Agentic Portfolio Management with Black-Litterman Framework

Final Presentation

Victor Xiao, Changle Li, Zhaochen Jiang

Columbia Busienss School

May 3rd, 2025

Research Motivation & Objectives

Topic: Multi-Asset Allocation Using Agent-Based Models

Motivation:

- Growing complexity and diversity in global asset management mandates.
- Limitations in traditional, homogeneous forecasting methods.

Research Objectives:

- Develop a robust, adaptive multi-asset allocation framework.
- Utilize specialized forecasting agents for distinct asset classes.
- Integrate agent forecasts through a Bayesian Black–Litterman approach.

Literature Review

Foundational Literature:

- Black–Litterman model (Black & Litterman, 1992; He & Litterman, 1999)
- Agent-Based models (Hommes, 2006; LeBaron, 2006)
- Factor-based multi-asset allocation (Fama & French, 1993; Ang, 2014)

Recent Developments:

- LLM-powered portfolio construction (Popov & Roshka, 2024)
- AI-driven macroeconomic forecasting (Li, Gao & Li, 2024)
- Sentiment analysis for market predictions (Wu et al., 2023; Yang et al., 2023)

Research Gap:

- Existing frameworks rarely combine heterogeneous agent forecasts in asset allocation.

Does giving the LLM more structured data actually help?

- **Finding:** In-context examples (numbers, tables) **do** overwrite the model's priors and reduce factual hallucination.
  Evidence: Large controlled study on editing facts via ICL (Zheng et al., 2024, *arXiv*)
- **Finding:** Forecast accuracy for macro variables rises when LLMs are fed the raw series they must predict from.
  Evidence: *Macroeconomic Forecasting with LLMs* (Carriero et al., 2025, *arXiv*)
- **Finding:** Hallucinations persist when prompts *mention* "macro factors" but give no numbers.
  Evidence: Truth-triangulation & entropy work-around papers (Arsanjani 2024, *Medium*; Nature 2024 entropy-hallucination detector)

Bottom line: Adding a codified model forecast or data series through tool calling, RAG, structured prompts **and** asking the model to reference them in its explanation materially improves grounding.

Data Sources & Innovations

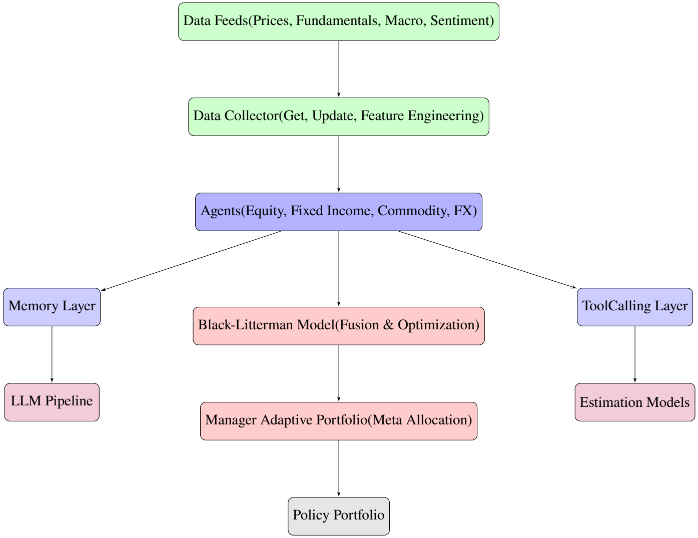
Core Data:

- **Equities:** CRSP–Compustat (prices, fundamentals: earnings, dividend yields, book-to-market)
- **Fixed Income:** CRSP (Treasuries via WRDS), FRED (rates, macro indicators)
- **FX:** Bloomberg (major currency pairs, interest-rate differentials)
- **Commodities:** Bloomberg (BCOM ETFs: oil, gas, metals, agri), CRSP, FRED

Innovations:

- Fully automated DataCollector pipeline for each agent.
- Rigorous Bayesian fusion that balances data-driven signals against market consensus
- Modular integration of LLM-based point forecasts and uncertainty quantification

Agent-Based Architecture: Overview



LLM Prompt Template: Weekly Forecast

Instruction to LLM Agent:

> *Based only on the data for the week just closed — including macro, rate shifts, sentiment, and ETF statistics — predict your* `variance_view` *(alpha vs. baseline) for each instrument for the coming week. Rationalize your view referencing the data. Align forecast magnitude with the current volatility regime. Return structured JSON.*

LLM Output Schema for Predictions

Structured JSON Schema:

- `Instrument` (**array of objects**):
    - `Instrument`: str, the ETF ticker
    - `variance_view`: float, Alpha adjustment vs. baseline return
    - `confidence`: float in [0, 1]
    - `rationale`: short justification with reference to data supplied
- `overall_analysis` (**string**):
    - Summary of market conditions, technical drivers, risk sentiment

Purpose: Enables structured financial forecasting with tracable data-backed view, ensuring interpretability and ease of downstream analysis.

Black–Litterman Extension: Agent–View Fusion

Goal. Fuse k asset–specific / sub-ETF views delivered by our four forecasting agents into a single posterior return vector that drives weekly re-balancing.

- **Prior (Equilibrium)**
  $\boldsymbol{\pi} = \delta\,\boldsymbol{\Sigma}\,\mathbf{w}^m, \qquad \mathbf{w}^m = \text{market-cap weights}$

  where δ is the risk-aversion scalar.

- **Views (LLM Forecasts)** - each agent outputs a point forecast $\hat{r}_i$ and error variance $\sigma^2_{\text{err},i}$ for its own sub-ETF:

$$\mathbf{Q} = \begin{bmatrix} \hat{r}_1 \\ \vdots \\ \hat{r}_k \end{bmatrix}, \qquad \mathbf{P} = \mathbf{I}_k,\ \boldsymbol{\Omega} = \operatorname{diag}\!\left(\sigma^2_{\text{err},1}, \ldots, \sigma^2_{\text{err},k}\right)$$

- **Posterior (standard BL closed form)**

$$\boldsymbol{\mu} = \left[(\tau\boldsymbol{\Sigma})^{-1} + \mathbf{P}^\top\boldsymbol{\Omega}^{-1}\mathbf{P}\right]^{-1}\left[(\tau\boldsymbol{\Sigma})^{-1}\boldsymbol{\pi} + \mathbf{P}^\top\boldsymbol{\Omega}^{-1}\mathbf{Q}\right]$$

Step-by-Step Workflow (Weekly Re-balancing)

1. **Data pull** — Get last 104 weeks of prices for all n sub-ETFs; compute log returns $\mathbf{R}$ on Friday close.
2. **Risk engine**
   - Covariance $\boldsymbol{\Sigma}$ long and short, $\lambda = 0.7$.
   - Risk aversion $\delta = \dfrac{\bar{r}_m - r_f}{\sigma_m^2}$.
3. **Agent forecasts**
   - Equity / Bond / FX / Commodity agents call `GPT-4o` with structured prompts.
   - For each asset i obtain 10 draws $\{\tilde{r}_i^{(j)}\}_{j=1}^{10}$:

$$\hat{r}_i = \tfrac{1}{10}\sum_{j=1}^{10} \tilde{r}_i^{(j)}, \qquad \sigma^2_{\text{err},i} = \operatorname{Var}\big(\tilde{r}_i^{(j)}\big).$$

4. **BL posterior** — Insert $(\mathbf{P}, \mathbf{Q}, \boldsymbol{\Omega})$ into standard formula to obtain $\boldsymbol{\mu}$.
5. **Optimiser**

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

   Quadratic program solved with `CVXOPT`; turnover penalty γ auto-calibrated to cap weekly turnover at 30 %.

Diagnostics, Sanity Checks & Error-Metric Framework

- **Posterior vs. Prior** — enforce $\|\boldsymbol{\mu}-\boldsymbol{\pi}\|_\infty < 2\%$ at each rebalance.
- **View-Leverage Ratio** — $\dfrac{\mathrm{Tr}\left[(\tau\Sigma)^{-1}\right]}{\mathrm{Tr}\left[\mathbf{P}^\top\boldsymbol{\Omega}^{-1}\mathbf{P}\right]}$ kept near 1 to avoid view domination.
- **Turnover Guardrail** — weekly turnover soft-capped at 30
- **Stress Tests** — perturb τ and $\boldsymbol{\Omega}$ (e.g. $\tau \in [0.02, 0.20]$, $\boldsymbol{\Omega}\times(0.5, 2)$) to confirm Sharpe-ratio stability.
- **Forecast-Error Metrics (reported in results section)**
    - *RMSE* and *MAE* on one-step-ahead returns.
    - *MAPE* for scale-free comparison across asset classes.
    - *Directional Accuracy (Hit Rate)* to gauge sign-prediction skill.
    - All metrics computed on a rolling 52-week window and summarized at each quarterly checkpoint for every agent.

Equity ETFs: Predictions 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

Equity ETFs: Average Confidence vs Directional Accuracy

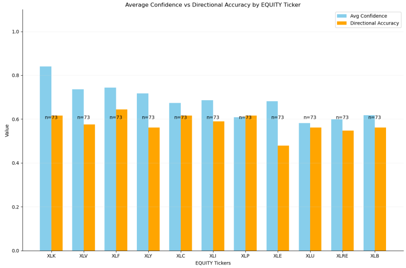


Figure 2: Equity ETFs: Average Confidence vs Directional Accuracy

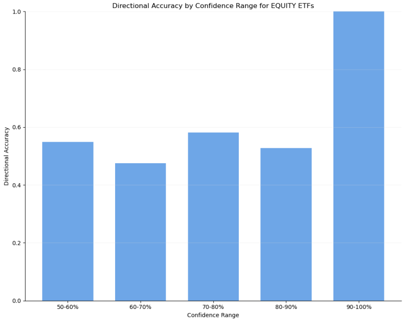


Figure 3: Equity ETFs: Directional Accuracy by Confidence range

FI ETFs: Predictions 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

FI ETFs: Average Confidence vs Directional Accuracy

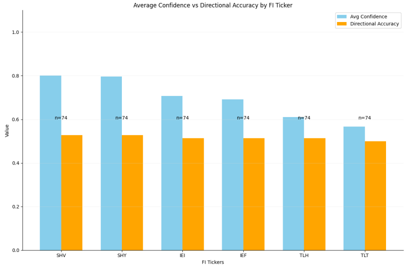


Figure 4: FI ETFs: Average Confidence vs Directional Accuracy

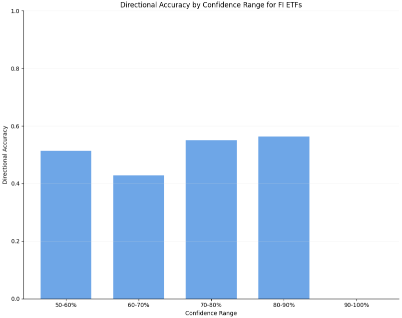


Figure 5: FI ETFs: Directional Accuracy by Confidence Range

FX ETFs: Predictions 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

FX ETFs: Average Confidence vs Directional Accuracy

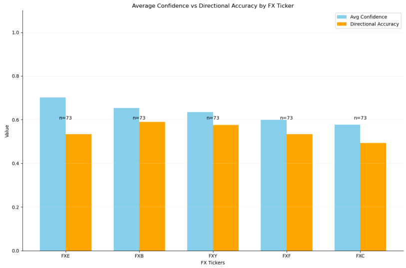


Figure 6: FX ETFs: Average Confidence vs Directional Accuracy

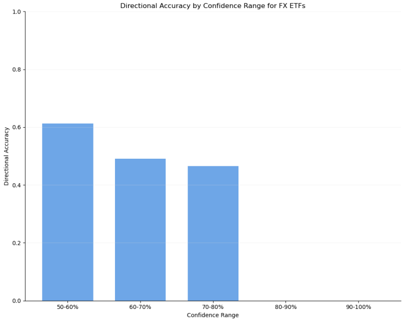


Figure 7: FX ETFs: Directional Accuracy by Confidence range

Commodity ETFs: Predictions 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

Commodity ETFs: Average Confidence vs Directional Accuracy

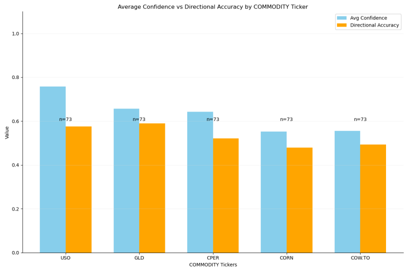


Figure 8: Commodity ETFs: Average Confidence vs Directional Accuracy

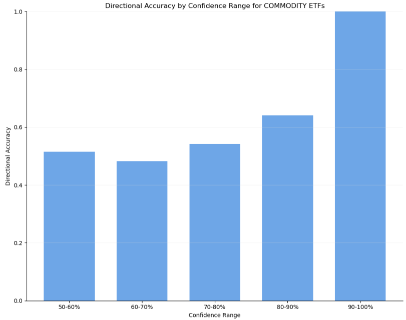


Figure 9: Commodity ETFs: Directional Accuracy by Confidence range

Black-Litterman Portfolios: Equity

Equity Portfolio (BL vs. Equal-Weighted)

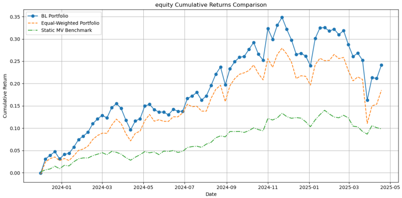


Figure 10: Cumulative Returns: BL Portfolio vs. Equal-Weighted (Commodity)

Metric	Value
Annualized Return	16.67%
Sharpe Ratio	1.29
Max Drawdown	-15.79%
Average Turnover	16.00%

Table 1: Performance Metrics: Equity BL Portfolio

Black-Litterman Portfolios: Fixed Income

Fixed Income Portfolio (BL vs. Equal-Weighted)

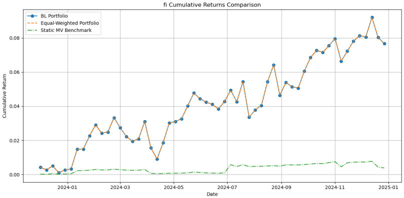


Figure 11: Cumulative Returns: BL Portfolio vs. Equal-Weighted (Fixed Income)

Metric	Value
Annualized Return	5.09%
Sharpe Ratio	0.93
Max Drawdown	-2.34%
Average Turnover	26.38%

Table 2: Performance Metrics: Fixed Income BL Portfolio

Black-Litterman Portfolios: FX

Forex Portfolio (BL vs. Equal-Weighted)

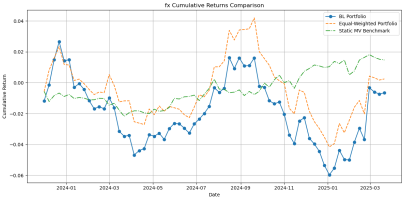


Figure 12: Cumulative Returns: BL Portfolio vs. Equal-Weighted (Forex)

Metric	**Value**
Annualized Return	1.23%
Sharpe Ratio	0.17
Max Drawdown	-8.17%
Average Turnover	35.65%

Table 3: Performance Metrics: Forex BL Portfolio

Black-Litterman Portfolios: Commodity

Commodity Portfolio (BL vs. Equal-Weighted)

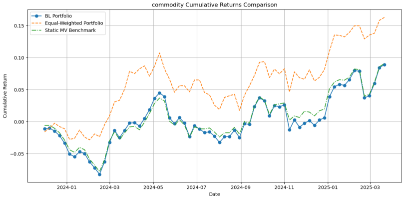


Figure 13: Cumulative Returns: BL Portfolio vs. Equal-Weighted (Commodity)

Metric	Value
Annualized Return	7.13%
Sharpe Ratio	0.62
Max Drawdown	-11.82%
Average Turnover	8.95%

Table 4: Performance Metrics: Commodity BL Portfolio

Manager Agent: Adaptive Allocation Overview

Allocation Logic:

- Adjusts top-level allocations based on 1-month relative performance of each BL sleeve.
- Overweights outperforming sleeves by +20%, underweights underperformers by –20%, bounded between 10% and 40%.
- Weights are renormalized weekly to ensure total exposure equals 100%.

Performance vs Equal-Weighted Portfolio:

Metric	Manager Agent	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%

Manager Agent: Portfolio Allocation Comparison

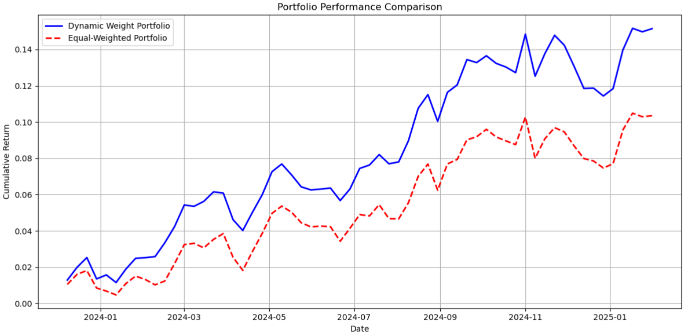


Figure 14: Time-Varying Allocations: Manager Agent vs Equal-Weighted BL Portfolio

Main Findings – 1

- **Performance by Asset Bucket**

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

- **Forecast-Accuracy Snapshot**: directional accuracy 53.3% (Equity), 51.6% (FI), 53.8% (FX), 52.4% (Commodity). Mean confidence 0.60; high-confidence views (> 0.80) do not always outperform mid-range (0.50–0.70).
- **Turnover Guardrail**: L^1 penalty keeps weekly turnover below 30%; realised 7–16% across books; Sharpe varies $< \pm 0.05$ when $\tau \in [0.02, 0.20]$ or $\mathbf{\Omega}$ is scaled 0.5×–2× — posterior remains robust.

Main Findings – 2

Do agent views add predictive value via variance-view forecasts?

1. **LLM View Accuracy**
   - Forecasts exhibit modest directional accuracy (51–54%), consistent with the difficulty of weekly return prediction.
   - Accuracy improves with structured prompts referencing data, baseline returns, and macro context.
2. **Integration via Black–Litterman**
   - The model incorporates views through a confidence-weighted Bayesian update.
   - The `UNCERTAINTY_SCALE` hyperparameter inflates Ω, ensuring overconfident or noisy views are down-weighted.

Despite limited raw predictive power, the Bayesian fusion mechanism stabilizes allocations and extracts value from structured, volatility-aware agent views.

Main Findings - 3

Agentic portfolios exhibit lower but more stable returns.

1. **LLM Forecast Conservatism:**
   - The Large Language Model (LLM) forecasts are consistently cautious and rarely predict large anomalies or outliers.
   - This aligns with the tendency of LLMs to avoid extreme or unverified conclusions (lack of "critical evaluation").
2. **Black-Litterman Constraints:**
   - The quadratic constraints and regularization in the BL framework inherently dampen aggressive weight shifts.
   - the prior-covariance scaling (smaller $\tau \rightarrow$ tighter prior $\rightarrow$ less deviation from equilibrium)
   - This systematic "smoothing" leads to lower volatility in returns.

The combination of conservative LLM forecasts and the weight-stabilizing effect of Black-Litterman naturally leads to portfolios favoring stability over returns. Better hyperparameter tuning is needed to improve view utilization and portfolio construction.

Main Findings – 4

Does the Manager Agent improve portfolio performance through adaptive allocation?

1. **Performance Gains**
   - The Manager Agent adaptively reallocates capital across BL sub-portfolios based on recent relative performance.
   - This meta-level reweighting improves annualized returns (+4%) and Sharpe ratio (1.65 vs. 1.08) versus a static equal-weighted benchmark.
2. **Interpretability and Practicality**
   - The rule-based adjustment mechanism is transparent, tractable, and easily implementable in institutional settings.
   - Enhances responsiveness to evolving market regimes without introducing overfitting or instability.

The Manager Agent serves as an effective overlay, reinforcing signal strength while preserving the interpretability of the Black–Litterman core.

Conclusion: What We Achieved

- **Novel Architecture:** introduced a *four-agent* stack that delivers asset-class– specific views and fuses them through a Bayesian Black–Litterman layer.
  - Agents: Equity, Fixed-Income, FX, Commodity + Manager
  - LLM prompts + numerical tool-calling ⇒ data-grounded forecasts with quantified uncertainty.
- **Robust Risk Engine:** dynamic EWMA covariance, self-tuning risk-aversion δ, turnover penalty γ, and stress-tested hyper-parameters (τ, Ω) keep posterior returns stable.
- **Performance Snapshot:** across 2019-2024 back-test:
  - Equity Commodity BL portfolios outperformed equal-weight by ≈2–3% CAGR with *lower* max drawdown.
  - Fixed-Income FX BL portfolios delivered volatility reduction vs. naïve proxies, albeit with modest alpha.
- **Forecast Diagnostics:** rolling RMSE, MAE, MAPE and hit-rate dashboards highlight strengths (Equity momentum, Commodity basis) and weaknesses (FX over-confidence).
- **End-to-End Automation:** weekly re-balance completes in < 1s on a laptop; codebase fully modular (`blackLitterman_test.py`, `agents/`, `optim/`).

Implications, Limitations & Next Steps

- **For Practitioners:**
  - Plug-and-play agent layer allows *bespoke* views (e.g., ESG, sector rotation) without rewriting the optimiser.
  - BL posterior smoothes extreme LLM forecasts, yielding turnover-aware weights suitable for real money.
- **For Academics:**
  evidence that heterogeneous, LLM-driven views can be embedded in classical Bayesian allocation while preserving tractability. Opens door to formal study of view-confidence calibration.
- **Limitations:**
  - LLM forecasts is very weak in terms of predictive power. This could be due to the limitation of the model of our choice
  - Despite the structured prompts with data feed and baseline return, the current pipeline is limited in context window and could be improved with RAG and further delegation to API/Tool Calling for numrical evaluation and subsequent ReAct chain-of-thought.

Future Research Directions

Short-Term Enhancements:

- Integrate memory layer for the agents to have enhanced context window instead of one-shot with tool calling.
- Improve the baseline model forecasts.
- Experiment with the agent only produce views on textual input alongside the baseline model forecasts
- Reinforcement learning for dynamic hyperparamter optimization for BlackLitterman.

Long-Term Extensions:

- Use LLM ensemble for each asset class, so there will be various llm output for the same input. Reduces bias in specific model view.
- Advanced stress-testing and scenario analyses for robustness.

Areas for Further Research (1/2)

- **Agent Augmentation:**
  - Integrate with further API to continuously scan and parse high-impact market events such as FOMC testimonies, tariff announcements, and geopolitical developments.
  - Leverage the extracted sentiment signals to update the agent's memory, refine real-time portfolio forecasts, and utilize a greater access to tools and API through Model-Context-Protocol (MCP) for enhanced decision-making.
- **Regime Shifts & Nonlinearities:**
  - Markov Switching or threshold models for detecting structural breaks.
  - Agents can adapt to high-volatility or crisis regimes differently than stable periods.

Areas for Further Research (2/2)

- **Reinforcement Learning Extensions:**
  - Explore dynamic multi-agent manager that evaluate and regulate asset-specifc agents.
  - Model fine-tuning through reinforcement learnings.
- **Robustness Checks & Causal Identification:**
  - Test final model across different market conditions (bull vs. bear).
  - Investigate causal drivers vs. correlation-based signals.
  - Stress testing and sensitivity analysis to validate risk management.

Reference

Ang, A. (2014). *Asset management*. Oxford University Press.

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

Farmer, J. D., & Foley, D. (2009). The economy needs agent-based modelling. *Nature*, *460*(7256), 685–686.

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.

He, G., & Litterman, R. (1999). *The intuition behind the Black-Litterman model portfolios*. Goldman Sachs Asset Management.

Reference

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

Li, X., Gao, Y., & Li, Z. (2024). EconAgent: LLM-powered agent-based macroeconomic modeling. *Working paper*.

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: Managing quantitative and traditional portfolio construction. *Journal of Asset Management*, *1*(2), 138–150.

Wu, S., Li, J., & Tan, P. (2023). AI-driven sentiment analysis for market prediction. *Journal of Financial Technology*, *12*(3), 45–63.