Computational Global Macro with AI for Risk and Portfolio Management

Marius Ciepluch
marius.ciepluch@gmail.com
This version: July 2026

Abstract

This paper evaluates AI inference-guided macro-factor awareness for portfolio optimization, a new option that comes from advances in LLMs. Using LLMs for quantitative finance is promising for some use cases but needs specific features for simulation-based testing. The architecture must provide Point in Time (PiT) inference to remove model look-ahead bias. To achieve that, model inputs are anonymized, timestamp-free, dated strictly before each rebalance, and discounted by a Memorization Confidence Score (MCS). It must also allow macro-factor de-risking: inflation velocity, growth expectations, and credit stress are z-scored and generate bounded Black–Litterman views that tilt a Hierarchical Risk Parity (HRP) and Conditional Value-at-Risk (CVaR) baseline, followed by a Sparse Jump Model (SJM) regime overlay and a macro-stress crisis gate. On the common 2016 to 2026 walk-forward the SJM de-risk line gives up Compound Annual Growth Rate (CAGR), from 15.4% to 9.4%, but lowers annualized volatility from 12.9% to 7.2% and maximum drawdown from -19.6% to -8.2%. The Sharpe ratio remains the same, while the appraisal ratio rises from 0.81 to 1.06. The results indicate that AI-inferred macro-factors can improve returns per unit of risk, and calm an otherwise volatile tech portfolio.

Keywords: Point in Time inference, look-ahead bias, macro factors, Black–Litterman, Hierarchical Risk Parity, CVaR, regime switching, portfolio de-risking, large language models, tear sheet metrics

JEL classification: G11, G17, E44, C45, C58

1. Introduction

Factor investing rests on a definition broad enough to carry macro variables: “any investment characteristic relating a group of securities that is important in explaining their risk and return behavior” (Bender et al., 2013, p. 2). Fama and French (Fama & French, 1993) marked a seminal moment for factor research, even though their model is built to explain the returns of stocks and bonds rather than a macro-overlay. Today the commonly known core four equity factors are value, momentum, quality, and volatility. Macro factors can be added as overlays to de-risk a portfolio that is neither currency-hedged nor market-neutral. Equity factors often depend on the market regime, especially for the technology sector, which is subject to changing funding activity. Consumer Price Index (CPI) year-over-year inflation, the 10-year minus 2-year Treasury spread, and the high-yield option-adjusted spread stand for inflation velocity, growth expectations, and credit stress. The question is twofold: Can macro signals de-risk a technology-focused (ETF) portfolio? And how useful are LLMs for this task? To approach this, a custom architecture using PiT inference (via open-weight models) has been developed. The research objective is optimizing return per unit of risk, not maximum raw return or direct price forecasting (predictive alpha).

2. Two pillars of the hybrid architecture

The architecture has two pillars. Inference engineering, the first pillar, controls what information reaches the portfolio optimization framework relative to the simulated decision time (walk-forward). It also measures whether the AI model looks up information (for a look-ahead) in its weights. The portfolio construction system, the second pillar, gives the model a bounded economic task: infer the macro risk and the level of control needed. Figure 1 reads top to bottom with the control layer marked in dashed red. The key technical requirements are “zero look-ahead” and “no recall”.

At each monthly rebalance the system (via Riskfolio-Lib (Cajas, 2026)) computes an HRP portfolio under a CVaR risk measure and scales the risk sleeves (excluding the 25% in the T-bill sleeve (BIL)). An open-weight language model (openai/gpt-oss-20b), selected on a measured holdout MCS area under the receiver operating characteristic curve (MCS-AUC) of 0.926, reads three anonymized rolling z-scores and returns five bounded regime loadings: inflation, growth, credit stress, policy, and risk appetite. A fixed exposure map converts these loadings into absolute single-asset Black–Litterman views (Litterman & Black, 1991). View magnitude is scaled by conviction and, where a score is available, by (1 − MCS). The long-only Black–Litterman portfolio is blended 30% against its 70% HRP base, with the per-asset tilt capped at 0.5. A Sparse Jump Model (SJM) regime overlay then labels the market state bull, bear, or neutral from Exponentially Weighted Moving Average (EWMA) log returns and caps the Black–Litterman posterior (Shu & Mulvey, 2024). A macro-stress crisis gate overrides the allocation when several regime indicators flag extreme conditions at once.

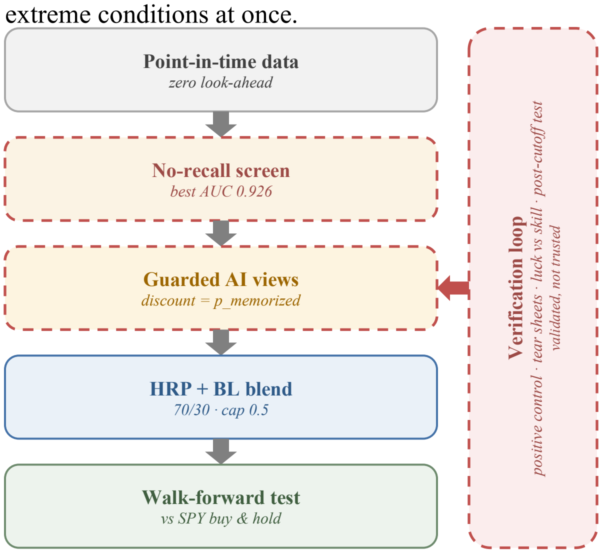

Figure 1. Architecture of the hybrid allocation model that allows controlled AI macro-factor tilts

3. Look-ahead bias and PiT inference

The main adoption problem for AI in historical portfolio optimization and simulation is *contamination*. "Has the model seen the test set?" is the LLM version of "was the backtest in-sample?" All prices, returns, and macro rows supplied to the allocation function are dated strictly before the rebalance. Ticker identities, calendar dates, and raw price levels are withheld; the model receives a normalized numeric macro state. This way the AI model inference is set up to spot patterns an analyst would overlook or not follow due to human bias.

Technically this architecture prevents direct use of future-dated rows. The process is PiT inference: the model must reason from the information state available at the simulated decision point; the rebalance cycle is a separate optimization concern outside this step. Every screened (open-weight) model shows statistically significant recall of the pre-cutoff macro data history. openai/gpt-oss-20b was selected deliberately, to be run recall-guarded rather than presumed clean. The MCS estimates the probability that a response is memorized (p_memorized) from the token log-probabilities. Where an MCS is available, view magnitude is multiplied by (1 − MCS); weak or unavailable scores pass the raw exposure through. Responses judged to come from the weights are discarded outright; only open-weight models expose the required token log-probabilities at this time. Model response parse and Black–Litterman failures fall back to HRP-CVaR. recall_guard implements ideas from MemGuard-Alpha and Look-Ahead-Bench to approach the LLM model PiT inference problem (Benhenda, 2026; Roy, 2026). Model responses are sampled rather than deterministic; recall_guard therefore ensembles per-draw scores into a single p_memorized with an interval, keeping the memorization signal distinct from sampling noise. The control reduces and measures recall risk. It does not prove that look-ahead bias has been removed.

4. Economic tailwinds and inherent performance

A high CAGR can come from favorable economic conditions alone, that is, from equity beta. No realized metric fully isolates inherent portfolio performance bias, so the metrics are read together and in a fixed order (Paleologo, 2021, Chapter 7): risk is described first, performance is then expressed per unit of that risk, and the result is tested for significance. CAGR records the return achieved; annualized volatility, Value-at-Risk (VaR), Conditional Value-at-Risk (CVaR), and maximum drawdown describe the risk carried; the Sharpe and Calmar ratios compress return per unit of that risk. Beta and correlation to the S&P 500 show how much market dependence remains, the Information Ratio (IR) separates benchmark-relative skill from plain risk reduction, and the appraisal ratio ($\alpha/\sigma\varepsilon$) prices alpha per unit of idiosyncratic risk. The Sharpe Stability Ratio (SSR) is read as temporal consistency within each realized line, not as proof that one strategy adds stable skill (Traver & Rodríguez Domínguez, 2026).

The simulation uses a macro sensitive portfolio consisting of the MSCI World ETF, a tech-sector ETF (State Street Technology Select Sector), a Gold ETF and 1-3 months T-Bill UCITS ETF. Buy and hold is simulated with equal weights (25% allocation per asset), the portfolio optimizations adjust the weights monthly.

5. Tear sheet metrics

Table 1 reports the realized tear sheet in three blocks: return against risk, dependence on the S&P 500, and factor risk. Five figures carry the headline argument – CAGR, annualized volatility, Sharpe, maximum drawdown, and SSR – while the appraisal ratio ($\alpha/\sigma\varepsilon$) and the IR decide whether the remaining performance is scalable skill or only a smaller share of the market.

Read in conjunction with figures 2 and 3, the SJM variant de-risks the portfolio rather than buying return with more of it. It carries the lowest annualized volatility, the smallest maximum drawdown at -8.2% against -19.6% for buy-and-hold, a Sharpe in line with the other two lines, and the highest appraisal ratio at 1.06 against 0.81. The benchmark block qualifies that reading. The IR is negative for all three lines and most negative for the two macro-aware ones, -0.33 and -0.41: against the S&P 500 they give up active return at a higher tracking error, so the gain is risk reduction rather than benchmark-relative skill. R^2 against the market falls from 0.71 to 0.13, which is the same result seen from the factor side. SSR sits at 0.10 to 0.13 across all three lines, which reads as steady rather than a lucky stretch.

The ETF portfolio acts as a proxy that isolates the simulation conditions: a small but diversified exposure to the USD. Equity factors can effectively be ignored here, which keeps the focus on macro-factors.

6. Results

Buy-and-hold finishes with the highest realized return. The AI macro-factor and its SJM variant de-risk lines accept lower CAGR in exchange for lower loss scale and market dependence.

On the common 2016 to 2026 window, buy-and-hold produces 15.4% CAGR at 12.9% annualized volatility, with a -19.6% maximum drawdown. The AI macro-factor line gives up return but reduces risk: 10.9% CAGR, 8.4% volatility, 1.05 Sharpe, and a -11.4% drawdown. The SJM de-risk line moves further in the same direction: 9.4% CAGR, 7.2% volatility, 1.03 Sharpe, and a -8.2% drawdown. Calmar improves from 0.79 to 1.15, beta falls from 0.60 to 0.14, and correlation to the S&P 500 falls from 0.84 to 0.36. Lower loss scale does not remove tail risk: SJM de-risk skew is -0.75 and excess kurtosis is 33.85, against -0.11 and 9.60 for buy-and-hold. The ETF universe was selected in-sample, the overlay holdout was not blind, and costs, taxes, capacity, and live execution are not modeled.

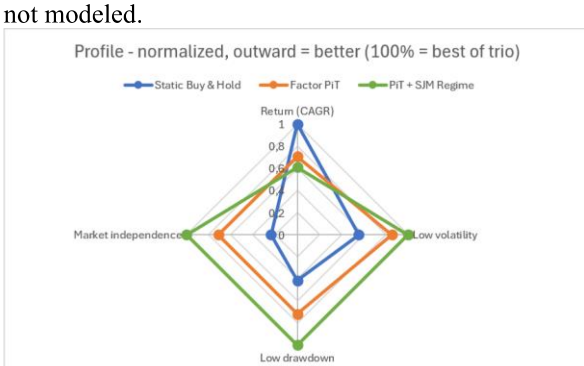

Figure 2. Normalized metric profile, outward = better (100% = best of the three lines). The SJM de-risk line leads on low volatility, low drawdown, and market independence, while buy-and-hold leads on return.

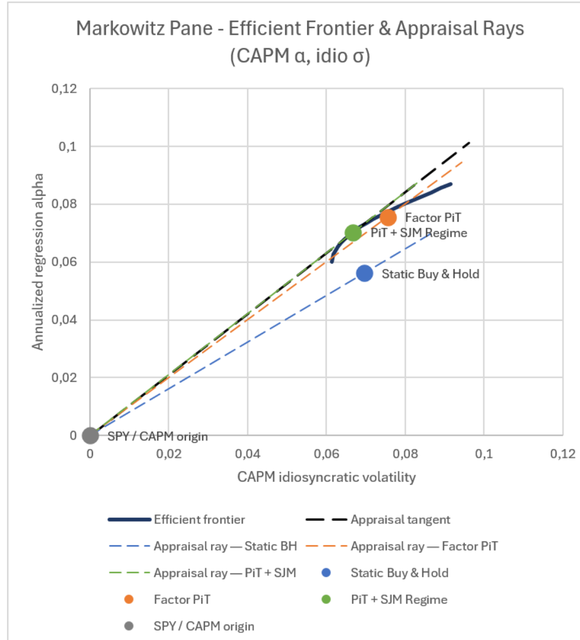

Figure 3. Markowitz pane with efficient frontier and appraisal rays (CAPM α against idiosyncratic volatility). The macro-aware lines lie on steeper appraisal rays than static buy-and-hold, with the appraisal ratio rising from 0.81 to 1.01 and 1.06.

7. Conclusion and outlook

In this simulation, PiT-inferred macro-factor tilts reduce volatility, drawdown, and market dependence. This construction is deliberate: it is optimized for the target objective of smoother returns and scalable risk vs. return. In conjunction with contemporary challenges of adopting AI for finance decisions, the hybrid allocation architecture serves as a blueprint on how to integrate AI systems while remaining in control and aware of the limitations.

The 4-ETF portfolio is not designed to provide excess returns over the S&P 500, or to cater towards Tiger Cub style investment scenarios. The hybrid allocation architecture provides a controlled research foundation, not proof of 100% PiT inference or predictive-alpha forecasting via LLMs.

The wider application matters because leveraged derivatives, including leveraged ETFs, are readily accessible. Leverage magnifies gains and path-dependent losses, making raw-return ranking especially misleading. Return per unit of risk is therefore central to choosing exposure, leverage, and de-risking rules.

Online LLMs are not required to include inferred macro-factor awareness for monthly (or more frequent) portfolio rebalances. Future research can emphasize factor mining, AI-assisted extrapolation of factors or rules, for example via FURIA or RIDRA approaches (Mollá et al., 2022).

8. Disclaimer

Parts of the research have been simulated, tested and documented using AI models from Anthropic (Opus and Fable).

9. References

Bender, J., Briand, R., Melas, D., & Subramanian, R. A. (2013). *Foundations of Factor Investing*.

Benhenda, M. (2026). *Look-Ahead-Bench: a Standardized Benchmark of Look-ahead Bias in Point-in-Time LLMs for Finance*. http://arxiv.org/abs/2601.13770

Cajas, D. (2026). *Riskfolio-Lib (7.3)*. https://github.com/dcajasn/Riskfolio-Lib

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

Litterman, R., & Black, F. (1991). *Asset Allocation: Combining Investor Views with Market Equilibrium*. www.iijournals.com

Mollá, N., Rabasa, A., Rodríguez-Sala, J. J., Sánchez-Soriano, J., & Ferrándiz, A. (2022). Incremental decision rules algorithm: A probabilistic and dynamic approach to decisional data stream problems. *Mathematics*, *10*(1). https://doi.org/10.3390/math10010016

Paleologo, G. A. (2021). *Advanced Portfolio Management: A Quant's Guide for Fundamental Investors*. Wiley.

Roy, A. (2026). *MemGuard-Alpha: Detecting and Filtering Memorization-Contaminated Signals in LLM-Based Financial Forecasting via Membership Inference and Cross-Model Disagreement*.

Shu, Y., & Mulvey, J. M. (2024). *Dynamic Factor Allocation Leveraging Regime-Switching Signals*. http://arxiv.org/abs/2410.14841

Traver, M. B., & Rodríguez Domínguez, A. (2026). *The Sharpe Stability Ratio Temporal Consistency of Risk-Adjusted Performance*. www.barclayhedge.com.

Table 1. Realized tear sheet on the common 2016 to 2026 walk-forward window (2,468 trading days). Return-vs-risk rows are annualized on the 252 trading-day basis; benchmark and factor-risk rows are as computed in the simulation on the VectorBT 365-day basis.

Metric	Buy-and-hold	AI macro-factor	SJM de-risk
Return vs. risk			
CAGR	15.4%	10.9%	9.4%
Ann. volatility	12.9%	8.4%	7.2%
Sharpe	1.04	1.05	1.03
Max drawdown	-19.6%	-11.4%	-8.2%
Calmar	0.79	0.96	1.15
SSR	0.13	0.12	0.10
Benchmark vs. S&P 500			
Correlation to S&P 500	0.84	0.43	0.36
Beta	0.60	0.20	0.14
Active return (ann.)	-0.8%	-5.4%	-6.9%
Tracking error (ann.)	10.0%	16.2%	16.8%
Information Ratio	-0.08	-0.33	-0.41
Factor risk and performance			
R^2 (CAPM, 1-factor)	0.71	0.19	0.13
Idio vol $\sigma\varepsilon$ (ann., CAPM)	7.0%	7.6%	6.7%
α (regression, ann.)	5.6%	7.6%	7.1%
Appraisal ratio ($\alpha/\sigma\varepsilon$)	0.81	1.01	1.06
R^2 (4-ETF basket)	0.92	0.81	0.75
Residual vol basket (ann.)	5.0%	4.8%	4.9%