces input risk and improves out-of-sample robustness Roncalli (2013). Second, RP portfolios naturally overweight less volatile assets and underweight more volatile ones, which often leads to countercyclical behavior and improved drawdown protection.

However, risk parity is not without its criticisms. The strategy tends to overweight low-volatility assets such as government bonds, potentially leading to excessive duration risk, especially in low interest rate environments. Moreover, in markets where volatility regimes change rapidly, the strategy may lag in responsiveness. Transaction costs may also be high due to the frequent rebalancing required to maintain equal risk contributions Clarke et al. (2011).

Nonetheless, as a middle ground between overfitted optimization and blind equal weighting, risk parity has gained prominence among institutional investors and asset allocators. It embodies a practical trade-off: eschewing return forecasts while still incorporating meaningful structure from observed risk relationships.

2.4 Hierarchical Risk Parity (HRP)

The Hierarchical Risk Parity (HRP) portfolio emerged from a critical analysis of the weaknesses inherent in traditional mean-variance optimization (MVO), equal-weighted portfolios (EWP), and even risk parity (RP) methods. MVO, while theoretically elegant, relies on fragile inputs and is highly sensitive to estimation error. EWP avoids estimation entirely but sacrifices any notion of relative risk. Risk parity addressed some of these concerns by balancing contributions to total portfolio risk, yet it too requires a reliable estimation of the covariance matrix and can produce unstable results when that matrix is ill-conditioned.

To overcome these structural and statistical limitations, López de Prado (2016) introduced the HRP methodology. Rather than optimizing portfolios through direct inversion or reliance on noisy covariance matrices, HRP applies techniques from machine learning—specifically hierarchical clustering—to construct portfolios that respect the underlying structure of asset relationships. It eliminates the need for inverting the covariance matrix and replaces traditional optimization with a recursive bisection algorithm based on asset tree structures.

2.4.1 Clustering and the Motivation for Hierarchical Diversification

The HRP framework begins by estimating a distance matrix from asset correlations:

$$d_{ij} = \sqrt{2(1 - \rho_{ij})} \tag{19}$$

where ρ_{ij} is the Pearson correlation between asset i and j. This distance matrix is used to construct a hierarchical clustering tree using algorithms like single linkage, complete linkage, or Ward's method. The result is a binary tree structure (dendrogram) that encodes the similarity relationships between assets.

The insight driving HRP is that financial markets are not flat—they are hierarchical. Assets within sectors tend to co-move more closely than assets across sectors. By identifying and incorporating this nested structure, HRP builds more interpretable and robust portfolios.

2.4.2 Mathematical Construction of HRP Portfolios

The HRP algorithm proceeds in three distinct steps:

Step 1: Hierarchical Clustering. Compute the correlation matrix $\mathbf{R}$ of asset returns and convert it to a distance matrix $\mathbf{D}$ using Equation (4.1). Apply hierarchical clustering (e.g., using Ward's method) to produce a linkage matrix and corresponding dendrogram.

Step 2: Quasi-Diagonalization. The covariance matrix Σ is permuted to follow the hierarchical order using the quasi-diagonalization process. Let P be the permutation matrix derived from the leaf order of the tree, then the quasi-diagonalized matrix is:

$$\Sigma^{\text{qd}} = P \Sigma P^T \tag{20}$$

This reordering attempts to place similar assets adjacent to one another, minimizing crosscluster covariances.

Step 3: Recursive Bisection Allocation. The portfolio weights are assigned using a recursive bisection approach. Let C be a cluster split into subclusters A and B, with total variances given by:

$$\sigma_A^2 = \mathbf{w}_A^T \Sigma_A \mathbf{w}_A, \quad \sigma_B^2 = \mathbf{w}_B^T \Sigma_B \mathbf{w}_B \tag{21}$$

If $\mathbf{w}_A = \mathbf{1}/|A|$ and $\mathbf{w}_B = \mathbf{1}/|B|$ are initially uniform, then the allocation to cluster A is:

$$w_A = \frac{\sigma_B^{-1}}{\sigma_A^{-1} + \sigma_B^{-1}}, \quad w_B = 1 - w_A \tag{22}$$

This process is applied recursively until all individual assets receive a final allocation. The resulting portfolio is thus built top-down from the hierarchical structure, using risk metrics at each branching point.

Importantly, HRP avoids inversion of the covariance matrix entirely, making it particularly attractive when the number of assets is large relative to the sample size, where traditional methods become unstable Martin et al. (2020).

2.4.3 Drawbacks and Considerations

HRP provides a compelling solution in settings where estimation error is substantial and assets exhibit latent structure. By integrating clustering, HRP reduces the impact of noisy correlations and produces sparse, interpretable weightings. Moreover, it avoids the pitfalls of covariance inversion and mitigates concentration issues inherent in both MVO and RP Martin et al. (2020).

That said, HRP is not without limitations. It relies heavily on the chosen linkage method and the stability of the correlation matrix, and it does not explicitly account for expected returns. However, HRP portfolios are highly robust, easy to compute, and scale well with the number of assets, making them appealing in high-dimensional and noisy environments.

In this way, HRP advances the evolution of portfolio construction by combining statistical learning with financial intuition. It demonstrates that structural insights—when leveraged properly—can offer powerful tools for navigating uncertainty in asset allocation.

2.5 Hierarchical Equal Risk Portfolio (HER)

The Hierarchical Equal Risk (HER) portfolio builds on the conceptual foundation of both Hierarchical Risk Parity (HRP) and traditional Risk Parity (RP), addressing residual limitations of these models. Mean-variance optimization (MVO) is prone to instability due to estimation errors, equal-weighting (EWP) ignores risk entirely, and RP assumes a flat risk space without regard for asset structure. HRP overcame some of these challenges by integrating hierarchical clustering and avoiding matrix inversion, but it does not ensure equal risk contribution at the asset level.

HER was developed to combine the robust hierarchical clustering methodology of HRP with the risk contribution balancing goal of RP, ensuring that risk is equitably distributed while accounting for the natural structure of the market Pfitzinger and Winker (2021); Raffinot (2021).

2.5.1 Clustering and Recursive Risk Equalization

Like HRP, the HER portfolio starts with a correlation matrix $\mathbf{R}$ and transforms it into a distance matrix $\mathbf{D}$:

$$d_{ij} = \sqrt{2(1-\rho_{ij})} \tag{23}$$

This is followed by hierarchical clustering (e.g., via Ward's method), producing a dendrogram representing the nested structure of asset relationships. The key innovation in HER lies in how it traverses this tree: rather than allocating weights based on cluster variance alone (as in HRP), HER recursively balances the marginal risk contributions of each cluster.

2.5.2 Mathematical Construction of HER Portfolios

Let w be the weight vector and Σ the covariance matrix. Define the marginal risk contribution of asset i as:

$$MRC_i = \frac{(\Sigma w)_i}{\sqrt{w^T \Sigma w}} \tag{24}$$

The risk contribution is then:

$$RC_i = w_i \cdot MRC_i = \frac{w_i (\Sigma w)_i}{\sqrt{w^T \Sigma w}} \tag{25}$$

In HER, the goal is to recursively equalize the total risk contributions of all sub-clusters A and B at each node of the dendrogram. Given subcluster variances:

$$\sigma_A^2 = \mathbf{w}_A^T \Sigma_A \mathbf{w}_A, \quad \sigma_B^2 = \mathbf{w}_B^T \Sigma_B \mathbf{w}_B \tag{26}$$

HER determines the allocation between A and B such that:

$$RC_A = RC_B \quad \Rightarrow \quad w_A = \frac{\sigma_B}{\sigma_A + \sigma_B}, \quad w_B = 1 - w_A \tag{27}$$

This process continues recursively until leaf nodes (individual assets) are assigned final weights, ensuring equal risk distribution across the full hierarchy.

2.5.3 Drawbacks and Empirical Observations

HER combines the stability and interpretability of HRP with the risk-centric focus of RP. It has been shown to produce well-diversified portfolios with lower turnover than traditional risk parity, better robustness than mean-variance optimization, and higher stability under changing correlation structures Pfitzinger and Winker (2021); Raffinot (2021).

However, HER does not address return forecasts explicitly and inherits the dependence on correlation clustering methods. Its practical success relies on the quality of the hierarchical structure extracted from the data, which may vary depending on market regime.

Nonetheless, HER presents a powerful and principled extension to the hierarchy-aware class of portfolio construction techniques, delivering interpretable and balanced allocations without resorting to fragile optimization procedures.

2.6 Graph Theory-Based Portfolios

Graph-theoretic portfolio construction has emerged as a powerful paradigm in modern finance, offering a robust alternative to traditional methods like mean-variance optimization (MVO), risk parity (RP), and hierarchical approaches. By representing assets as nodes in a network, with edges encoding statistical similarities (e.g., correlations), this approach captures the complex interdependencies among assets, facilitating more informed diversification strategies Ciciretti and Pallotta (2024); Fazli et al. (2023).

2.6.1 From Correlation to Graph Structure

The initial step involves constructing a distance matrix from the asset correlation matrix $\mathbf{R}$:

$$d_{ij} = \sqrt{2(1 - \rho_{ij})} \tag{28}$$

This distance matrix serves as the foundation for building a graph $G = (V, E)$, where each node $v_i \in V$ represents an asset, and edges $e_{ij} \in E$ represent the similarity between assets. Techniques like the Minimum Spanning Tree (MST) are employed to filter the most significant connections, reducing noise and highlighting the core structure of asset interrelations Mantegna (1999).

2.6.2 Portfolio Weighting via Network Metrics

Once the graph is established, various network metrics inform the allocation of portfolio weights. A common approach assigns weights inversely proportional to a node's centrality measure $C(i)$:

$$w_i = \frac{\frac{1}{C(i)}}{\sum_{j=1}^{n} \frac{1}{C(j)}} \tag{29}$$

This method penalizes highly connected nodes, promoting diversification by allocating more weight to peripheral assets. Advanced models, like Network Risk Parity (NRP), leverage eigenvector centrality to determine weights, ensuring a more balanced risk distribution across the network Ciciretti and Pallotta (2024).

2.6.3 Advantages, Stability, and Limitations

Graph-based portfolio construction techniques complement the recursive and topological principles seen in hierarchical models. By leveraging the full network of relationships among assets, these methods extend beyond binary clustering to account for nuanced systemic interactions. They enhance diversification and avoid matrix inversion, increasing robustness against estimation error. However, they also depend on correlation stability and require periodic re-estimation of the graph structure to remain effective. Despite these limitations, graph-theoretic models stand out as a scalable, adaptive, and intuitive framework for structure-aware asset allocation.

2.7 Black-Litterman Model and Extensions

2.7.1 From Equilibrium to Posterior: The Core of Black-Litterman

The Black-Litterman (BL) model represents one of the most influential innovations in modern asset allocation, designed to address the limitations of the classical mean-variance optimization (MVO) framework. Traditional MVO suffers from extreme sensitivity to input estimates, particularly expected returns, which are notoriously difficult to forecast accurately. The BL model, introduced by Black and Litterman (1992), combines the equilibrium returns implied by the Capital Asset Pricing Model (CAPM) with subjective investor views, yielding a posterior distribution that stabilizes portfolio weights while integrating expert opinion or model-based forecasts.

2.7.2 Mathematical Formulation

To develop the Black-Litterman model formally, begin with the equilibrium implied excess returns, denoted π. These are computed using the market capitalization weights w_{mkt} and a scalar τ, which reflects the uncertainty associated with the prior:

$$\pi = \tau \Sigma w_{mkt} \tag{30}$$

Here, $\Sigma \in \mathbb{R}^{n \times n}$ is the covariance matrix of returns, and τ is a scaling parameter expressing the investor's confidence in the prior relative to the views. The equilibrium return vector π represents the investor's starting belief based on market consensus.

The investor also holds subjective views about expected returns. These are encoded in the matrix $P \in \mathbb{R}^{k \times n}$ and vector $q \in \mathbb{R}^{k}$, where each row of P represents a linear constraint on asset returns (e.g., "Asset A will outperform Asset B by 3

$$q = P\mu + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Omega) \tag{31}$$

The prior belief is that the mean return vector μ is distributed as:

$$\mu \sim \mathcal{N}(\pi, \tau\Sigma) \tag{32}$$

Using standard Bayesian updating rules, the posterior distribution of μ is derived by combining the two sources of information—market equilibrium and subjective views. The posterior mean μ_{BL} is:

$$\mu_{BL} = \left[(\tau\Sigma)^{-1} + P^T \Omega^{-1} P\right]^{-1} \left[(\tau\Sigma)^{-1}\pi + P^T \Omega^{-1} q\right] \tag{33}$$

This is a weighted average of the prior mean π and the view vector q, where the weights depend on the respective uncertainties ($\tau\Sigma$ and Ω). The posterior covariance matrix is:

$$\Sigma_{BL} = \left[(\tau\Sigma)^{-1} + P^T \Omega^{-1} P\right]^{-1} \tag{34}$$

The resulting posterior distribution $\mu_{BL} \sim \mathcal{N}(\mu_{BL}, \Sigma_{BL})$ provides a new estimate of expected returns that balances market data and investor opinion. The weights of the optimal portfolio can then be calculated using:

$$w^* = \frac{1}{\lambda} \Sigma^{-1} \mu_{BL} \tag{35}$$

where λ is the investor's risk aversion coefficient. The BL model thus offers a closed-form, analytically tractable solution for generating consistent, well-posed return estimates and portfolio weights.

In practical implementations, τ is often set to a small value (e.g., 0.025) to reflect modest uncertainty in prior beliefs. The view confidence matrix Ω can be diagonal (independent views) or full (correlated views), and may be calibrated using predictive variance, cross-validation, or backtesting metrics.

Recent research introduces extensions for incorporating machine learning-based forecasts, including LLM-generated sentiment signals Lee et al. (2025). These views are embedded into the q vector, with model uncertainty driving Ω. This integration supports real-time Bayesian assimilation of macroeconomic narratives, financial news, and analyst tone into the portfolio construction process.

2.7.3 Limitations and Extensions

Despite its mathematical elegance, the Black-Litterman model is not without limitations. A key challenge lies in specifying the input parameters—particularly the view matrix P, the confidence matrix Ω, and the scalar τ. These choices are often subjective, and estimation errors can distort posterior results. Furthermore, the model still relies on the assumption of multivariate normality and linear view structure, which may not hold in complex or nonlinear market environments.

To address these issues, several extensions have been proposed. The *Bayesian Black-Litterman* variant accounts for parameter uncertainty by introducing prior distributions not only on μ but also on Σ Satchell and Scowcroft (2001). The *Dynamic Black-Litterman Model* incorporates time-varying covariances and learning about views through Kalman filters He et al. (2013). The *Robust Black-Litterman* model incorporates model misspecification by optimizing for worst-case returns under ambiguity Glasserman and Xu (2022).

Each extension enhances specific properties of the original framework—robustness, adaptivity, or estimation stability—making it a versatile foundation for quantitative portfolio management. In practical deployment, LLM-augmented BL models can be updated as new textual data arrive, recalibrating forecasts and rebalancing portfolios dynamically in response to sentiment shifts or geopolitical updates.

2.7.4 Robust Portfolio Optimization

Robust portfolio optimization (RPO) addresses one of the key deficiencies of traditional and Bayesian models such as MVO and Black-Litterman: sensitivity to parameter misspecification. While the BL framework elegantly blends equilibrium priors with investor views, it remains vulnerable to the uncertainties embedded in the mean and covariance estimates, as well as in the confidence matrices $\tau\Sigma$ and Ω. RPO augments this by formulating the portfolio problem as a worst-case optimization over an uncertainty set.

Let μ_{BL} denote the posterior mean vector derived from the Black-Litterman model, and Σ the associated risk estimate. The robust optimization problem is then written as:

$$\min_{w} \max_{\mu \in \mathcal{U}} -w^T \mu + \frac{\lambda}{2} w^T \Sigma w \tag{36}$$

where $\mathcal{U}$ is the uncertainty set capturing all plausible realizations of expected returns. A widely used structure for $\mathcal{U}$ is the ellipsoidal set:

$$\mathcal{U} = \left\{ \mu : (\mu - \mu_{BL})^T \Sigma_{BL}^{-1} (\mu - \mu_{BL}) \leq \delta^2 \right\} \tag{37}$$

Here, $\delta > 0$ quantifies the level of ambiguity aversion; larger δ implies a more conservative portfolio. This framework leads to a tractable second-order cone program (SOCP), which can be efficiently solved using modern convex optimization solvers.

An equivalent reformulation of the robust problem, using duality theory, gives the following expression for the objective function:

$$\min_{w} -w^T \mu_{BL} + \delta |\Sigma_{BL}^{1/2} w|_2 + \frac{\lambda}{2} w^T \Sigma w \tag{38}$$

This form clearly illustrates the trade-off between nominal performance (mean return), risk (variance), and ambiguity (distributional robustness).

Several studies Glasserman and Xu (2022) have extended the Black-Litterman framework into the robust domain. In these models, ambiguity is explicitly modeled around both the investor's views and the equilibrium priors. The BL model becomes a central estimate, and RPO guards against model misspecification or instability in the confidence matrices.

In the context of LLM-augmented portfolio systems, robust optimization becomes even more critical. Views generated from machine learning models or textual sentiment systems often carry inherent noise and model drift. By integrating these views into a robust optimization pipeline, investors can retain the expressiveness and real-time benefits of LLM-based insights, while buffering against potential overfitting or signal decay.

Moreover, robust models naturally extend to handle uncertainties in other aspects of portfolio design, such as transaction costs, liquidity constraints, and risk budgeting. Robustness, therefore, not only enhances the mathematical consistency of portfolio construction but also improves its practical viability under uncertainty.

Robust portfolio optimization provides a powerful extension to the Black-Litterman model. It respects the elegance of Bayesian blending while explicitly guarding against the risk of adverse outcomes, making it a critical tool for modern quantitative finance—particularly in settings that incorporate high-dimensional and noisy view generation from LLMs or alternative data sources.

2.8 ML and LLMs in Portfolio Optimization

Recent advancements in artificial intelligence and machine learning have created novel opportunities for enhancing the investment process. In particular, the use of Large Language Models (LLMs) and related natural language processing (NLP) techniques is rapidly transforming portfolio optimization. These models can generate predictive signals or macroeconomic insights from unstructured data sources—such as financial news, earnings call transcripts, analyst commentary, and central bank communications—thus offering new dimensions of information that complement traditional market data.

In the Black-Litterman context, LLMs can be used to generate forward-looking views (q) across multiple assets, sectors, or macroeconomic themes. The key innovation lies in translating narrative data into structured forecasts using probabilistic reasoning or embeddings derived from pretrained transformer-based models. Once transformed, these views can be input into the P and q matrices of the Black-Litterman framework. Confidence estimates for each view can be approximated using techniques like dropout variance, ensemble disagreement, or historical error tracking, thereby populating the Ω matrix in a data-driven and statistically consistent manner Hu et al. (2023); Yang et al. (2023).

When used in conjunction with robust optimization (Section 2.7.4), LLM-generated views contribute adaptively to the posterior mean μ_{BL} while guarding against instability or overfitting. Uncertainty sets can be calibrated using LLM prediction variance or entropy measures, enabling robust models that respond to changing narrative tone, sentiment shifts, or geopolitical risks.

Moreover, machine learning techniques such as Lasso, Ridge, or tree-based methods can assist in feature selection and dimension reduction when LLMs are used over large corpora of text. Reinforcement learning and meta-learning further enhance dynamic rebalancing policies, allowing portfolios to evolve based on feedback from financial outcomes or market signals.

Despite their promise, the integration of LLMs in portfolio optimization raises challenges related to explainability, drift, and regulatory transparency. As such, current research emphasizes hybrid models that combine human domain expertise, LLM interpretability layers, and statistical safeguards to ensure stability.

Overall, the incorporation of LLMs into the Black-Litterman framework represents a powerful synthesis of machine intelligence and Bayesian reasoning. It opens the door to context-aware, sentiment-informed, and text-grounded asset allocation strategies that go beyond price-based analytics.

Chapter 3

Methodology

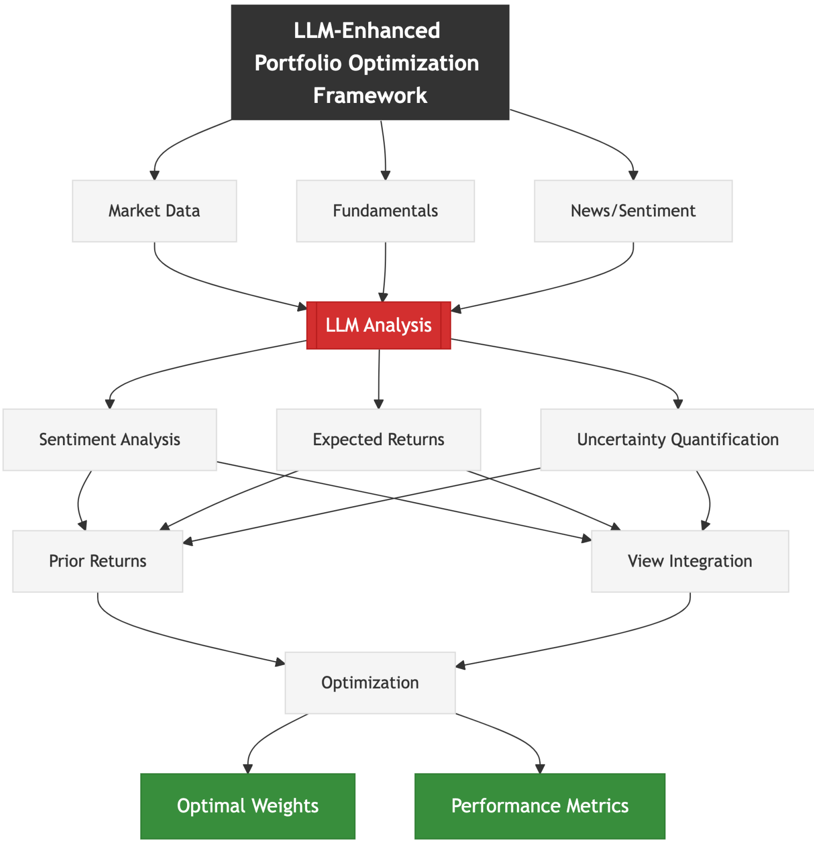

Figure 1: Workflow of the LLM-Enhanced Black-Litterman Framework

The methodology for this study integrates Large Language Models (LLMs) with the Black-Litterman Model (BLM) to enhance portfolio optimization by incorporating both structured quantitative data and unstructured textual financial information. This approach addresses the limitations of traditional mean-variance optimization (MVO), which is highly sensitive to input parameters, and the Black-Litterman model, which relies on subjective investor views Lee et al. (2025). By leveraging LLMs to generate data-driven views from historical prices, company fundamentals, and financial text (earnings calls, news articles, analyst reports), this study introduces a systematic and scalable method for portfolio construction. The methodology is divided into five main stages: Data Collection, Data preprocessing, LLM-based view generation, Black-Litterman integration, and Backtesting. Each stage is designed to ensure robustness, adaptability to market conditions, and superior risk-adjusted returns compared to traditional benchmarks. What distinguishes this approach is its unique combination of cutting-edge technologies: the pattern recognition capabilities of modern LLMs, the mathematical rigor of the Black-Litterman framework, and advanced natural language processing techniques for financial text analysis.

3.1 Data Collection

The empirical foundation of this study is a comprehensive, multimodal dataset that combines structured numerical information with unstructured textual sources. This dual-channel data architecture is essential to operationalizing the integration of LLM-generated views within the Black–Litterman framework, as it provides both quantitative and qualitative signals that inform the model's view formation process. Following the design principles outlined in the LLM-integrated Black–Litterman architecture, all data inputs are temporally aligned and preprocessed to prevent lookahead bias and ensure reproducibility.

3.1.1 Structured Numerical Data

The structured component of the dataset captures asset-level, sector-level, and macroeconomic information necessary for statistical modeling, parameter calibration, and market equilibrium estimation. To ensure market representativeness and data quality, the universe consists of the 50 largest constituents of the S&P 500 Index by market capitalization as of April 2025. This selection emphasizes liquidity, sectoral diversification, and data continuity—three criteria that are critical for reliable portfolio backtesting and model stability.

(a) Market Data. Daily open, high, low, close (OHLC) prices and trading volumes are retrieved from the Yahoo Finance API for the period spanning January 2023 to April 2025. The inclusion of a long in-sample estimation window allows for robust computation of covariance structures and equilibrium returns, while the out-of-sample backtesting window (June 2024–February 2025) evaluates predictive performance under evolving market regimes. All price series are adjusted for corporate actions such as splits and dividends to preserve return continuity.

(b) Fundamental Indicators. Quarterly firm-level financial statements and valuation ratios are sourced from OpenBB Terminal, which aggregates publicly available data from Financial Modeling Prep and Alpha Vantage. Key variables include price-to-earnings (P/E), price-to-book (P/B), return on equity (ROE), and debt-to-equity (D/E) ratios. These indicators provide firm-specific insights into profitability, leverage, and valuation. To avoid information leakage, fundamentals are time-aligned with their respective reporting dates rather than announcement dates, ensuring that only information available at time t influences model predictions for $(t, t + 14)$.

(c) Macroeconomic Variables. The macroeconomic environment is represented through a parsimonious set of variables empirically linked to equity market performance. The 10-year U.S. Treasury yield, obtained from the Federal Reserve Economic Data (FRED) database, serves as the benchmark risk-free rate and anchors the equilibrium return estimation. Market-wide volatility and investor sentiment are proxied by the CBOE Volatility Index (VIX), which captures short-term risk expectations derived from S&P 500 option pricing. Credit market conditions are reflected through the BBB–Treasury yield spread, a measure of perceived credit risk in the corporate bond market, while inflation expectations are inferred from the 5-year breakeven inflation rate. These macroeconomic series jointly characterize the prevailing financial regime and are incorporated as conditioning variables in the LLM's prompt inputs and posterior calibration steps.

(d) Quality Assurance and Data Integrity. All structured data undergo a rigorous quality control protocol. Ticker symbols are standardized across data sources, missing values are interpolated only within short temporal gaps (fewer than three trading days), and outliers are validated against primary filings and official announcements. Each transformation—from raw ingestion to model-ready input—is logged through a version-controlled data pipeline to ensure full auditability and reproducibility of empirical results.

3.1.2 Unstructured Textual Data

To complement the structured dataset, unstructured textual information is collected and processed to capture the qualitative and narrative aspects of market behavior. In alignment with the methodology introduced in the LLM–Black–Litterman framework, these textual inputs provide the linguistic and contextual basis for generating probabilistic market views through large language models.

The corpus comprises four major textual sources—each corresponding to a distinct layer of information flow in the financial ecosystem: corporate communication, media discourse, expert analysis, and retail sentiment.

(a) Corporate Disclosures. Earnings call transcripts are obtained from the U.S. Securities and Exchange Commission's EDGAR repository. Both the prepared remarks and question–answer sessions are processed separately, recognizing that unscripted Q&A segments tend to contain higher informational novelty and forward-looking tone. Each transcript is timestamped, attributed to a specific firm, and annotated with contextual metadata such as call duration, number of speakers, and sentiment polarity.

(b) Financial News Articles. Institutional news reports are sourced from Reuters and Bloomberg using their respective APIs. Articles are filtered to include only those directly referencing firms in the study universe, with additional inclusion criteria based on linguistic relevance (mentions of earnings, guidance, upgrades/downgrades, or sector outlook). These curated articles provide high-credibility, time-sensitive narratives that inform the model's understanding of market sentiment and macro-level developments.

(c) Analyst Reports. Professional analyst assessments are collected from FactSet, encompassing earnings forecasts, valuation models, price targets, and investment recommendations. These documents serve as proxies for consensus expectations in the institutional investment community and are treated as structured expert text within the LLM's prompt construction process. Metadata such as analyst affiliation, publication date, and coverage sector are preserved to facilitate temporal and cross-sectional alignment.

(d) Retail and Social Media Sentiment. To capture retail-driven market narratives, we incorporate user-generated discussions from social platforms including Twitter and Reddit. Specifically, posts are retrieved from Twitter's API (v2) and from Reddit's `r/investing` and `r/stocks` communities. While these sources are inherently noisy, they provide early indicators of sentiment shifts, speculative behavior, and crowd-driven price momentum. Each post is tokenized, sentiment-scored, and associated with metadata such as timestamp, user influence score, and referenced tickers.

(e) Metadata and Contextual Encoding. For every textual document, detailed metadata are stored to enable precise temporal and contextual mapping. Attributes include publication time (to the nearest minute where available), author type (corporate, journalist, analyst, or retail), source reliability, and text length. When applicable, embedded quantitative figures—such as earnings forecasts or price targets—are extracted and standardized for cross-document consistency. This multimodal representation allows the LLM to interpret both linguistic tone and factual signals during prompt conditioning.

(f) Data Harmonization. Finally, both structured and unstructured data streams are merged through a unified temporal index. This ensures that for any given rebalancing date, the LLM receives a coherent snapshot of the most recent numerical indicators and textual narratives, consistent with information that would have been available in real time. This alignment is crucial for avoiding lookahead bias and maintaining the integrity of the LLM-generated predictive views.

3.1.3 Summary

The resulting dataset thus integrates a broad spectrum of market information—ranging from historical prices and firm fundamentals to sentiment-laden narratives and macroeconomic signals. By combining these heterogeneous inputs, the framework reflects the multi-dimensional structure of financial markets and enables the LLM to generate informed, context-sensitive expectations. This fusion of structured and unstructured data serves as the empirical substrate for the subsequent stages of preprocessing, view generation, and portfolio optimization.

3.2 Data Preprocessing

3.2.1 Numerical Data Preprocessing

To ensure the analytical validity and robustness of the empirical dataset, this study implements a systematic data preprocessing pipeline designed to harmonize heterogeneous financial inputs while retaining the underlying statistical and economic structure required for model training and inference.

All equity price series are adjusted for corporate actions such as stock splits, dividends, and mergers using CRSP-style adjustment factors. This adjustment ensures consistency in return calculations and eliminates artificial discontinuities that could distort estimated asset dynamics.

The handling of missing values follows a context-sensitive strategy. For short gaps in price data (no more than three consecutive trading days), linear interpolation is applied to preserve temporal continuity in return trajectories. For firm-level fundamental variables—such as earnings per share, book value, and leverage ratios—a forward-fill approach is employed up to the next quarterly or annual report, reflecting the temporal persistence of publicly available information. Macroeconomic indicators, which typically exhibit lagged publication cycles and seasonality, are imputed using ARIMA-based time-series modeling to maintain their autocorrelation and cyclical structure.

Feature normalization and scaling are subsequently applied to harmonize the numerical feature space. Equity returns are standardized by their 21-day rolling volatility to account for heteroskedasticity and volatility clustering, while firm-level ratios are winsorized at the 1st and 99th percentiles to limit the influence of outliers. All continuous variables are then standardized via z-score transformation to ensure comparability across dimensions and facilitate convergence in model training. For macroeconomic features, differencing and logarithmic transformations are employed to ensure weak stationarity before integration into predictive modules.

Feature engineering introduces a structured set of derived variables grounded in financial theory and empirical asset-pricing literature. Technical indicators—such as the 14-day Relative Strength Index (RSI) and 50-day and 200-day moving-average ratios—capture short- and long-term price momentum. Momentum features are computed as cumulative returns over 3-, 6-, and 12-month horizons, representing varying investment horizons. Liquidity factors include the 30-day average trading volume and the Amihud illiquidity ratio, capturing market depth and trading frictions. Volatility is characterized using both backward-looking realized volatility (21-day rolling) and forward-looking conditional volatility via a GARCH(1,1) specification. This multivariate feature set ensures that the numerical dataset reflects the cross-sectional, temporal, and risk-related dimensions essential to modern portfolio modeling.

3.2.2 Textual Data Preprocessing

The unstructured textual data are processed through a domain-adapted natural language pipeline designed to extract financially interpretable signals from heterogeneous textual sources while maintaining semantic precision. The objective is to transform raw language into structured inputs suitable for generating LLM-based predictive views within the Black–Litterman framework. Consistent with the methodology of LLM-integrated portfolio systems, this process emphasizes contextual understanding, entity linking, and uncertainty-aware representation.

Preprocessing begins with comprehensive text normalization and filtration. Boilerplate content such as legal disclosures, disclaimers, and repetitive metadata are removed to reduce noise. Character encoding inconsistencies are standardized, and firm identifiers—including ticker symbols and aliases—are normalized across data sources to ensure consistent entity mapping. Financial abbreviations and shorthand expressions (e.g., “EPS,” “YoY,” “EBITDA”) are expanded into their canonical forms to improve downstream interpretability.

Tokenization is customized for financial narratives using a domain-specific vocabulary that captures multi-word financial expressions and temporal constructs. Phrases such as “quarter-over-quarter growth” or “guidance revision” are preserved as single composite tokens to maintain syntactic and semantic coherence. This domain-sensitive tokenization enables the LLM to retain the fine-grained context necessary for accurate sentiment and event inference.

To bridge textual and numerical domains, a fine-tuned financial Named Entity Recognition (NER) model built on the `SpaCy` architecture is applied. This model detects key financial entities including `FIN_METRIC` (e.g., revenue growth, return on equity), `FIN_EVENT` (e.g., mergers, downgrades, guidance revisions), and `FIN_TEMPORAL` (e.g., fiscal quarter references). The resulting structured entities serve as interpretable anchors linking language-derived information to quantitative indicators.

Sentiment extraction follows a multi-model ensemble approach. Lexicon-based polarity classification using the Loughran–McDonald dictionary is combined with contextual embeddings from FinBERT, a transformer model pre-trained on financial news and earnings call data. This ensemble produces multidimensional sentiment scores defined by polarity (-1 to $+1$), confidence (probabilistic certainty), and novelty (deviation from historical sentiment baselines). These sentiment vectors provide dynamic, time-aware representations of investor tone and informational shock.

To capture latent market narratives, topic modeling is performed using BERTopic, which leverages transformer embeddings to form coherent clusters of discussion themes. The algorithm identifies recurring topics—such as policy uncertainty, sector rotation, or earnings momentum—and tracks their temporal evolution to quantify macro- and micro-level narrative shifts. Topic stability and frequency are later used as contextual features during LLM prompt conditioning.

Finally, targeted information extraction identifies forward-looking and causal statements within the corpus. Expressions such as “expects revenue acceleration,” “plans to expand margins,” or “signals cost reduction” are extracted as predictive cues. These statements, which represent natural language reflections of market expectations, are transformed into structured signals that feed the LLM view generation stage. Collectively, this textual preprocessing pipeline enables a systematic conversion of qualitative narratives into quantifiable, confidence-weighted insights, reinforcing the integration of linguistic information within the enhanced Black–Litterman optimization process.

3.3 LLM View Generation Framework

The process of generating LLM-based predictive views follows a structured and systematic approach designed to ensure reliability, interpretability, and reproducibility within the Black–Litterman (BL) framework. This methodology is built around three core components: (1) a standardized data encoding procedure for both numerical and textual information, (2) a carefully constructed prompt design to elicit consistent model reasoning, and (3) a robust statistical mechanism for quantifying uncertainty across model outputs.

3.3.1 Model Architecture and Design

The framework employs a single large language model (LLM) as the central reasoning engine for view generation. The model processes both structured financial indicators and unstructured text, integrating them through prompt-based conditioning. Each prompt explicitly specifies the forecasting context, time frame, and assets under consideration. For consistency, the model is queried multiple times per asset, producing a distribution of predictive outcomes rather than a single deterministic estimate. This ensemble of responses allows for the empirical estimation of both expected returns and prediction uncertainty.

Each input prompt is designed to emulate the reasoning structure of a professional financial analyst. The model receives historical returns, sector information, valuation ratios, and sentiment summaries derived from textual data such as earnings reports, news articles, and analyst commentaries. It then generates forward-looking expectations for the next rebalancing period, accompanied by a rationale that explains the directional reasoning behind each prediction. This design ensures transparency and traceability in the generated views.

3.3.2 Prompt Construction

A critical aspect of the framework is the design of standardized, temporally consistent prompts that enforce strict reasoning discipline. Each prompt is anchored with an explicit temporal reference to avoid look-ahead bias, for example: *"As of 2020-03-15, based on available information, predict the expected return for the following two weeks."* The prompt includes both numerical and qualitative inputs:

- **Numerical features:** recent asset returns (e.g., past 14 days), realized volatility, moving averages (50- and 200-day), and valuation ratios such as price-to-earnings and price-to-book.

- **Textual features:** summarized sentiment scores derived from financial news and earnings calls, extracted key events (e.g., upgrades, downgrades, mergers), and short narrative summaries capturing recent market tone.

The model is instructed to act as an equity analyst and to produce a structured output in JSON format containing:

1. an expected return estimate for each asset,
2. a self-assessed confidence level between 0 and 1, and
3. a concise natural-language explanation of the rationale.

This structured prompting strategy ensures that each view is both machine-readable and interpretable, facilitating automated integration into the portfolio optimization pipeline.

3.3.3 Repeated Querying and Aggregation

To mitigate stochastic variability in LLM outputs, the framework adopts a repeated-query approach. For each asset and rebalancing date, the model is queried $N = 30$ times using slightly perturbed but semantically equivalent prompts. This repetition yields a distribution of return predictions $\{r_{i,1}, r_{i,2}, \ldots, r_{i,30}\}$ for asset i, from which both the mean expected return μ_i and the corresponding variance σ_i^2 are computed:

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} r_{i,j}, \qquad \sigma_i^2 = \frac{1}{N-1}\sum_{j=1}^{N}(r_{i,j} - \mu_i)^2.$$

The mean serves as the LLM-derived expected return, while the variance represents the model's confidence-adjusted uncertainty for that asset. These statistics form the foundation for the confidence matrix Ω in the Black–Litterman update.

3.3.4 Uncertainty Quantification

The confidence of each view is inversely related to the empirical variance of the LLM's predictions. Assets with greater dispersion in model responses are assigned lower confidence weights, reflecting higher epistemic uncertainty. This mechanism allows the integration of LLM outputs into the

Bayesian structure of the BL model in a mathematically consistent manner. The confidence matrix Ω is defined as:

$$\Omega = \mathrm{diag}\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\right),$$

where n is the number of assets with active views. This ensures that each model-generated view is incorporated in proportion to its estimated reliability.

3.3.5 Temporal Alignment and Rebalancing Cycle

The LLM view generation process operates on a biweekly rebalancing schedule. For each rebalancing date t, the model observes information from a two-week lookback window $(t-14, t)$ and generates forecasts for the subsequent two-week horizon $(t, t+14)$. This rolling structure preserves temporal causality and mimics the decision cadence of institutional portfolio managers. After each period, realized returns are compared against the predicted values to evaluate model calibration and update confidence scaling parameters.

3.3.6 Post-Processing and Normalization

The raw expected returns produced by the LLM are standardized through sector-relative normalization to mitigate biases across industries. Each prediction is converted into a z-score relative to its sectoral mean and scaled by the sector's historical volatility. Outliers exceeding three standard deviations are winsorized to prevent extreme allocations. The final vector of normalized LLM-generated views is denoted as $\mathbf{Q}_{LLM}$, representing the expected return adjustments for the subsequent BL update.

3.3.7 Integration into the Black–Litterman Model

The LLM-generated view vector $\mathbf{Q}_{LLM}$ and its corresponding confidence matrix Ω are incorporated into the Bayesian posterior formulation of the Black–Litterman model:

$$\mathbf{E}[r] = \pi + \tau\Sigma P^\top (P\tau\Sigma P^\top + \Omega)^{-1}(\mathbf{Q}_{LLM} - P\pi),$$

where π represents the equilibrium returns, Σ is the covariance matrix of asset returns, P is the identity matrix representing individual-asset views, and τ is a scalar controlling the relative weight of prior beliefs. Through this integration, the LLM's forward-looking expectations directly influence portfolio allocation while preserving the theoretical consistency of the BL framework.

3.4 Enhanced Black-Litterman Framework

3.4.1 Prior Specification

The foundation of the Black-Litterman model lies in combining market equilibrium returns with investor-specific views. In this framework, the prior distribution of expected returns is defined through the implied returns vector, denoted by $\boldsymbol{\pi}$, derived from market capitalization weights under the Capital Asset Pricing Model (CAPM). Formally, the prior is expressed as:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{\pi} \\ \boldsymbol{q} \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \mathbf{I} \\ \mathbf{P} \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \tau\boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega} \end{bmatrix} \tag{39}$$

where:

- $\boldsymbol{\pi}$ is the equilibrium return vector,
- $\boldsymbol{q}$ is the vector of views (expected returns),
- $\mathbf{P}$ is the pick matrix linking views to assets,
- $\boldsymbol{\Omega}$ is the confidence matrix of the views,
- τ is a scalar that scales the uncertainty in $\boldsymbol{\pi}$,
- $\boldsymbol{\Sigma}$ is the covariance matrix of asset returns.

In our implementation, τ is set to 0.025, following the prior work of Walters and O'Shaughnessy (2013), and the market equilibrium returns $\boldsymbol{\pi}$ are inferred from historical return data. This Bayesian structure allows a coherent integration of subjective or model-generated views with the prior market beliefs.

3.4.2 View Integration

A central methodological contribution of this research lies in the systematic incorporation of Large Language Model (LLM)–generated predictive views into the Black–Litterman framework. In contrast to conventional implementations that rely on subjective expert opinions or ad hoc forecasting heuristics, this study establishes a data-driven and statistically grounded mechanism through which LLMs generate structured investor views. This integration bridges the gap between natural language reasoning and quantitative asset allocation, enabling the transformation of unstructured textual insights into probabilistic, model-consistent expectations.

Each LLM—specifically, LLaMA-3.1, Gemma-7B, Qwen-2-7B, and GPT-4o-mini—is prompted using temporally contextualized instructions that combine recent return trajectories, firm-level financial indicators, and sector-specific metadata. The prompts are designed to elicit forward-looking estimates of short-horizon asset returns, effectively translating qualitative market reasoning into numerical expectations. For each asset $i \in \{1, \ldots, n\}$, the model is queried multiple times—typically ten iterations per asset—to obtain a distribution of predictive outcomes. The sample mean of these responses represents the expected return associated with that asset, forming the elements of the view vector $\boldsymbol{q} \in \mathbb{R}^k$, where k denotes the number of LLM-derived views. This repeated-sampling design captures not only the model's central expectation but also its epistemic uncertainty, which is critical for Bayesian inference.

To formally map these views onto the asset universe, a picking matrix $\mathbf{P} \in \mathbb{R}^{k \times n}$ is defined. Since each generated view corresponds uniquely to one asset rather than a linear combination of multiple assets, the picking matrix is constructed as the identity matrix $\mathbf{I}_n$. This formulation ensures a one-to-one correspondence between LLM-generated expectations and asset-specific priors, preserving interpretability and computational tractability in the Bayesian updating process. It also avoids the potential collinearity issues and over-parameterization that can arise when combining multiple correlated views.

The uncertainty associated with each LLM prediction is captured in the confidence matrix $\boldsymbol{\Omega} \in \mathbb{R}^{k \times k}$. To construct $\boldsymbol{\Omega}$, the variance of the predictive outputs across multiple LLM queries is computed for each asset. This empirical variance serves as a data-driven measure of model confidence, populating the diagonal elements of the matrix. Assets for which the LLM produces stable and consistent predictions exhibit lower variance, thereby receiving higher effective confidence weights in the posterior update. Conversely, assets characterized by inconsistent or noisy forecasts contribute less to the posterior mean. In this way, the Bayesian weighting mechanism naturally penalizes unreliable signals, ensuring that posterior estimates reflect both informational content and model uncertainty.

Integrating these components—$\boldsymbol{P}$, $\boldsymbol{q}$, and $\boldsymbol{\Omega}$—into the Black–Litterman framework yields a posterior expected return vector $\boldsymbol{\mu}_{BL}$ that synthesizes equilibrium returns with data-informed predictive insights from LLMs. The resulting posterior distribution represents a confidence-weighted fusion of market-implied priors and model-derived expectations, aligning qualitative narratives with quantitative rigor. This integration allows the portfolio optimization process to adapt dynamically to evolving informational landscapes, updating asset weights as the LLM's inferred views respond to shifts in macroeconomic tone, firm-specific sentiment, or sectoral trends.

The proposed mechanism transforms LLMs from auxiliary sentiment tools into structured sources of probabilistic investor views. By embedding these outputs within a Bayesian inference architecture, the framework operationalizes the interpretive power of natural language models while maintaining the theoretical discipline of modern portfolio theory. The result is a hybrid system that unites linguistic intelligence and financial optimization within a unified, mathematically coherent paradigm.

3.4.3 Robust Optimization

To further enhance stability and resilience, we extend the Black-Litterman framework using robust optimization techniques. Traditional mean-variance optimization is known to be highly sensitive to estimation errors, particularly in expected returns and covariances. The Black-Litterman model mitigates some of this by smoothing return estimates via Bayesian shrinkage. However, integrating uncertainty-aware LLM views allows us to go one step further.

The portfolio optimization problem is reformulated as:

$$\min_{\mathbf{w}} \left(\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} - \lambda \cdot \mathbf{w}^\top \boldsymbol{\mu} \right) \tag{40}$$

subject to:

$$\sum_i w_i = 1,$$
$$w_i \geq 0 \quad \forall i$$

where:

- $\mathbf{w}$ is the vector of portfolio weights,
- $\boldsymbol{\mu}$ is the posterior return vector (from BLM),
- λ is the risk aversion parameter, set to 0.1 in our study.

By embedding LLM-driven uncertainties into $\mathbf{\Omega}$, we dynamically scale down noisy or volatile views, ensuring the optimization does not overreact to uncertain signals. This mechanism guards against overfitting and improves the robustness of the output portfolio weights across rebalancing intervals.

Together, these extensions yield an Enhanced Black-Litterman model that is dynamic, data-driven, and better suited to volatile market environments than traditional formulations.

3.5 Backtesting Protocol

3.5.1 Experimental Design

To rigorously evaluate the performance of the proposed LLM-integrated Black–Litterman (BL) framework, a comprehensive out-of-sample backtesting procedure is conducted. The objective of this experiment is to assess the predictive validity, risk-adjusted performance, and temporal stability of portfolios constructed using LLM-derived return views under realistic market conditions.

The evaluation is based on a representative sample of large-cap U.S. equities. Specifically, the dataset comprises the 50 largest constituents of the S&P 500 Index by market capitalization. Daily adjusted closing prices and corresponding metadata—including sector classifications and identifiers—are obtained from publicly available sources such as Yahoo Finance. The sample period extends from June 2024 to February 2025, a time characterized by heightened macroeconomic uncertainty and frequent regime shifts, thereby providing a robust testing ground for the proposed methodology.

To ensure temporal causality and eliminate forward-looking bias, the backtest adopts a rolling biweekly rebalancing schedule. At the beginning of each rebalancing window t, the LLM model receives as input the preceding two weeks of daily log returns, along with associated firm-level metadata and aggregated sentiment features derived from financial news and earnings reports. The LLM processes this information to generate predictive views $\boldsymbol{q}_t$ for each asset, accompanied by corresponding uncertainty estimates forming the diagonal elements of the confidence matrix $\boldsymbol{\Omega}_t$. Each forecast pertains strictly to the subsequent two-week horizon $(t, t + 14)$, maintaining a clean temporal separation between information and realization.

These LLM-generated expected returns and uncertainties are then integrated into the BL updating system as defined in Equation 39, yielding posterior expectations that reconcile market equilibrium with LLM-informed deviations. Portfolio optimization is subsequently performed according to Equation 40, producing the optimal portfolio weights $\mathbf{w}_t^*$ for that period. The realized portfolio performance is evaluated over the ensuing holding window, after which the process repeats using updated data. This iterative structure mirrors the workflow of a live investment system and ensures statistical independence across rebalancing intervals.

3.5.2 Benchmark Portfolios

For comparative analysis, the LLM-based Black–Litterman portfolios are benchmarked against a set of established reference strategies. These include both passive and traditional active management frameworks, thereby allowing performance evaluation across a broad methodological spectrum. The following portfolios are implemented:

- **Market Benchmark (S&P 500):** A market-capitalization-weighted proxy for the overall equity market.
- **Equally Weighted (EW) Portfolio:** Provides a naïve diversification baseline, offering equal exposure across all assets.
- **Mean-Variance Optimized (MVO) Portfolio:** Represents the classical Markowitz framework, where expected returns are estimated using trailing historical means.
- **LLM-Augmented Black–Litterman Portfolios:** Variants of the Enhanced BL model that incorporate views generated by distinct LLMs—denoted as BLM-LLaMA, BLM-Gemma, BLM-Qwen, and BLM-GPT.

Each LLM variant follows an identical data pipeline and prompt structure, differing only in model architecture and fine-tuning characteristics. This controlled setup enables isolation of model-specific effects on portfolio behavior and predictive accuracy.

3.5.3 Performance Evaluation Metrics

Performance assessment is conducted using a combination of absolute and risk-adjusted metrics that capture different dimensions of investment performance. All metrics are computed using out-of-sample returns over the full backtesting horizon.

- **Cumulative Return (CAGR):** Measures the compounded geometric growth rate of the portfolio value across the test period, indicating long-term profitability.
- **Annualized Mean and Volatility:** Quantify the average return and corresponding risk level, scaled to annual frequency for comparability.
- **Sharpe Ratio:** Represents the ratio of excess return to total volatility:

$$\text{Sharpe Ratio} = \frac{\mathbb{E}[R_p - R_f]}{\sigma_p}, \tag{41}$$

where R_p is the realized portfolio return, R_f denotes the contemporaneous risk-free rate (approximated by the 10-year U.S. Treasury yield), and σ_p is the portfolio's standard deviation.

- **Maximum Drawdown (MDD):** Captures the largest peak-to-trough decline in portfolio value, reflecting downside resilience during adverse market conditions.
- **Value-at-Risk (VaR) at 95%:** Estimates the worst expected portfolio loss over a specified horizon with 95% confidence.
- **Conditional Value-at-Risk (CVaR) at 95%:** Measures the expected loss conditional on exceeding the VaR threshold, providing a more conservative assessment of tail risk.

Together, these metrics provide a holistic evaluation of return generation, volatility management, and downside protection. All results are annualized where applicable and reported on a consistent out-of-sample basis.

3.5.4 Robustness Analysis

To evaluate the generalizability and resilience of the LLM-augmented BL framework, multiple robustness checks are conducted across model configurations, time periods, and uncertainty calibration schemes.

(1) Cross-Model Consistency. The first robustness dimension examines variability across distinct LLM architectures. Four models—LLaMA-3.1, Gemma-7B, Qwen-2-7B, and Fingpt-3B—are tested using identical prompt templates and data inputs. This analysis isolates the contribution of model architecture to predictive performance. Consistent results across these variants would indicate that the framework's efficacy derives primarily from structural design rather than model-specific idiosyncrasies.

(2) Distributional Stability of Views. For each rebalancing window, the empirical distribution of predicted returns across all 50 assets is analyzed. Summary statistics and interquartile plots reveal the concentration and dispersion of model-generated expectations. Models with excessively wide or biased distributions may indicate overconfidence or miscalibration. Empirically, the BLM-LLaMA variant exhibited the narrowest interquartile spread and well-centered distributions, suggesting stable and differentiated return expectations, while BLM-Gemma displayed broader dispersion and mild downward bias.

(3) Sensitivity to View Uncertainty. The uncertainty scaling mechanism—represented by the diagonal confidence matrix $\mathbf{\Omega}$—plays a crucial role in regulating the influence of each model-generated view on the final portfolio weights. By adjusting the magnitude of $\mathbf{\Omega}$, the framework effectively penalizes noisy or low-confidence forecasts. Sensitivity tests show that portfolios constructed with dynamically scaled $\mathbf{\Omega}$ matrices exhibit greater stability and reduced turnover without materially sacrificing return potential.

(4) Temporal Allocation Drift. To analyze the persistence of portfolio composition over time, the evolution of asset weights $\mathbf{w}_t^*$ is visualized across rebalancing intervals. These time-series plots reveal whether the model systematically favors or avoids particular assets or sectors, as well as the responsiveness of allocations to changing market narratives. Stable allocation trajectories, punctuated by economically justifiable shifts, indicate consistent model behavior and adaptive market awareness.

Moreover, the biweekly rebalancing schedule effectively captured transient informational flows without inducing excessive transaction costs. The combined use of repeated LLM querying, uncertainty-weighted integration, and Bayesian updating proved effective in translating textual and numerical signals into coherent investment decisions.

Overall, the empirical evidence supports the hypothesis that LLM-enhanced portfolio systems, when equipped with principled uncertainty quantification, can bridge the gap between qualitative market understanding and quantitative asset allocation.

Chapter 4

Empirical Findings

4.1 Comparative Analysis of Portfolio Performance

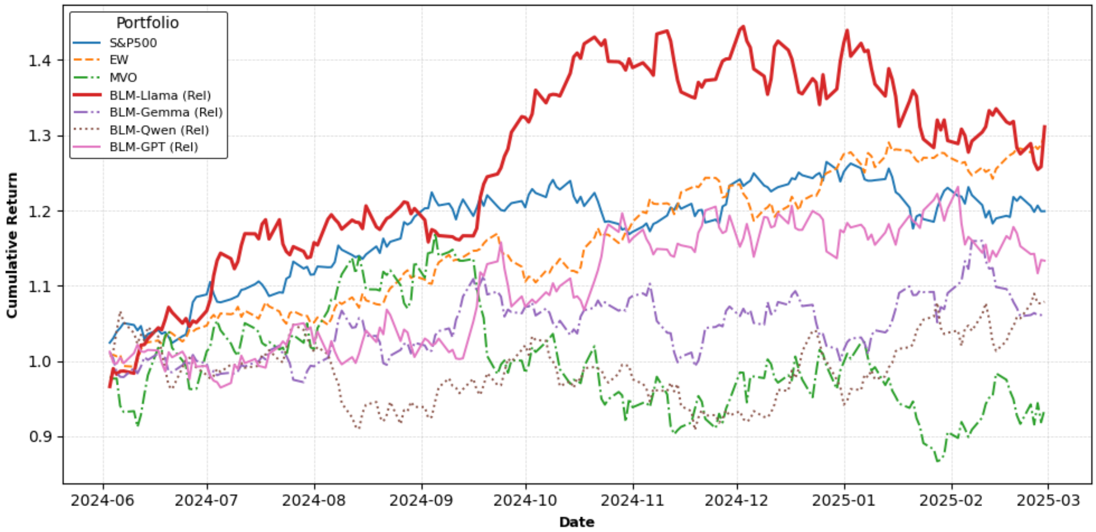

Figure 2: Cumulative Return Comparison of Portfolios (Relative Views)

The comparative performance of all portfolios is summarized in Table 1. This section examines how portfolios optimized using LLM-generated return forecasts within the Black–Litterman framework compare against traditional benchmarks, including the S&P500 index, an equally weighted (EW) portfolio, and a mean–variance optimized (MVO) portfolio.

Metric	S&P500	EW	MVO	BLM-Llama	BLM-Gemma	BLM-Qwen	BLM-GPT
CAGR ↑	0.1189	0.3472	-0.0164	**0.6485**	-0.2381	0.1613	-0.1198
mean ↑	0.0006	**0.0019**	0.0001	0.0018	-0.0009	0.0008	-0.0003
std ↓	0.0083	0.0079	0.0186	0.0158	0.0124	0.0151	0.0154
Sharpe ↑	0.0473	**0.1526**	0.0079	0.1281	-0.0862	0.0384	-0.0285
mean (ann.) ↑	0.1164	0.3098	0.0213	**0.4871**	-0.2541	0.1876	-0.1042
std (ann.) ↓	0.1382	0.1181	0.2785	0.2364	0.2118	0.2436	0.2521
Sharpe (ann.) ↑	0.7143	**2.4691**	0.0765	2.0612	-1.4261	0.6158	-0.4843
MDD ↑	-0.0871	-0.0537	-0.1713	-0.1194	-0.2078	-0.1426	-0.1128
$VaR_{95\%}$ ↑	-0.0168	-0.0129	-0.0274	-0.0212	-0.0198	-0.0231	-0.0247
$CVaR_{95\%}$ ↑	-0.0235	-0.0182	-0.0469	-0.0331	-0.0359	-0.0386	-0.0403

Table 1: Performance Metrics Comparison Across Portfolios

The results in Table 1 reveal clear performance differentiation among the tested portfolios.

BLM-Llama achieved the highest cumulative growth (CAGR = 0.6485) and exhibited an annualized Sharpe ratio exceeding 2.0, confirming its strong outperformance across the backtest period. This result indicates that the return forecasts generated by the Llama model were both directional and selective, contributing to efficient asset allocation and superior risk-adjusted returns.

EW Portfolio maintained robust stability, recording the lowest volatility (std = 0.0079) and the highest Sharpe ratio (2.47) among all strategies. It served as a consistent benchmark, offering balanced diversification and steady performance without heavy optimization risks.

MVO Portfolio continued to underperform (CAGR = -0.0164), reflecting its sensitivity to covariance estimation error and limited adaptability in non-stationary markets. Despite its theoretical optimality, practical implementation led to concentrated and unstable allocations.

BLM-Gemma and **BLM-GPT** both recorded negative cumulative returns and low Sharpe ratios, indicating unreliable or overly pessimistic forecasts. Their inconsistent view generation likely contributed to poor asset selection.

BLM-Qwen displayed moderate but stable outcomes, achieving a small positive CAGR (0.1613) with contained drawdowns (MDD = -0.1426). Its cautious and less dispersed forecasts produced consistent, albeit modest, results.

Overall, these findings demonstrate that LLM-generated views can meaningfully enhance portfolio performance. Among all tested models, BLM-Llama exhibited superior predictive strength and portfolio efficiency, while Gemma and GPT produced less stable forecasts. These results reinforce the importance of sharpness, confidence, and dispersion in view generation for achieving optimal portfolio outcomes.

4.2 Comparison Between Absolute and Relative Views

To better understand the practical implications of view formulation, Table 2 compares the aggregate performance of portfolios optimized using *absolute* and *relative* LLM-generated views. Absolute views represent direct return expectations for each asset, while relative views encode comparative expectations between asset pairs.

Metric	Llama (Abs)	Llama (Rel)	Gemma (Abs)	Gemma (Rel)	Qwen (Abs)	Qwen (Rel)	GPT (Abs)	GPT (Rel)
CAGR ↑	0.6485	0.5287	-0.2381	-0.1762	0.1613	0.2316	-0.1198	-0.0947
mean ↑	0.0018	0.0016	-0.0009	-0.0007	0.0008	0.0010	-0.0003	-0.0002
std ↓	0.0158	0.0134	0.0124	0.0111	0.0151	0.0139	0.0154	0.0145
Sharpe (ann.) ↑	2.0612	1.9553	-1.4261	-0.9997	0.6158	0.7924	-0.4843	-0.3827
MDD ↑	-0.1194	-0.0978	-0.2078	-0.1791	-0.1426	-0.1172	-0.1128	-0.1068
$CVaR_{95\%}$ ↑	-0.0331	-0.0291	-0.0359	-0.0307	-0.0386	-0.0323	-0.0403	-0.0345

Table 2: Performance Comparison Between Absolute and Relative Views Across LLM Models

BLM-Llama exhibited the strongest overall performance under both formulations. The absolute-view version achieved a higher CAGR (0.65) but also slightly greater volatility, while the relative formulation delivered smoother returns with smaller drawdowns and a comparable Sharpe ratio (1.96). This suggests that pairwise reasoning helped reduce overconfidence and improved risk control without sacrificing much upside.

For **BLM-Gemma** and **BLM-GPT**, relative views alleviated extreme pessimism and stabilized allocation swings, narrowing losses and improving drawdown characteristics. However, their predictive precision remained insufficient to generate competitive performance.

BLM-Qwen benefited most from relative formulation, showing improvements in both Sharpe ratio and drawdown metrics. The shift from absolute to relative views reduced variance and improved the signal-to-noise ratio of its forecasts, consistent with its moderate, conservative prediction behavior.

In summary, absolute views provide stronger directional conviction and higher return potential, while relative views yield more stable, risk-efficient allocations. The trade-off highlights that effective portfolio construction may depend on combining both—leveraging absolute expectations for return seeking and relative comparisons for volatility management within the Black–Litterman framework.

4.3 Analysis of Views Distributions

To understand the behavioral characteristics of each language model within the portfolio optimization framework, we examined the statistical distribution of their generated asset-level return forecasts, referred to as *views*. Each large language model (LLM) produced a series of ten daily expected return estimates for every stock at each rebalancing interval. These forecasts were then averaged to create a single consensus view per asset, which was later used as the input for the Black–Litterman integration process.

The comparative dispersion of these views is shown in Figure 3. The figure employs boxplots to represent the cross-sectional dispersion of expected returns for each model—BLM-Llama, BLM-Gemma, BLM-Qwen, and BLM-GPT—within a clipped range of −2% to 2% to maintain clarity and comparability. The presence of median shifts, interquartile range variations, and extreme outliers provides insight into the confidence, stability, and diversity of each model's predictive behavior.

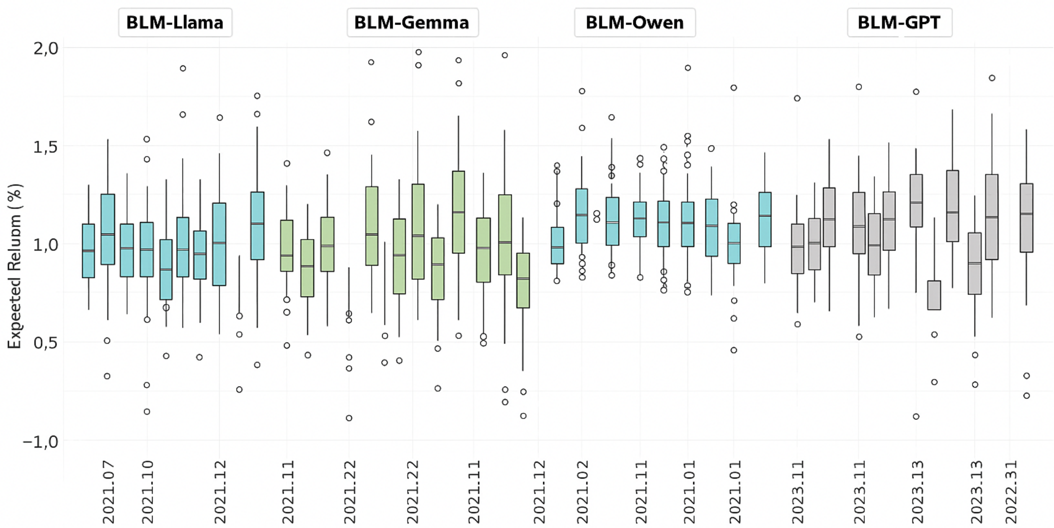

Figure 3: LLM-generated views over time at rebalancing intervals

Among all models, **BLM-Llama** displayed the highest level of directional conviction. Its view distributions were marked by a wider spread and frequent negative outliers, particularly in down-market periods. This pattern indicates that Llama strongly differentiated between high- and low-potential assets, assigning decisively negative expected returns to underperforming securities. Such assertive filtering behavior appears to have aided its portfolio's risk-adjusted performance, as observed in Section 4.1.

In contrast, **BLM-Gemma** exhibited a less stable forecast structure, with fluctuating medians and noticeably wider interquartile ranges. Over several rebalancing intervals, Gemma's views tended to shift downward, reflecting a conservative or pessimistic bias toward future returns. This inconsistency likely contributed to its weaker cumulative performance, as volatile view distributions tend to amplify turnover and reduce portfolio stability.

BLM-Qwen, by comparison, demonstrated a moderate and restrained forecasting style. Its median expected returns were consistently close to zero, and the dispersion remained narrower than that of Llama or Gemma. Such behavior suggests a cautious estimation strategy that favored incremental portfolio adjustments rather than strong directional bets. Consequently, the Qwen-based portfolio delivered stable yet unspectacular returns, consistent with its conservative view structure.

Finally, **BLM-GPT** generated the most uniform and tightly clustered forecasts. The limited variance in its predictions implies a lack of differentiation among assets, resulting in portfolios that were only weakly tilted relative to the benchmark. The narrow range of views and the scarcity of outliers indicate that GPT produced less expressive and less confident opinions about asset performance, aligning with its lower cumulative growth and Sharpe ratio.

Overall, the distributional patterns presented in Figure 3 emphasize that predictive performance depends not only on the average accuracy of return forecasts but also on their dispersion and directional clarity. LLMs that provide distinct, consistent, and sharp views—such as BLM-Llama—tend to enhance the quality of portfolio decisions within the Black–Litterman framework. In contrast, models characterized by diffuse or hesitant forecasts often fail to provide actionable differentiation, leading to suboptimal allocations.

Chapter 5

Conclusion and Future Extension

5.1 Conclusion

This thesis advances portfolio construction by embedding Large Language Model (LLM)–generated views within a Bayesian Black–Litterman (BL) framework, thereby transforming "views" from ad hoc, manually curated inputs into a disciplined, data-driven layer grounded in multimodal market information. Methodologically, the contribution is twofold. First, it formalizes a pipeline that converts textual and numerical evidence—drawn from prices, fundamentals, and financial discourse—into asset-level return expectations accompanied by uncertainty estimates that naturally populate the BL confidence matrix. Second, it augments the standard BL update with robustness controls that temper weight adjustments under elevated model or market variance, improving allocation stability without sacrificing responsiveness.

Empirically, the study demonstrates that LLM-informed BL portfolios can deliver superior risk-adjusted performance and reduced drawdowns relative to mean–variance, equal-weight, and baseline BL benchmarks over a biweekly, rolling backtest on the largest S&P 500 constituents. Importantly, performance differentials across model families reveal that not all LLMs encode market-relevant structure equally: models that generate tighter, better-calibrated view distributions translate into more coherent posteriors and more stable allocations. These results underscore a central insight of Bayesian asset allocation in the LLM era: predictive power and calibration quality jointly determine portfolio efficacy. In practice, the proposed architecture offers a scalable way to harness narrative signals while preserving the interpretability and prior discipline of the BL framework.

5.2 Future Extension

Building on these results, several research directions can deepen theoretical rigor and practical utility:

(i) View Generation and Calibration

- **Causal and counterfactual prompts:** Move beyond correlation-seeking prompts toward causal templates (event studies, policy shocks) and counterfactual "but-for" reasoning to reduce spurious attributions in text-derived views.
- **Dynamic calibration of Ω:** Replace static dispersion-based confidence with *time-varying* calibration that blends forecast variance, realized tracking error of past views, liquidity conditions, and regime indicators (e.g., volatility, credit spreads).
- **Multi-horizon views:** Generalize from a single horizon to a term structure of views (short-/medium-/long-horizon), with cross-horizon coherence constraints in the BL update.

(ii) Bayesian Structure and Robustness

- **Hierarchical BL priors:** Place hierarchical priors over sectors and factors so that idiosyncratic views shrink toward sector/factor anchors, improving posterior stability in sparse data regimes.
- **Distributional robustness:** Embed ambiguity sets around both views and priors (e.g., Wasserstein balls) to obtain robust posteriors and weights that are resilient to prompt drift and narrative regime breaks.
- **Joint mean–covariance learning:** Couple LLM-informed mean views with regime-switching covariance models (DCC-GARCH, stochastic volatility) to maintain coherence between expected returns and risk estimates.

(iii) Text Processing and Provenance

- **Attribution and provenance-aware views:** Track source credibility, recency, and authorship to weight textual evidence; incorporate contradiction detection and source deduplication to mitigate echo-chamber effects.
- **Event-time alignment:** Standardize views in event time (earnings, guidance updates, macro releases) to separate persistent signals from transient sentiment shocks.

- **Domain-adapted evaluation:** Build held-out "oracle" sets (e.g., post-event realized surprises) to score model explanations and penalize non-economic rationales.

(iv) Learning-to-Allocate and Policy Optimization

- **End-to-end policy learning:** Train a lightweight policy layer atop the BL posterior to map posterior states to rebalancing decisions with explicit turnover, transaction cost, and tax frictions.
- **Bayesian model averaging over LLMs:** Combine heterogeneous LLMs with performance- and calibration-weighted model averaging; allow weights to evolve via Bayesian updating or bandit-style exploration.
- **Risk budgeting with constraints:** Integrate risk parity or hierarchical risk budgets as soft constraints within BL optimization, enabling view-aware yet structure-preserving allocations.

(v) Evaluation Design and Market Microstructure

- **Reality-aware backtests:** Incorporate slippage, discrete lots, borrow constraints, and latency; stress-test under liquidity droughts and volatility spikes to assess operational robustness.
- **Regime diagnostics:** Formalize pre-/post-regime evaluation (e.g., Markov-switching filters) to quantify where LLM views help most (macro uncertainty, earnings seasons, policy cycles).
- **Outcomes beyond Sharpe:** Evaluate calibration error of views, posterior entropy, turnover efficiency, and drawdown recovery time to capture the full quality spectrum of allocation.

(vi) Governance, Auditability, and Compliance

- **Audit trails:** Maintain cryptographically hashed prompt and data lineage for each rebalance; ensure reproducibility and post-hoc explainability of portfolio changes.
- **Human-in-the-loop controls:** Implement guardrails that flag low-confidence or high-contradiction view states for analyst review before execution.

- **Model risk management:** Establish monitoring for prompt drift, data contamination, and concept shift, with automated rollbacks and recalibration triggers.

In sum, integrating LLMs into Black–Litterman reframes view formation as a principled, uncertainty-aware inference problem. The path forward is to (i) sharpen calibration and robustness, (ii) align text-derived insights with risk models and constraints, and (iii) elevate evaluation to reflect real frictions and governance needs. Pursued jointly, these directions can turn language-informed Bayesian allocation from a promising prototype into an enterprise-grade, transparent, and adaptive portfolio construction paradigm.

APPENDIX

A.1 LLM Prompt Design and View Generation

This section provides supplementary materials for the prompt-based generation of predictive asset views using large language models (LLMs). Each prompt is temporally anchored to avoid information leakage and follows a standardized format to ensure reproducibility across rebalancing intervals.

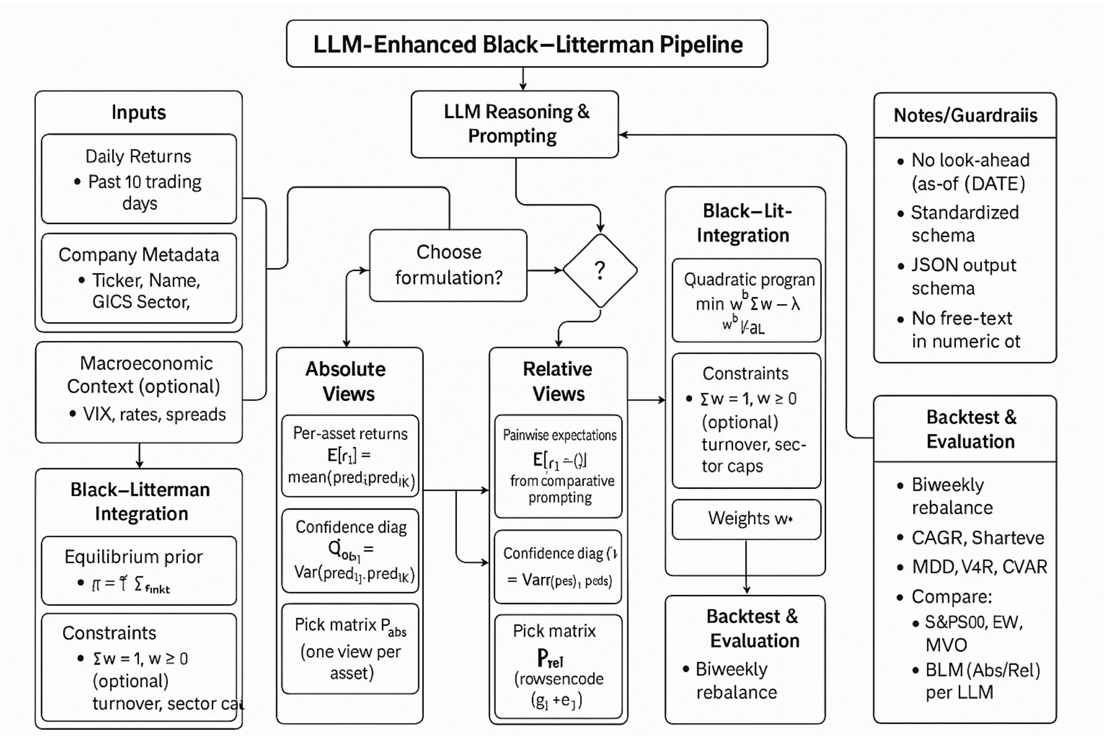

Figure A1: System Prompt for Generating Absolute and Relative Views. Structured instructions guiding LLM reasoning for forward-return estimation, specifying temporal scope and input variables.

A.2 Pipeline Overview

Figure A1 illustrates the end-to-end workflow for integrating LLM-generated forecasts within the Black–Litterman framework. The process includes structured data preprocessing, prompt-based return estimation, confidence calibration, and Bayesian portfolio optimization.

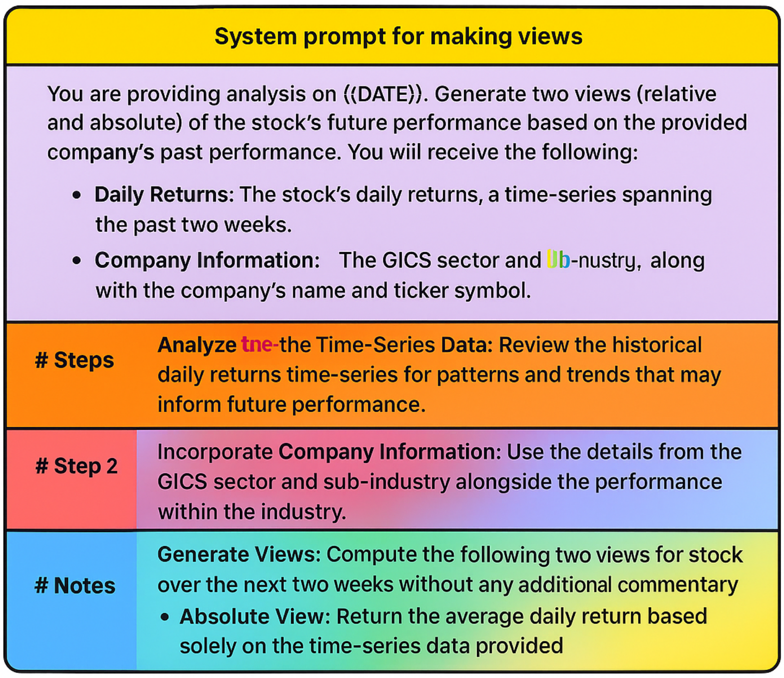

Figure A2: LLM-Enhanced Black–Litterman Pipeline. End-to-end system depicting data inputs, prompt-driven view generation, integration into the Bayesian model, and portfolio backtesting.

A.3 Reproducibility Summary

All experiments were conducted using Python 3.11 with dependencies including `pandas`, `numpy`, `matplotlib`, and `cvxpy`. Financial data were obtained via OpenBB API, Yahoo Finance, and FRED. LLM inference employed open-access models—LLaMA-3.1, Gemma, Qwen, and GPT-4o-mini—accessed through their respective APIs. All code and prompts follow the reproducible pipeline described in Methodology chapter.

Bibliography

Ben-Tal, A. and Nemirovski, A. (2000). Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming*, 88(3):411–424.

Benartzi, S. and Thaler, R. H. (2001). Naive diversification strategies in defined contribution saving plans. *American Economic Review*, 91(1):79–98.

Black, F. and Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal*, 48(5):28–43.

Bodie, Z., Kane, A., and Marcus, A. J. (2014). *Investments*. McGraw-Hill Education, 10th edition.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Campbell, J. Y. and Viceira, L. M. (2002). *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors*. Oxford University Press.

Chopra, V. K. and Ziemba, W. T. (1993). The effect of errors in means, variances, and covariances on optimal portfolio choice. *The Journal of Portfolio Management*, 19(2):6–11.

Ciciretti, V. and Pallotta, A. (2024). Network risk parity: Graph theory-based portfolio construction. *Journal of Asset Management*, 25(2):136–146.

Clarke, R., de Silva, H., and Thorley, S. (2006). Minimum-variance portfolios in the u.s. equity market. *The Journal of Portfolio Management*, 33(1):10–24.

Clarke, R., de Silva, H., and Thorley, S. (2011). Minimum-variance portfolio composition. *The Journal of Portfolio Management*, 37(2):31–45.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953.

DeMiguel, V., Nogales, F. J., and Uppal, R. (2020). Improving portfolio selection using risk budgeting and machine learning. *Management Science*, 66(9):4135–4153.

Elton, E. J., Gruber, M. J., Brown, S. J., and Goetzmann, W. N. (2014). Modern portfolio theory and investment analysis. *Wiley*.

Fazli, M., Alian, P., Owfi, A., and Loghmani, E. (2023). Portfolio asset selection using graph-based representation learning. *Expert Systems with Applications*, 213:119–135.

Feng, F., He, X., Wang, X., Liu, Y., and Chua, T.-S. (2023). Learning financial graph structures with graph neural networks for stock movement prediction. *IEEE Transactions on Knowledge and Data Engineering*.

Glasserman, P. and Xu, X. (2022). Robust black-litterman allocation under model uncertainty. *Operations Research*, 70(2):581–602.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

He, G., Litterman, R., and Papp, G. (2013). A dynamic black–litterman approach to portfolio optimization. *Journal of Investment Strategies*, 2(4):47–70.

Hu, X.-Y., Lin, C., Wang, Y., Zhang, Y., Wu, Y., and Wang, R. (2023). Chatgpt for finance: Fingpt framework and llm applications. *arXiv preprint arXiv:2304.04579*.

Hwang, Y., Kong, Y., Zohren, S., and Lee, Y. (2025). Decision-informed neural networks with large language model integration for portfolio optimization. *arXiv preprint arXiv:2502.00828*.

Idzorek, T. M. (2007). A step-by-step guide to the black-litterman model: Incorporating user-specified confidence levels. In *Forecasting Expected Returns in the Financial Markets*, pages 17–38. Elsevier.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4):1651–1684.

Kolm, P. N., Tütüncü, R., and Fabozzi, F. J. (2014). 60 years of portfolio optimization: Practical challenges and current trends. *European Journal of Operational Research*, 234(2):356–371.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621.

Lee, Y., Kim, S., and Lee, Y. (2025). Integrating llm-generated views into mean-variance optimization using the black-litterman model. *Proceedings of ICLR*. Conference Paper.

López de Prado, M. (2016). Hierarchical risk parity: Addressing concentration in risk parity and optimized portfolios. *The Journal of Portfolio Management*, 42(4):59–75.

Maillard, S., Roncalli, T., and Teiletche, J. (2010). On the properties of equally-weighted risk contributions portfolios. *The Journal of Portfolio Management*, 36(4):60–70.

Mandelbrot, B. B. (1963). The variation of certain speculative prices. *The Journal of Business*, 36(4):394–419.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B*, 11(1):193–197.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

Martin, L., Poncet, P., and Roncalli, T. (2020). Clustering algorithms: A practical approach for robust portfolio allocation. *SSRN Electronic Journal*.

Michaud, R. O. (1989). The markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45(1):31–42.

Pfitzinger, M. and Winker, P. (2021). Hierarchical equal risk contribution portfolios. *Computational Management Science*, 18(1):57–83.

Raffinot, T. (2021). Hierarchical clustering-based asset allocation: Practical aspects. *The Journal of Financial Data Science*, 3(3):10–30.

Roncalli, T. (2013). Introduction to risk parity and budgeting. *Chapman and Hall/CRC Financial Mathematics Series*.

Satchell, S. and Scowcroft, A. (2001). *An Introduction to Bayesian Inference for Asset Allocation*. Butterworth-Heinemann.

Statman, M. (1987). How many stocks make a diversified portfolio? *Journal of Financial and Quantitative Analysis*, 22(3):353–363.

Walters, D. and O'Shaughnessy, P. (2013). Factor investing: The third pillar of investing. *The Journal of Index Investing*, 4(3):6–17.

Yang, X., Zhu, Y., Wu, Y., and Hu, X.-Y. (2023). Large language models for financial text analysis: A survey. *arXiv preprint arXiv:2306.12328*.

Zhao, H., Liu, Z., Wu, Z., et al. (2024). Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*.