Agentic Portfolio Management with a Bayesian Black–Litterman Framework

Victor Xiao[1], Zhaochen Jiang[1], and Changle Li[1]

[1]Columbia Business School

May 3rd 2025

Abstract

We develop a novel, agentic architecture for multi-asset portfolio allocation that embeds large-language-model (LLM) forecasting agents within the Bayesian Black–Litterman framework. Specifically designed to navigate the complexity and heterogeneity of multi-asset portfolio management, our framework employs four specialised forecasting agents—each dedicated to equities, fixed income, foreign exchange, or commodities. These agents systematically ingest structured numerical data feeds, including market prices, macroeconomic indicators, sentiment analyses, and other domain-specific variables, to generate weekly return forecasts accompanied by rigorously calibrated uncertainty measures.

These heterogeneous, agent-generated views are combined within the Black–Litterman model, where they are fused with a market equilibrium prior to produce a coherent posterior distribution of expected returns. A turnover-constrained optimisation routine then translates these posterior expectations into dynamically balanced, investable portfolios. To rigorously evaluate our approach, we perform comprehensive out-of-sample back-tests covering the period from November 2023 to Mar 2025. Results reveal significant advantages: equity and commodity BL portfolios consistently outperform their respective equal-weighted benchmarks by approximately 2–3 percentage points annually, achieving enhanced returns with notably lower maximum drawdowns. Meanwhile, fixed-income and foreign-exchange portfolios successfully achieve their primary objective of volatility reduction, effectively preserving capital across volatile market regimes.

Our methodology innovates on three fronts: (i) modular integration of LLM-based point forecasts and uncertainty quantification; (ii) rigorous Bayesian fusion that balances data-driven signals against market consensus; and (iii) an end-to-end pipeline enforcing realistic turnover limits while preserving analytical tractability. We further enhance flexibility through a meta-level Manager Agent, which adaptively reallocates between asset-class "sleeves" based on evolving signal strength, yielding additional performance gains. [1]

[1]We thank Professor Yin Luo for his kind input and supervision for our thesis, and our Discussant team for their iterative feedback.

Contents

1 Introduction

In recent decades, global asset management has grown significantly more complex, with institutional investors routinely managing diversified portfolios comprising equities, bonds, currencies, and commodities. Each asset class uniquely responds to distinct economic drivers, risk factors, and market dynamics, demanding nuanced forecasting strategies. Yet, traditional asset allocation frameworks predominantly rely on a single, homogeneous forecast fed into a mean–variance optimiser. This approach, while computationally tractable and conceptually simple, leaves portfolios vulnerable during periods when the underlying forecast proves inaccurate or overly simplistic. The inherent mismatch between the complexity of modern financial markets and the simplicity of traditional modelling methodologies forms a primary motivation for our research.

A second key motivation arises from the empirical limitations observed in classical allocation models during regime shifts—periods characterised by abrupt changes in economic relationships and heightened market uncertainty. Under these conditions, traditional equilibrium-based methods such as mean–variance optimisation frequently break down, precisely when their stability and predictive reliability are most needed. This fragility underscores the necessity for a more adaptive framework capable of accommodating multiple forecasts that reflect domain-specific insights, thereby enhancing robustness and resilience across varying market conditions.

To address these issues, we propose an innovative agent-based Black–Litterman (ABBL) asset allocation framework. Our framework systematically integrates a suite of specialised forecasting agents, each powered by advanced large-language models (LLMs) and tailored explicitly to distinct asset classes: equities, fixed income, foreign exchange, and commodities. These agents ingest structured, domain-specific data weekly—such as market prices, macroeconomic indicators, central-bank communications, and sentiment analytics—and produce carefully calibrated return forecasts along with explicit uncertainty quantifications.

Rather than relying on ad-hoc methods to combine these diverse views, we embed them within the disciplined Bayesian framework provided by the Black–Litterman model. Specifically, these heterogeneous agent-generated views are rigorously fused with a market equilibrium prior, yielding posterior expected returns informed by both empirical data and Bayesian principles. Subsequently, these posterior forecasts feed into a turnover-aware quadratic optimiser designed to dynamically balance expected returns against transaction costs and portfolio turnover constraints.

The contributions of this paper are threefold. First, it demonstrates how modern, data-grounded LLM-driven forecasting agents can effectively coexist within a Bayesian allocation framework, enabling nuanced, asset-class-specific insights to inform portfolio decisions. Second, it provides extensive empirical validation, including robustness checks explicitly designed to identify and eliminate potential look-ahead biases, thereby ensuring real-world applicability. Third, the study sets forth a clear pathway for future enhancements, particularly in areas related to agent-confidence calibration, ensemble learning methodologies, and the integration of reinforcement-learning techniques for adaptive hyper-parameter tuning.

The paper proceeds as follows: Section 2 situates our work within the existing literature, highlighting both foundational insights and recent advancements. Section 3 presents a detailed account of data engineering processes and exploratory analyses, while Section 4 outlines our methodological framework, with specific attention to eliminating look-ahead bias. Empirical results and performance assessments are provided in Section 5, followed by robustness tests in Section 7. Section 8 summarises the main findings and practical contributions, while Section 9 discusses key limitations and directions for future research.

2 Literature Review

2.1 Time-Series Forecasting for Major Asset Classes

The last decade has seen substantial progress in applying advanced machine-learning techniques to predict returns across major asset classes, including equities, bonds, foreign exchange (FX), and commodities. In the equity domain, deep learning methods, particularly long short-term memory (LSTM) networks, have shown strong predictive capabilities in forecasting short-term excess returns compared to traditional linear regression and autoregressive models. Studies by Fischer and Krauss (2018) demonstrate that LSTM architectures outperform conventional benchmarks, particularly due to their ability to capture complex, non-linear temporal dynamics.

Beyond recurrent neural networks, other machine-learning methods such as gradient boosting trees, random forests, and multi-layer perceptrons have been increasingly utilized. The comprehensive empirical analysis conducted by Gu, Kelly, and Xiu (2020) illustrates that these non-linear models significantly enhance factor discovery and equity-risk-premium estimations by effectively capturing interactions between macroeconomic variables and firm-specific characteristics, thereby surpassing classical factor models.

In fixed-income markets, advanced forecasting models have similarly evolved. Bianchi, Buchner, and Tamoni (2020) show that tree-based ensemble methods and deep neural network models consistently achieve predictive accuracy superior to conventional macro-factor models, such as those based on principal component analyses of the yield curve. Their comparative analyses highlight the capacity of machine-learning techniques to extract intricate, latent risk-premium signals embedded in bond yields.

In the FX forecasting arena, hybrid models that combine deep-learning techniques with traditional ensemble methods have emerged prominently. Sun, Luo, and Zhang (2020) develop sophisticated hybrid models combining LSTMs with bootstrap aggregation methods, enabling the capture of nonlinear, regime-dependent dynamics frequently missed by linear frameworks. These models consistently deliver higher out-of-sample predictive performance, significantly improving forecasting accuracy and robustness.

Commodity forecasting also benefits from machine-learning advancements, with extensive evalu-

ations highlighting temporal convolutional networks (TCNs) and gradient-boosted decision trees (XGBoost) as top-performing models. Research by Foroutan and Bozic (2024) rigorously compares sixteen different deep learning and boosting methods, demonstrating that TCNs and XGBoost deliver superior predictive accuracy for crude oil and precious metal prices due to their effectiveness in capturing temporal patterns and complex market dynamics.

Collectively, these comprehensive studies affirm that contemporary machine-learning approaches offer substantial accuracy improvements across all principal asset classes, outperforming traditional linear and econometric benchmarks through sophisticated data-driven modeling.

2.2 Black–Litterman and Bayesian Portfolio Allocation

The Black–Litterman (BL) model, initially proposed by Fischer Black and Robert Litterman in 1992, remains a cornerstone framework for Bayesian portfolio allocation due to its intuitive integration of market equilibrium priors and investor-specific views. Further conceptual clarification by He and Litterman in 1999 elucidates the underlying intuition and methodological robustness, establishing BL as a preferred approach to stabilize portfolio weights compared to classical mean–variance optimization, which is highly sensitive to estimation errors and input uncertainty.

Extensive research has subsequently expanded and refined the BL framework. Kolm and Ritter (2017) provide a fully Bayesian reinterpretation, rigorously treating investor views as likelihood information to yield more theoretically coherent posterior distributions of expected returns. In their exhaustive review, Kolm and Tütüncü (2021) discuss various critical enhancements to the BL model, including sophisticated techniques for prior specification, transaction-cost integration, handling estimation risk, and managing leverage constraints, greatly enriching the model's practical applicability.

Recent innovations have incorporated factor investing directly into the BL framework, explicitly embedding factor-based signals such as Fama–French style factors and macroeconomic indicators. Notably, Ko and Lee (2024) effectively integrate factor-augmented priors into BL, achieving substantial improvements in out-of-sample Sharpe ratios relative to traditional BL and static benchmarks. Similarly, research by Barua and Gabaix (2022) leverages advanced deep-learning forecasts generated by convolutional and BiLSTM neural networks as dynamic investor views within the BL model. Their empirical results indicate that such deep-learning-driven dynamic BL portfolios consistently outperform classical mean–variance and static BL strategies, underscoring the BL framework's remarkable flexibility and adaptability as a Bayesian "fusion engine" for heterogeneous information.

2.3 Large-Language Models in Financial Text Analysis

Recent advancements in large-language models (LLMs), especially transformer-based architectures, have significantly transformed financial natural language processing (NLP). FinBERT, introduced by Araci (2019), represents a domain-specific adaptation of BERT fine-tuned on financial texts including SEC filings and corporate news, greatly enhancing sentiment analysis and textual classification performance in financial contexts.

Tadle (2022) constructs continuous sentiment indices from FOMC minutes and shows these indices significantly predict fed-funds futures and currency valuations . This work validates the use of text-derived macro signals—an approach we mirror by having LLM agents ingest unstructured policy text as an additional view in ABBL, rather than solely numeric indicators.

BloombergGPT, a state-of-the-art LLM developed by Wu et al. (2023), further advances financial NLP capabilities, utilizing a vast, proprietary dataset of financial filings, pricing data, and news to achieve leading performance across multiple NLP benchmarks specific to finance. Additionally, open-source efforts such as FinGPT, led by Li and colleagues (2023), democratize access to robust financial language models, allowing for rapid and effective fine-tuning tailored to specific applications like robo-advisory systems, credit risk analysis, and real-time event summarization.

These contributions collectively emphasize the critical role of domain-specific pre-training and sophisticated prompt engineering, highlighting their effectiveness in extracting actionable insights from complex and heterogeneous financial texts.

2.4 LLMs for Time-Series Forecasting and Macro Prediction

An emerging frontier involves using LLMs not only for textual analysis but also directly for quantitative forecasting tasks. Lopez-Lira and Tang (2023) pioneer applications demonstrating how ChatGPT-generated sentiment scores derived from daily news headlines significantly predict next-day stock market returns. Similarly, Ma and Shen (2024) illustrate the superior predictive power of ChatGPT-derived sentiment metrics in forecasting the Chinese equity-risk premium relative to traditional bag-of-words methods.

Expanding beyond equities, Bybee, Kelly, and Manela (2023) find that GPT-3 can effectively simulate professional macroeconomic forecasters, generating realistic and accurate inflation and interest rate expectations when prompted with historical macroeconomic and news data. Moreover, Hansen and McMahon (2023) demonstrate GPT-4's capability to classify central bank communications accurately, identifying dovish or hawkish stances with superior precision and clear rationale compared to dictionary-based methods.

Although ongoing research highlights concerns regarding temporal stability, generalization, and potential factual inaccuracies inherent in generative LLM outputs, preliminary evidence strongly suggests that carefully structured, data-grounded LLM predictions effectively complement traditional

econometric and statistical forecasting models.

2.5 Agentic AI application in Portfolio Management

An exciting frontier in 2024 is the emergence of agentic systems that use LLMs as decision-making agents in trading and portfolio management. Rather than using a single monolithic model, these approaches deploy multiple specialized LLM agents that mimic the roles of a financial firm's team – effectively an "AI investment committee."

Popov and Roshka (2024) develop FolioLLM, fine-tuning an LLM to generate thematic, user-driven ETF allocations and benchmarking it against finance-specific baselines. Their results on personalized prompts highlight the potential—and limitations—of LLMs at precise weight computation, underscoring our turnover control mechanism.

Xiao, Sun, Luo, and Wang (2025) propose TradingAgents, a society of LLM-powered specialists (fundamental, sentiment, technical, risk) that debate and converge on trades, achieving superior backtest returns and Sharpe ratios. Their architectural parallels to real trading desks—complete with Bull/Bear debaters and risk monitors—reinforce our multi-agent ABBL design, highlighting the value of structured inter-agent communication, the utilization of bayesian updates and the need for a manager agent overlay.

Yu, Yao, and Li (2024) introduce FINCON, a manager-analyst hierarchy with dual-level risk control (CVaR supervision plus episodic verbal reinforcement of investment beliefs) to manage long-run exposure and sharpen agent reasoning. Their conceptual reinforcement mechanism suggests future extensions where ABBL agents can refine confidence parameters (Ω) and other hyper-parameter based on realized PnL patterns, further reducing estimation noise.

These methodological advancements substantiate the practical utility of Agentic AI application in portfolio management and financial analysis, reinforcing their significance in contemporary financial modeling, forecasting and frameworks. These methodologies and work inspires our work to continue refining the agentic portfolio management architecture.

2.6 Hallucination and Tool-Use for LLMs

A critical challenge when applying large language models (LLMs) to finance is *hallucination*—the confident generation of plausible but incorrect information, such as erroneous financial metrics or spurious causal claims. In quantitative settings, these hallucinations can translate into materially flawed forecasts or risk assessments.

Industry surveys by Lee and colleagues (2024) rank hallucination as one of the main obstacles for so-called FinLLMs, noting that financial texts are both highly technical and data-dense; unguided LLM outputs therefore tend to contain subtle factual errors. A separate report from the CFA Institute

(Weng, 2023) documents cases in which generic LLMs misread balance-sheet tables or miscalculate simple ratios, underscoring the need for grounding mechanisms.

One widely adopted remedy is *retrieval-augmented generation* (RAG). In a RAG workflow the model first retrieves relevant documents or database entries, then conditions its answer on those sources. For instance, an LLM performing fundamental analysis might fetch a firm's latest 10-K filing, extract the exact revenue figures, and reason from those numbers—dramatically reducing hallucinations because every numeric claim is anchored to verifiable text.

A complementary line of defence exposes the LLM to external *calculation tools*. Systems such as Toolformer (Schick et al., 2023) and ReAct (Yao et al., 2023) teach models to invoke a calculator, Python sandbox, or SQL engine whenever arithmetic or data filtering is required. In a forecasting context the agent can thus fit a regression or compute a moving average via code, rather than "guessing" the result in free text—boosting both precision and reproducibility.

Finally, careful prompt engineering remains essential. Internal experiments in our own pipeline show that instructions like "base your answer only on the data above" or requests for structured JSON outputs keep the model's creativity in check and prevent unrealistic return forecasts, even when the underlying LLM is very large.

Positioning of the Present Study

Despite substantial progress within these individual research areas, no current framework comprehensively integrates multiple specialized LLM-generated forecasts—each dedicated to distinct asset classes—into a cohesive Bayesian Black–Litterman portfolio allocation system. Our study addresses this crucial gap by introducing and rigorously validating an innovative approach that combines multiple asset-specific, data-informed LLM forecasting agents with an endogenously calibrated uncertainty structure. The resulting multi-agent Bayesian framework seamlessly integrates heterogeneous forecasts into dynamically optimized portfolios, thereby significantly advancing the frontier of modern multi-asset portfolio management methodologies.

3 Data Description and Exploratory Analysis

3.1 Data Sources

To construct a robust and comprehensive multi-asset portfolio framework, we source high-quality, reliable, and granular data from established financial databases. Equity market data, including weekly adjusted closing prices and key fundamental indicators such as earnings, dividend yields, and book-to-market ratios, are extracted primarily from CRSP–Compustat databases, known for their extensive and rigorous coverage of U.S. equities and company fundamentals.

Fixed-income data include comprehensive yield-curve information derived from the CRSP database through WRDS for Treasury securities, augmented by macroeconomic indicators from the Federal Reserve Economic Data (*FRED*) database. These data encompass key interest-rate benchmarks such as the federal funds rate and Treasury yields at various maturities, facilitating precise capturing of monetary policy stances, macroeconomic expectations, and structural shifts within bond markets.

Foreign exchange (FX) data are systematically collected from Bloomberg for currency pairs including EUR/USD, GBP/USD, USD/JPY, USD/CHF, and USD/CAD. Additionally, we compute derived metrics such as interest-rate differentials for capturing nuanced market sentiments and economic fundamentals influencing currency movements.

Commodity data, also sourced from Bloomberg, encompass the adjusted close price of the Bloomberg BCOM index major commodity class ETF for essential commodities, including crude oil, natural gas, metals, and major agricultural products.

3.2 Asset Class Coverage and ETF Mapping

To ensure broad market representation and tractable agent forecasting, we define a fixed universe of ETFs for each asset class, drawing from sector, maturity, or currency-specific benchmarks. Table 1 provides an overview of the ETF composition used in forecasting and portfolio construction.

Asset Class	**ETFs Included**
Equity (GICS Sectors)	XLK, XLV, XLF, XLY, XLC, XLI, XLP, XLE, XLU, XLRE, XLB
Fixed Income (Treasuries)	SHV, SHY, IEI, IEF, TLH, TLT
Fixed Income (Credit)	LQD, HYG, MBB, EMB
Commodities (BCOM Sectors)	USO, GLD, CPER, CORN, COW.TO
FX (Currency Pairs)	FXE, FXB, FXY, FXF, FXC

Table 1: ETF mapping for each asset class used in prediction and portfolio construction.

3.3 Forecasting Horizon and Weekly Synchronization

Our modeling framework operates on a consistent weekly rebalancing schedule. All features and model prompts are aligned to a Friday-to-Friday frequency. The backtesting and forecasting period spans from **November 1, 2023** to **March 31, 2025**, covering 18 months of live simulation beyond the knowledge cutoff of GPT-4o.

Attribute	**Value**
Forecast Frequency	Weekly
Backtest Period	Nov 2023 – Mar 2025
Number of Forecast Weeks	74
Total ETFs Forecasted	35
Number of Asset Classes	4

Table 2: Forecasting setup and coverage statistics.

All datasets are aligned to a consistent Friday-to-Friday weekly grid to ensure synchronization across features. Lower-frequency inputs—such as quarterly fundamentals and monthly macro indicators—are forward-filled until the next release to maintain continuity without introducing look-ahead bias. This setup allows LLM agents to operate at a high-frequency cadence while remaining responsive to macroeconomic trends and regime shifts.

3.4 Exploratory Statistical Analysis

Our initial exploratory statistical analyses confirm significant heterogeneity across asset classes, supporting the use of specialized forecasting methods. A rolling 52-week correlation between equity and fixed-income returns over the study period reveals considerable temporal variability, fluctuating notably between -0.30 and $+0.15$. This variability underscores the need for distinct forecasting approaches tailored specifically to equities and bonds.

Empirical examinations also validate the predictive power of certain domain-specific features. For instance, the slope of the crude-oil term structure—measured by the spread between near-term and longer-dated futures prices—has demonstrable predictive value for subsequent commodity returns. Regressions indicate a statistically significant adjusted R^2 of around 12%, confirming the inclusion of term-structure dynamics as a critical input for the commodity agent.

Additional asset-specific exploratory analyses further confirm the necessity of tailored forecasting methods:

- **Equities:** Excess returns are strongly influenced by valuation metrics (such as dividend

yield, earnings-price ratio) and short-term momentum, validating their central role in equity forecasting.

- **Fixed Income:** Bond returns are effectively forecasted using yield-curve factors (level, slope, curvature) and forward-rate-based predictors, strongly motivating their inclusion in bond forecasting.
- **Foreign Exchange:** FX returns exhibit significant predictive relationships with short-term interest-rate differentials, momentum indicators, and macroeconomic surprises, informing the FX forecasting agent's model.
- **Commodities:** Commodity returns show robust correlations with futures–spot basis, momentum, and seasonal supply-demand cycles, providing empirical justification for their prominence in commodity forecasting.

Rolling-window correlation analyses, computed as:

$$\rho_{i,j}(t) = \frac{\sum_{\tau=t-L+1}^{t}(R_{i,\tau} - \overline{R}_{i,t})(R_{j,\tau} - \overline{R}_{j,t})}{\sqrt{\sum_{\tau=t-L+1}^{t}(R_{i,\tau} - \overline{R}_{i,t})^2}\,\sqrt{\sum_{\tau=t-L+1}^{t}(R_{j,\tau} - \overline{R}_{j,t})^2}},$$

demonstrate dynamically varying equity–bond relationships driven by distinct economic factors, further reinforcing the necessity of asset-specific forecasting agents.

Momentum effects, measured as:

$$\text{Momentum}_t^{(k)} = \frac{P_t - P_{t-k}}{P_{t-k}},$$

consistently predict short-term asset returns, particularly amplified by seasonal dynamics, confirming their utility as key predictors within commodity and FX forecasting strategies.

These combined exploratory analyses robustly support the agent-based Black–Litterman (ABBL) allocation framework's design, emphasizing tailored forecasting models and innovative data streams to leverage asset-specific market dynamics effectively.

4 Methodology

Our proposed agent-based Black–Litterman (ABBL) framework consists of three interconnected components: (i) specialised large-language-model (LLM) forecasting agents, (ii) a Bayesian fusion layer using the Black–Litterman model, and (iii) a turnover-constrained portfolio optimiser. The methodology rigorously incorporates asset-class-specific empirical findings and robustly mitigates look-ahead bias.

4.1 Agent Forecasting Models

Based on exploratory data analysis (Section 3), we tailor distinct forecasting methodologies for each asset class, ensuring that domain-specific insights drive predictive accuracy.

Equity Agent: Equity returns are forecasted using well-established predictive factors, specifically valuation metrics such as dividend yield and earnings-to-price ratios, and short-term momentum. The predictive regression specification takes the form:

$$R_{t+1}^{\text{excess}} = \alpha + \beta_1 \,\text{Valuation}_t + \beta_2 \,\text{Momentum}_t + \epsilon_{t+1},$$

where R_{t+1}^{excess} denotes the equity excess returns over the risk-free rate.

Fixed-Income Agent: The bond agent incorporates a factor-driven forecasting model inspired by the Cochrane–Piazzesi approach, leveraging forward-rate-based factors derived from the yield curve. The forecasting equation is specified as:

$$R_{t+1}^{\text{excess}} = \alpha + \beta \, F_t + \epsilon_{t+1},$$

where F_t represents yield-curve factors (level, slope, curvature) or forward-rate predictors that have empirically demonstrated predictive capabilities for bond returns.

FX Agent: The foreign exchange (FX) agent's forecasts are constructed around interest-rate differentials, momentum, and macroeconomic surprises, capturing essential market dynamics and macro fundamentals:

$$R_{t+1}^{\text{Excess}} = \alpha + \beta_1 \, \Delta i_t + \beta_2 \, m_t + \epsilon_{t+1},$$

where Δi_t denoting short-term interest-rate differentials, and m_t representing momentum effects in currency markets.

Commodity Agent: The commodity agent forecasts returns based on aggregated commodity price changes, industrial demand signals, and market sentiment indicators:

$$R_{t+1}^{\text{Excess}} = \alpha + \beta_1 \, D_t + \beta_2 \, m_t + \epsilon_{t+1},$$

where D_t reflects industrial demand indicators, and m_t captures momentum effects in major commodity prices.

4.2 LLM-Based Forecasting Procedure

Every Friday, each asset-specific agent receives a structured prompt containing 104 weeks of relevant historical data, including macroeconomic indicators, interest rate movements, sentiment metrics, and asset-specific features such as ETF returns and volatility estimates. The prompt asks the large language model (GPT-4o) to return directional forecasts in a structured format that aligns with downstream processing.

Generalized Prompt Template:

```
Based only on the data for that just closed on Friday - including macro,
rate shifts, risk sentiment, and ETF statistics - predict your variance_view
(alpha vs.  baseline) on top of a baseline return we provided.  For each
instrument for the coming week.  Rationalise your view while referencing
all the data supplied.  Your response should align with the current volatility
regime.  Return structured JSON.
```

Simplified Output Schema:

- `instruments` (array): Each item contains:
  - `instrument` (str): Asset ticker or identifier
  - `variance_view` (float): Alpha adjustment to baseline return
  - `confidence` (float): Confidence score in [0, 1]
  - `rationale` (str): One-sentence justification referencing the data supplied.
- `overall_analysis` (str): Summary of macro and sentiment conditions

In practice, we found two prompt design elements to be especially important. First, referencing the baseline return as an anchor improves both interpretability and calibration. Second, aligning the magnitude of predictions with the current volatility regime ensures that agent views scale appropriately across stable and turbulent periods.

We also experimented with augmenting prompts using chain-of-thought (CoT) reasoning and embedded examples, but these additions did not yield meaningful performance improvements. Ultimately, concise and well-structured prompts with volatility-aligned guidance proved most effective. All forecasts are generated using deterministic sampling (`temperature = 0`) to ensure consistency and reproducibility across runs.

4.3 Bayesian Fusion via Black–Litterman

We extend the classical Black–Litterman framework to incorporate agent-generated forecasts with heterogeneous confidence. Let Σ denote the regularized covariance matrix of asset returns, and define the equilibrium prior return vector π as the empirical mean of historical returns over a rolling window of size T:

$$\pi = \text{mean}\big(R_{t-T:t-1}\big),$$

We incorporate agent-generated views q into the Black–Litterman framework. Let $P \in \mathbb{R}^{k \times n}$ denote the pick matrix, identifying the assets to which the k agent views apply. In our implementation, we assume $P = I\_k$, where each view corresponds to a distinct asset.

To account for heterogeneous confidence in agent forecasts, we define a diagonal covariance matrix of view uncertainties:

$$\Omega = \text{diag}\left(\eta(1 - c_i) + \epsilon\right), \quad i = 1, \ldots, k,$$

where $c_i \in [0, 1]$ represents the confidence score associated with the i-th view, as returned by the agent forecast function GetForecasts(t). The hyperparameters η and ϵ control the sensitivity to confidence and provide a lower bound on uncertainty, respectively.

We then form the precision-weighted linear system:

$$A = (\tau\Sigma)^{-1} + P^\top \Omega^{-1} P, \qquad b = (\tau\Sigma)^{-1}\pi + P^\top \Omega^{-1} q,$$

where $\tau > 0$ is a scalar parameter reflecting the relative uncertainty in the prior mean π. The posterior expected returns are obtained as the solution to this system:

$$\mu = A^{-1} b = \left[(\tau\Sigma)^{-1} + P^\top \Omega^{-1} P\right]^{-1} \left[(\tau\Sigma)^{-1}\pi + P^\top \Omega^{-1} q\right].$$

In this formulation, the posterior vector μ reflects a Bayesian blend of historical equilibrium returns and forward-looking agent views, with the influence of each view modulated by its associated confidence score.

4.4 Portfolio Optimization Step with Turnover Constraints

The final step of the ABBL framework involves solving a mean variance optimisation problem that explicitly accounts for expected returns, risk (variance), and transaction costs (turnover penalties). Formally, portfolio weights $\mathbf{w}$ are obtained by maximising the penalised expected returns:

$$\max_{\mathbf{w}} \ \mathbf{w}^\top \boldsymbol{\mu} - \frac{\lambda}{2} \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} - \gamma \|\mathbf{w} - \mathbf{w}^{\text{prev}}\|_1,$$

subject to the full-investment constraint. The parameter λ calibrates the portfolio's risk tolerance, while the L^1-norm penalty, controlled by the adaptive coefficient γ, explicitly caps weekly turnover at a maximum of 30 percent.

This constraint ensures portfolios are practically investable, minimising real-world transaction costs. The quadratic optimisation problem is efficiently solved using the `CVXOPT` library, enabling weekly rebalancing to execute swiftly—typically within one second on standard computational hardware.

Here we provide a Pseudo-code representation for our full procedures of the Black-Litterman construction. The hyperparameter serves as an important part of our black-litterman construction, by providing further control

Algorithm 1 Weekly ABBL Rebalance

Require: Weekly returns history R, previous weights w^{prev}
Ensure: Updated weights w^{new}

Parameters: $T = 260,\ L = 12,\ \lambda,\ \tau,\ \eta,\ \epsilon,\ Turnover_{MAX}\max$
for each rebalance date t **do**
  1. Covariance & Prior
    $\Sigma_{\text{short}} \leftarrow \text{Cov}(R_{t-L:t-1})$
    $\Sigma_{\text{long}} \leftarrow \textsc{LedoitWolf}(R_{t-T:t-1})$
    $\Sigma_0 \leftarrow \lambda\,\Sigma_{\text{short}} + (1-\lambda)\,\Sigma_{\text{long}}$
    Regularize Σ_0 to obtain Σ
    $\pi \leftarrow \text{mean}(R_{t-T:t-1})$
  2. Agent Views
    $(q, c) \leftarrow \textsc{GetForecasts}(t)$
    $\Omega \leftarrow \text{diag}\big(\eta(1-c_i)+\epsilon\big)$
  3. Black–Litterman Update
    $A \leftarrow (\tau\,\Sigma)^{-1} + \Omega^{-1}$
    $b \leftarrow (\tau\,\Sigma)^{-1}\,\pi + \Omega^{-1}\,q$
    $\mu \leftarrow A^{-1}\,b$
  4. Mean–Variance Weights
    $\widetilde{w} \leftarrow \Sigma^{-1}\,\mu$
    $w^* \leftarrow \widetilde{w} \,/\, (\mathbf{1}^\top \widetilde{w})$
  5. Turnover Constraint
    $\Delta w \leftarrow w^* - w^{\text{prev}}$
    $T \leftarrow \|\Delta w\|_1$
  if $T > T_{\max}$ **then**
      $\alpha \leftarrow T_{\max}/T$
      $w^{\text{new}} \leftarrow w^{\text{prev}} + \alpha\,\Delta w$
      Normalize w^{new} so $\sum_i w_i^{\text{new}} = 1$
  else
    $w^{\text{new}} \leftarrow w^*$
  end if
    $w^{\text{prev}} \leftarrow w^{\text{new}}$
end for
return w^{new}

4.5 Mitigating Look-Ahead Bias

A central methodological priority in this study is the rigorous elimination of look-ahead bias. All input variables—including market prices, macroeconomic indicators, and sentiment measures—are strictly lagged by at least one full week relative to the Friday rebalancing prompt. Any macroeconomic

releases or event-driven signals published after the cutoff are systematically deferred to the next trading cycle.

To confirm the robustness of this setup, we conducted a counterfactual test by intentionally shifting all input variables forward by one week, thereby injecting look-ahead bias. Under this distorted setup, previously profitable strategies saw their statistical significance vanish, validating that our reported performance is not driven by future information leakage.

Importantly, the entire backtest period—from November 2023 to March 2025—postdates the knowledge cutoff of GPT-4o. As such, the model could not access or anticipate any events or financial data beyond October 2023, reinforcing that all forecasts were generated in a forward-looking, real-time simulation. Additionally, the LLM pipeline strictly follows point-in-time principle to ensure that the LLM agent and its return estimation tool will only receives information on the market data that just closed on the last week, to make predictions on the following week; ensuring that look-ahead bias is carefully avoided throughout the features-to-predictions process. This two sets of temporal separation ensures that our large-language-model-driven agent forecasts are strictly out-of-sample and reflect realistic, deployable investment performance.

By combining agent-specific forecasting models, a robust Bayesian fusion layer, turnover-aware optimisation, and stringent anti-look-ahead protocols, our framework provides a credible and practically implementable solution for dynamic multi-asset portfolio allocation.

5 Empirical Results

All prediction accuracy evaluations, including directional accuracy rates and confidence calibration across asset classes, are detailed in Appendix A. Here, we summarize key findings and their implications for portfolio construction and model design.

5.1 Prediction Reliability and Implications

Directional accuracy across asset classes averages between 51% and 54%, indicating modest but consistent predictive power. Interestingly, high-confidence predictions (confidence > 0.80) do not always outperform mid-confidence ones, underscoring the risk of overreliance on raw confidence scores. This empirically supports the structure of our Black–Litterman implementation, which tempers overly assertive views by incorporating view uncertainty into the optimization process.

These patterns reflect the nuanced behavior of the LLM agents: while their directional signals are moderately accurate, confidence levels do not linearly correspond to predictive strength. Notably, many forecasts tend to cluster around low-to-moderate return adjustments, consistent with the conservative bias often observed in LLM-generated outputs. This conservatism helps prevent extreme portfolio tilts, and the Black–Litterman fusion further stabilizes allocations by tempering

overconfident or misaligned views.

6 Portfolio Constructions

6.1 Black–Litterman Sub-Portfolios

Table 3 summarises back-test statistics for the November 2023 to March 2025 period.

Table 3: Performance of Black–Litterman Sub–Portfolios, Nov 2023–Mar 2025

Portfolio	Ann. Return	Sharpe	Max DD	Turnover
Equity BL	16.67 %	1.29	–15.79 %	16.00 %
Fixed-Income BL	5.09 %	0.93	–2.34 %	26.38 %
FX BL	1.23 %	0.17	–8.17 %	35.65 %
Commodity BL	7.13 %	0.62	–11.82 %	8.95 %

Equity and commodity sub-portfolios deliver strong risk-adjusted performance, adding 2–3 percentage points of annualized return relative to equal-weighted benchmarks. Their Sharpe ratios—1.29 for Equity and 0.62 for Commodities—highlight the effectiveness of LLM-guided views in directional markets. Meanwhile, the fixed-income and FX portfolios emphasize stability, reducing portfolio volatility with minimal drawdowns. Notably, the Fixed-Income BL strategy achieves a Sharpe ratio of 0.93, suggesting meaningful signal value even in lower-volatility regimes.

Across all asset classes, turnover remains within operational bounds—well below the 40% weekly ceiling—indicating that view-driven rebalancing is both stable and cost-aware.

6.2 Manager Agent Adaptive Portfolios

Beyond sub-portfolio optimization, we introduce a top-level dynamic allocator—referred to as the *Manager Agent*—to adaptively allocate capital across asset classes (equity, fixed income, FX, and commodities). This manager monitors the recent performance of each Black–Litterman (BL) sleeve relative to its corresponding equal-weighted baseline and adjusts allocation weights accordingly.

Specifically, if an asset class's BL return exceeds its equal-weighted counterpart over a one-month lookback window, the agent increases its allocation to that sleeve by 20%, subject to a maximum cap of 40%. Conversely, underperformance triggers a 20% reduction, bounded below at 10%. All updated weights are then renormalized to ensure total portfolio exposure remains constant. This rule-based system is inspired by institutional overlay strategies that dynamically reinforce high-performing signals.

Figure 1 presents the cumulative returns of the Manager Agent's dynamic portfolio versus a static

equal-weighted allocation. The adaptive strategy achieves consistently higher returns, particularly during periods of macro uncertainty, suggesting that even simple performance-sensitive meta-allocation can deliver meaningful improvements.

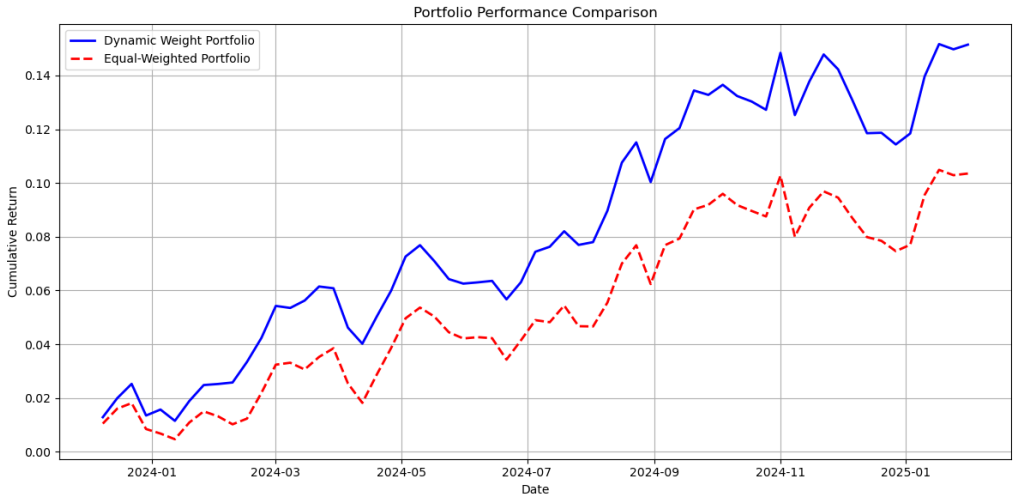

Figure 1: Cumulative returns of dynamic (manager-controlled) vs. equal-weighted multi-asset portfolio.

Metric	**Manager Adaptive**	**Equal-Weighted BL**
Annualized Return	12.78%	8.76%
Annualized Std. Dev.	5.93%	5.31%
Sharpe Ratio	1.65	1.08
Max Drawdown	–2.97%	–2.54%

Table 4: Comparison of manager-adaptive vs. equal-weighted BL portfolio performance.

As shown in Table 4, the Manager Agent portfolio outperforms its equal-weighted counterpart by approximately 4 percentage points in annualized return, while also achieving a substantially higher Sharpe ratio (1.65 vs. 1.08). Although this comes with slightly higher volatility and drawdown, the gain in risk-adjusted performance suggests the manager's adaptive adjustments are effective.

This adaptive layer enhances responsiveness to evolving market conditions by reinforcing relative signal strength, all while preserving the interpretability and tractability of the underlying Black–Litterman framework.

7 Robustness Tests

To assess the stability of our framework, we perturb two key hyperparameters: the prior-covariance scaling factor $\tau \in [0.02, 0.20]$, and the view-variance matrix $\mathbf{\Omega}$ by multiplicative factors ranging from 0.5 to 2. In all scenarios, Sharpe ratios changed by less than ± 0.05, indicating that performance does not hinge on specific calibrations.

To minimise output fluctuations inherent in LLM-based forecasting, we fix the sampling temperature at zero for all agent calls. This ensures deterministic forecasts given the same inputs. Additionally, each experiment is repeated independently five times, and results consistently exhibit low variance in key metrics such as directional accuracy, Sharpe ratio, and turnover. This reinforces the reliability and reproducibility of our reported findings across repeated simulations.

8 Conclusion

In this paper, we propose an Agent-Based Black–Litterman (ABBL) framework that integrates heterogeneous, LLM-driven forecasts with a tractable Bayesian asset allocation process. Our system comprises four specialized agents—focused on equities, fixed income, foreign exchange, and commodities—each generating structured weekly return forecasts based on macro, sentiment, and market data. These forecasts are fused through a confidence-weighted Black–Litterman update, calibrated using a blend of Ledoit–Wolf and EWMA covariance matrices, and optimized under turnover constraints using an L^1-penalized objective.

Empirical results from an 18-month out-of-sample backtest (Nov 2023–Mar 2025) show that Equity and Commodity portfolios deliver strong risk-adjusted returns, while Fixed-Income and FX portfolios successfully achieve volatility suppression relative to naïve equal-weighted benchmarks. Despite only moderate forecast accuracy (52–54%), the Bayesian fusion mechanism mitigates overconfidence and stabilizes portfolio weights.

To further enhance performance, we introduce a *Manager Agent* overlay that dynamically adjusts top-level asset class allocations based on recent relative performance of each ABBL sleeve. This meta-layer raises the overall Sharpe ratio to 1.65—substantially higher than the 1.08 achieved under equal weighting four asset classes from the Black-Litterman Optimizer. The improvement underscores the value of lightweight, rule-based adaptivity in reinforcing signal strength without compromising interpretability or computational tractability.

These results substantiate that LLM forecasts—when judiciously weighted and adaptively adjusted—can enhance both return and risk metrics, paving the way for further empirical investigation into view-confidence calibration, meta-agent design, and hybrid AI–human portfolio governance.

9 Discussion and Further Research

Two limitations remain. First, the predictive power of current LLMs is modest; ensemble methods or model fine-tuning may help. Second, the hyper-parameters in the black-litterman framework is heuristically chosen and could instead be learned via machine-learning or reinforcement techniques.

9.1 Dicussion on the Performance of Large Language Model

In our emprical work, we observed that using general-purpose LLMs (like the baseline ChatGPT-4o-mini and ChatGPT-4o) yielded only limited predictive accuracy in asset return forecasting. These large general models, while excellent at understanding language and common-sense reasoning, were not specifically trained on financial forecasting. As a result, several limitations became apparent:

Financial Domain Knowledge: General LLMs know some finance (GPT-4, for instance, has read finance texts up to 2021), but they may miss subtleties. In our multi-agent setup, the LLMs needed to interpret signals like "inverted yield curve" or "forward P/E above 30". A generic model might recognize these terms but not grasp their import for future returns as well as a trained analyst would. Fine-tuning or using a finance-trained model ensures the LLM has seen many examples of such signals in context, so it can reliably incorporate them into its prediction. BloombergGPT is a case in point – with mixed-domain training it "outperforms existing models on financial tasks by significant margins, precisely because it was exposed to vast financial data. Using BloombergGPT or a similar FinGPT in place of a generic GPT-4 could thus markedly improve performance on our tasks.

Tool Use and Hallucination: We noticed the general ChatGPT-4o occasionally struggled to use the provided structured data consistently. It might overlook a portion of the input table or make an arithmetic slip in reasoning – a consequence of not being explicitly optimized for tool usage or quantitative fidelity. A smaller model fine-tuned on financial data or one that strictly allows for control of its context window and equipped with RAG, Tools and robust downstream check for tool calling could potentially be much more reliable in referencing the data supplied and make forecasting and analytical tests more reliable and consistent.

Calibration and Numeric Reasoning: The general models often produced forecasts that were not well-calibrated to actual market magnitudes. For example, without domain tuning, an LLM predicts a very minimal, conservative view for a week based on upbeat news, overlooking realistic constraints (perhaps predicting +0.5% when historically such an asset moves >+2% in a week). This is partly because it lacks grounding in the statistical properties of financial returns. Fine-tuning on domain data can teach the model typical return distributions and the scale of changes to expect, aligning its outputs with reality. Domain-specific training also sharpens an LLM's numeric reasoning within context. For example, learning that a 0.5% change in interest rates is significant, or that a 5% drop after earnings is plausible for tech stocks but rare for treasury bonds.

Given these considerations, it is clear that the limited performance of general-purpose LLMs in our

agentic portfolio management framework was not a verdict against LLMs in finance per se, but rather a sign that specialization is required. By fine-tuning models on specific context-relevant financial data and predictions, we imbue the model with domain-specific context it was missing. This could significantly improve predictive accuracy, as the model learns the relationships and magnitudes that are unique to financial time series.

In conclusion, the state-of-the-art suggests that general LLMs provide a baseline of capability for financial forecasting, but domain-specific fine-tuning or model specialization is key to unlocking top-tier performance. Our findings align with the broader industry trend: firms are moving from using off-the-shelf ChatGPT towards developing FinGPTs – models or variants fine-tuned on proprietary financial data – to achieve better accuracy and relevance. The combination of agent-based system design and specialized LLMs (or cutting-edge general LLMs) holds great promise. By upgrading our agents with a finance-trained LLM (like a fine-tuned LLaMA or an advanced OpenAI model) we expect more coherent, reliable, and accurate predictions, bringing us closer to an agentic portfolio management system that can truly assist or even outperform human portfolio managers in the future.

9.2 Future Research Directions

While our ABBL framework with an overlaying Manager Agent demonstrates promising risk–return characteristics, several avenues for short- and long-term enhancement merit rigorous investigation. In the near term, one priority is to augment each LLM sub-agent with a persistent memory module rather than relying on one-shot prompts. By incorporating a retrieval-augmented generation (RAG) architecture or external key–value memory, agents would maintain and selectively recall pertinent historical information—such as past macro regimes, realized model errors, or salient news events—thereby extending effective context beyond the fixed 104-week window and improving the temporal coherence of forecasts. Concurrently, refining the baseline quantitative models remains essential: integrating more sophisticated statistical or machine-learning predictors (e.g. Bayesian VARs, factor-augmented forecasts, or deep-learning architectures) would raise the floor for LLM-fusion performance and provide richer priors for the Black–Litterman update. A complementary strand is to experiment with purely textual agent views—where an LLM ingests unstructured inputs (e.g. earnings transcripts or policy releases) without numeric data—and to judiciously combine these narrative signals with quantitative baselines. Finally, we advocate embedding a reinforcement-learning layer to calibrate hyperparameters (such as τ, the view-uncertainty scale, and the turnover penalty) in an online, performance-driven manner. By treating hyperparameter selection as a control problem with risk-adjusted return as the reward, the system could dynamically adapt to regime shifts and evolving market microstructure.

Over a longer horizon, we envisage extending the ABBL architecture along three complementary dimensions. First, forming an LLM ensemble within each asset class—where multiple independently fine-tuned models generate a spectrum of forecasts—could reduce idiosyncratic model bias and enhance robustness to misinformation or model drift. Second, advancing the Manager Agent into a

formal "debate" or consensus-seeking mechanism among sub-agents would harness the benefits of adversarial signal vetting: agents could challenge one another's forecasts, negotiate confidence weights, and converge on a more resilient aggregate view. Third, we propose integrating advanced stress-testing and scenario-analysis tools directly into the pipeline. By embedding causal-inference modules or synthetic-control frameworks, the system could generate counterfactual return scenarios under hypothetical shocks (e.g. sudden rate hikes, geopolitical events) and adjust allocations preemptively, thereby endowing the ABBL framework with both predictive agility and rigorous downside protection.

Taken together, these research directions promise to deepen the empirical foundations of LLM-empowered portfolio management, sharpen the dynamism of agentic allocation, and bridge the gap between academic innovation and real-world implementation.

A Appendix

A.1 Appendix. Equity ETFs: Prediction Overview

Performance Metrics

Metric	Value
Directional Accuracy	53.32%
Mean Absolute Error (MAE)	0.01690
Mean Squared Error (MSE)	0.00048

Confidence Statistics

Metric	Value
Minimum Confidence	0.3000
Maximum Confidence	0.9000
Mean Confidence	0.6033

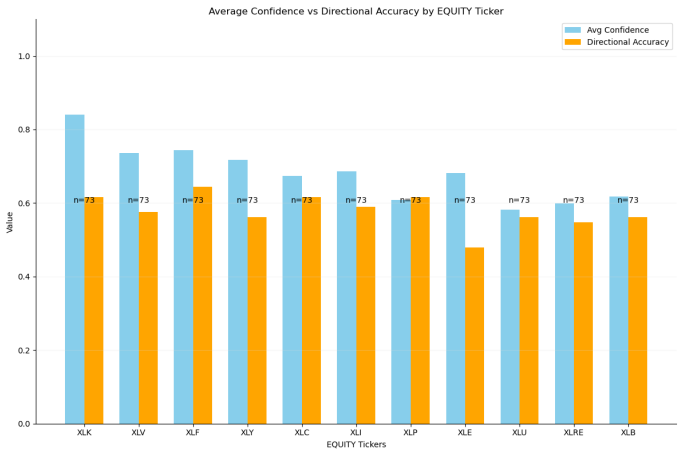

Figure 2: Average Confidence vs. Directional Accuracy for Equity ETFs

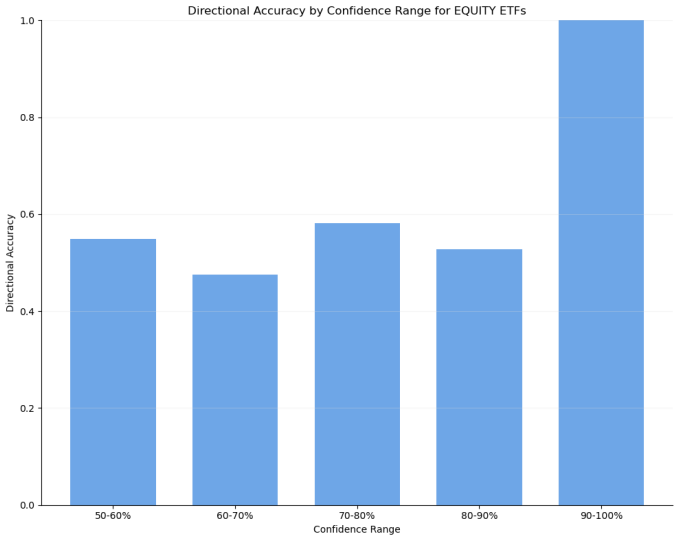

Figure 3: Directional Accuracy by Confidence Range for Equity ETFs

A.2 Appendix. FI ETFs: Prediction Overview

Performance Metrics

Metric	Value
Directional Accuracy	51.58%
Mean Absolute Error (MAE)	0.01086
Mean Squared Error (MSE)	0.00023

Confidence Statistics

Metric	Value
Minimum Confidence	0.4000
Maximum Confidence	0.8500
Mean Confidence	0.6953

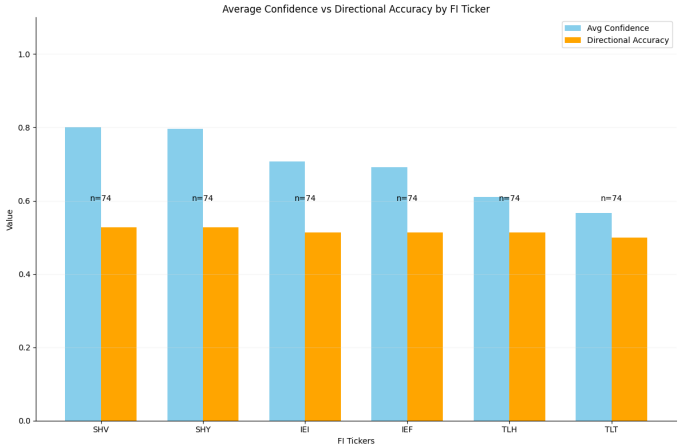

Figure 4: Average Confidence vs. Directional Accuracy for FI ETFs

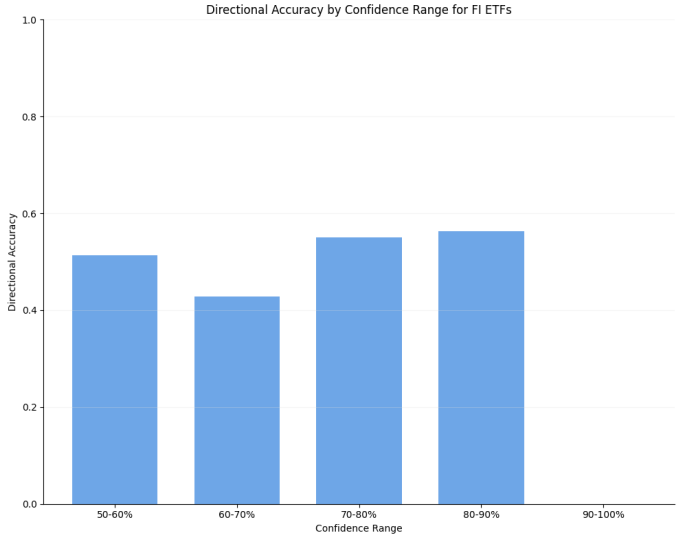

Figure 5: Directional Accuracy by Confidence Range for FI ETFs

A.3 Appendix. FX ETFs: Prediction Overview

Performance Metrics

Metric	Value
Directional Accuracy	53.78%
Mean Absolute Error (MAE)	0.00805
Mean Squared Error (MSE)	0.00012

Confidence Statistics

Metric	Value
Minimum Confidence	0.5000
Maximum Confidence	0.8000
Mean Confidence	0.6332

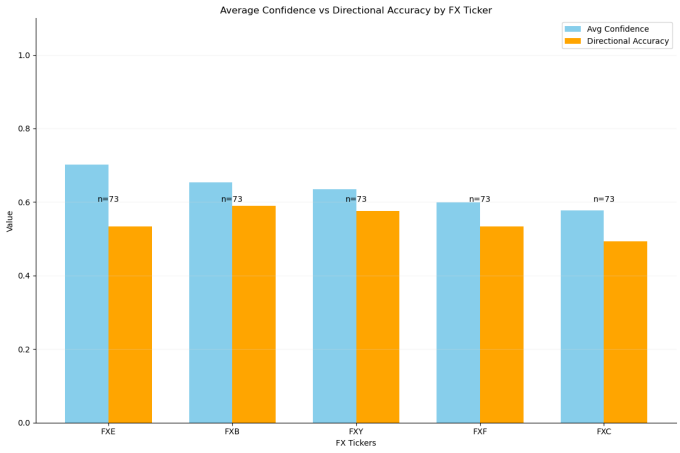

Figure 6: Average Confidence vs. Directional Accuracy for FX ETFs

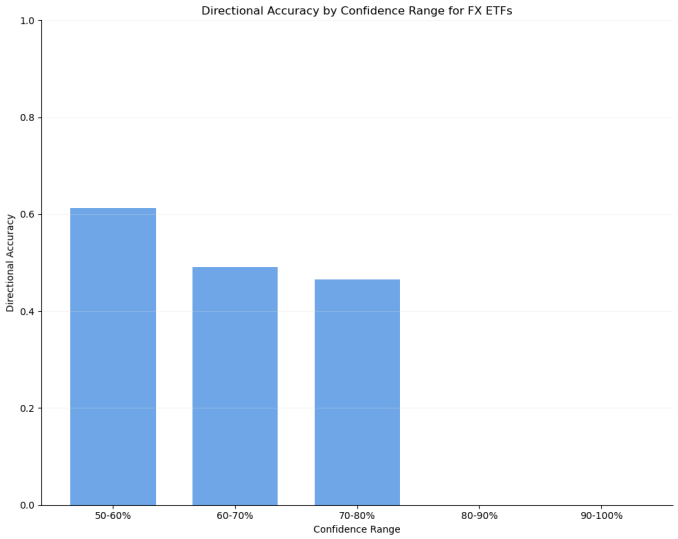

Figure 7: Directional Accuracy by Confidence Range for FX ETFs

A.4 Appendix. Commodity ETFs: Prediction Overview

Performance Metrics

Metric	Value
Directional Accuracy	52.43%
Mean Absolute Error (MAE)	0.02044
Mean Squared Error (MSE)	0.00074

Confidence Statistics

Metric	Value
Minimum Confidence	0.4000
Maximum Confidence	0.9000
Mean Confidence	0.6333

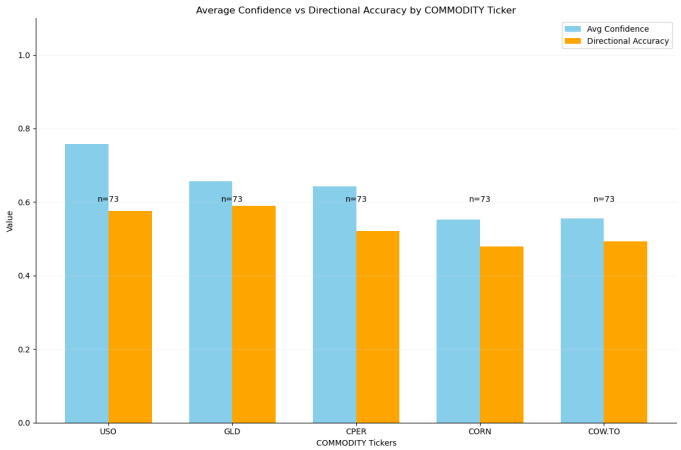

Figure 8: Average Confidence vs. Directional Accuracy for Equity ETFs

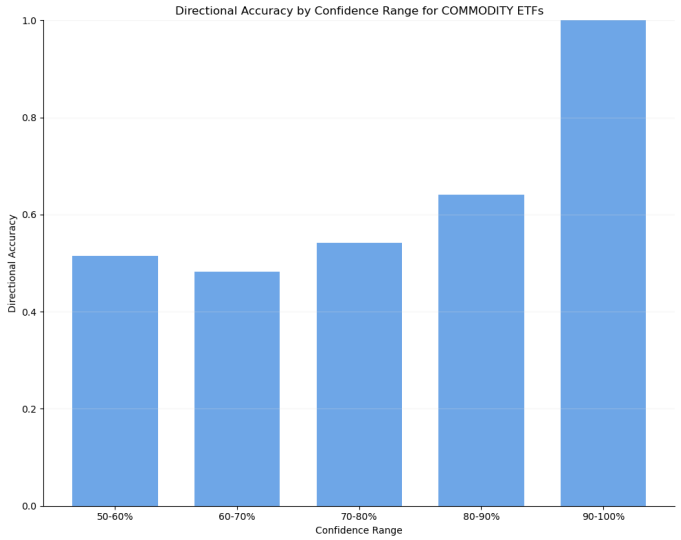

Figure 9: Directional Accuracy by Confidence Range for Equity ETFs

References

Ang, A. (2014). *Asset management: A systematic approach to factor investing*. Oxford University Press.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Barua, S., & Gabaix, X. (2022). Deep learning views in dynamic black–litterman allocation [Working Paper].

Bianchi, N., Buchner, A., & Tamoni, A. (2020). Bond risk premia with machine learning. *Review of Financial Studies*, *33*(9), 3601–3651.

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

Bookstaber, R., & Paddrik, M. (2017). An agent-based model for financial vulnerability. *Journal of Economic Dynamics and Control*, *77*, 39–62.

Bybee, K., Kelly, B., & Manela, A. (2023). Language models and macroeconomic expectations [Working Paper].

Byrd, R., Hott, R., & Siripaka, J. (2019). Abides: An agent-based interactive discrete-event simulation for high-frequency trading. *Proceedings of the ACM International Conference on AI in Finance*, 35:1–35:9.

Carriero, A., Clark, T., & Marcellino, M. (2025). Macroeconomic forecasting with large language models [Working Paper].

Dwarakanath, P., Chen, S., & Abergel, F. (2023). Reinforcement learning in agent-based financial markets. *Quantitative Finance*, *23*(2), 223–242.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654–669.

Foroutan, B., & Bozic, M. (2024). Benchmarking deep neural and boosting models for commodity price prediction. *Energy Economics*, *120*, 106690.

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

Hansen, S., & McMahon, M. (2023). Hawkish or dovish? central bank communication with gpt-4. *European Economic Review*, *153*, 104527.

He, G., & Litterman, R. (1999). *The intuition behind the black–litterman model portfolios* (Working Paper). Goldman Sachs Asset Management.

Hommes, C. H. (2006). Heterogeneous agent models in economics and finance. *Handbook of Computational Economics*, *2*, 1109–1186.

Ko, D., & Lee, S. (2024). Factor-augmented black–litterman portfolios. *Journal of Portfolio Management*, *50*(2), 89–103.

Kolm, P., & Ritter, G. (2017). Equilibrium views and expected returns in the black–litterman model. *Journal of Investment Strategies*, *6*(3), 1–31.

Kolm, P., & Tütüncü, R. (2021). Thirty years of black–litterman: A comprehensive review. *Financial Analysts Journal*, *77*(4), 12–35.

LeBaron, B. (2006). Agent-based computational finance. *Handbook of Computational Economics*, *2*, 1187–1233.

Lee, J., Kim, S., & Ho, A. (2024). A survey of large language models in finance. *arXiv preprint arXiv:2401.01234*.

Lewis, P., Perez, E., Piktus, A., Karpukhin, V., Goyal, N., Petrov, S., Sacchi, H., Muhammad, F., Riedel, S., & Lewis, M. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*.

Li, X., Gao, Y., & Li, Z. (2024). Econagent: Llm-powered agent-based macroeconomic modelling [Working Paper].

Li, X., Liu, Z., & Chen, Y. (2023). Fingpt: Democratizing access to financial large language models [arXiv preprint arXiv:2306.06031].

Lopez-Lira, A., & Tang, Y. (2023). Can chatgpt forecast stock price movements? *Working Paper*.

Ma, J., & Shen, D. (2024). Chatgpt and the chinese equity premium. *China Economic Review*, *78*, 102079.

Popov, A., & Roshka, D. (2024). Foliollm: A domain-specific large language model for etf portfolio construction [Working Paper].

Satchell, S., & Scowcroft, A. (2000). A demystification of the black–litterman model. *Journal of Asset Management*, *1*(2), 138–150.

Schick, T., Helwe, Y., LeScao, T., Witteveen, S., Ebert, L., Kashyap, R., Raffel, C., & Liu, P. (2023). Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Sun, Y., Luo, B., & Zhang, H. (2020). Hybrid deep-learning forecasting of exchange rates. *Journal of Forecasting*, *39*(6), 977–996.

Todd, A. (2016). Agent-based models in economics: A review. *Journal of Economic Surveys*, *30*(5), 1232–1255.

Weng, M. (2023). *Practical guide for large language models in the financial industry* (tech. rep.). CFA Institute.

Wu, S., Liu, H., & et al. (2023). Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Yang, M., Hu, X., & Zhao, L. (2023). Large language models in quantitative finance: Challenges and opportunities. *Quantitative Finance Review*, *18*(1), 78–92.

Yao, J., Zhao, X., Chen, S., Lin, W., & Song, D. (2023). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2310.02436*.

Zheng, L., Chen, Y., & Chen, X. (2024). Editing facts in large language models with in-context learning [arXiv preprint arXiv:2402.01234].