arXiv:submit/8190614 [cs.AI] 7 Oct 2026

SOTA: STOCK OPTIONS TRADING AGENTS GUIDED BY OPTION-IMPLIED RETURN DISTRIBUTIONS

Yizhen Xie*
Carnegie Mellon University

Mengyang Liu†
Amazon

ABSTRACT

As option markets grow and AI advances, agentic systems for option trading are gaining increasing attention. Language-model-based agents can reason over contextual information such as news, but option trading presents a particularly challenging decision problem: a single stock can have thousands of contracts, and the agent must decide both which contracts to trade and how to combine them. Existing approaches often sidestep this complexity by restricting the policy to a fixed strategy structure, such as a straddle, limiting their ability to switch strategies as market conditions change. We present SOTA (Stock Options Trading Agents), an agentic trading framework for structured option-strategy selection. SOTA abstracts the large option universe into strategy-level decisions while deterministic resolvers handle portfolio implementation. We develop SOTA by post-training Qwen3.8-27B with supervised fine-tuning followed by reinforcement learning. SOTA is evaluated on options on nine large-cap U.S. equities and SPY against rule-based and machine-learning strategy selectors in the same trading environment. Over a six-month out-of-sample period, SOTA earns an 18.3% total return with a Sharpe ratio of 1.60 and a maximum drawdown of 8.96%. We also document an asymmetric role of news: news improves frontier-teacher trajectories, but retaining news during reinforcement learning reduces out-of-sample return from 18.3% to −2.7%.

1 INTRODUCTION

Option trading has expanded rapidly in recent years (Bryzgalova et al., 2023), creating a natural setting for increasingly capable trading agents. Recent language-model trading systems formulate financial decision-making as sequential interaction with a market environment, where an agent observes market information and portfolio state and chooses actions such as buy, sell, or hold (Yang et al., 2023a; Yu et al., 2025; Zhang et al., 2024; Xiao et al., 2024). Extending this formulation to options is difficult because the decision space is much larger and changes over time.

For a single underlying, an agent may face thousands of contracts that differ in strike, maturity, and option type. Existing approaches reduce this complexity either by predicting returns for individual contracts (Bali et al., 2023) or by restricting the trading problem to a single strategy class, such as volatility trading (Sheu & Wei, 2011; Chen et al., 2025b). These formulations are useful, but they do not directly support dynamic selection across heterogeneous payoff structures.

Reinforcement learning (RL) is well suited to sequential trading problems with transaction costs, discrete rebalancing, and portfolio-state dependence (Sutton & Barto, 2018; Mnih et al., 2015; Bühler et al., 2019; Chen et al., 2025b). It has been applied to portfolio management, stock trading, and high-frequency trading (Jiang & Liang, 2017; Yang et al., 2020; Briola et al., 2021; Qin et al., 2024; Zong et al., 2024). In options, however, learning-based methods have focused primarily on hedging (Bühler et al., 2019) or on learning positions within a fixed volatility structure such as an at-the-money straddle (Tan et al., 2024; Chen et al., 2025b). Selecting among heterogeneous option strategies remains challenging.

* Correspondence to: yizhenxi@andrew.cmu.edu.
† Work conducted outside the author's role at Amazon.

In this paper, we present **SOTA (Stock Options Trading Agents)**, an agentic trading framework for structured option-strategy selection. SOTA separates strategy-level reasoning from contract-level implementation: the language model selects among nine economically meaningful strategy families and their parameters, while deterministic resolvers map each decision into exact contracts, position sizes, and hedges. This structured strategy-selection interface reduces the decision space without restricting the agent to a single payoff structure. The policy can also combine structured market information with contextual signals such as news before producing executable option positions.

The primary contributions of our work are:

- We formulate option trading as structured strategy selection rather than contract-level choice. The agent selects among nine strategy families and their parameters, while deterministic resolvers map these decisions into contracts, position sizes, and hedges. This reduces a large contract-level decision problem to a compact strategy space spanning directional, volatility, skew, and curvature trades.
- We construct a point-in-time trading environment whose state summarizes the option-implied return distribution, and train Qwen3.8-27B using supervised fine-tuning on anonymized frontier-model trajectories followed by reinforcement learning on portfolio returns net of transaction costs.
- SOTA earns an 18.3% total return with a Sharpe ratio of 1.60 over a six-month out-of-sample period on options on nine large-cap U.S. stocks and SPY, while all rule-based and machine-learning baselines generate negative returns. Ablations show that reinforcement learning improves on supervised fine-tuning alone, while news helps the frontier teacher but degrades performance when retained during reinforcement learning.

2 Related Work

2.1 Agentic Trading

Agentic trading systems treat trading as repeated interaction with a market: at each step the agent observes prices and its portfolio, takes an action, and receive feedback through profit and loss. Early systems learn this policy with reinforcement learning. They have been applied to trading stocks (Yang et al., 2020), cryptocurrency (Jiang & Liang, 2017; Qin et al., 2024; Zong et al., 2024), and high frequency trading using limit-order-book data (Briola et al., 2021), and shared environments such as FinRL-Meta make these settings easy to benchmark (Liu et al., 2022). These approaches operate primarily on numerical market data.

Language-model agents add the ability to process news, filings, and other text. Financial language models adapt general-purpose models to financial text (Yang et al., 2023a;b). Trading agents built on such models add memory of past events (Yu et al., 2025), multimodal inputs and tool use (Zhang et al., 2024), or multi-agent roles such as analyst, researcher, trader, and risk manager (Yu et al., 2024; Xiao et al., 2024), and agent platforms and market simulators support these designs (Yang et al., 2024; Zhang et al., 2026). When trading on their own, the performance of language-model agents remains mixed: on real stock data, StockBench finds that most language-model agents struggle to outperform a simple buy-and-hold baseline (Chen et al., 2025a).

In both lines of work, the action space is small and fixed: buy, sell, or hold a single stock or cryptocurrency, or choose portfolio weights over a fixed set of assets. Options do not fit this mold. Each underlying may have thousands of contracts, the available set changes as contracts expire and new ones are listed, and a position often combines several contracts. SOTA keeps the agentic loop but changes the action: the agent selects an option strategy and its parameters, and deterministic tools turn that decision into contracts, position sizes, and hedges.

2.2 Option-Implied Return Distributions

Option prices encode the market's distribution of future returns (Breeden & Litzenberger, 1978). Implied volatilities across strikes and maturities reveal the volatility, skewness, and kurtosis of this distribution (Corrado & Su, 1996; Dennis & Mayhew, 2002; Bakshi et al., 2003; Dumas et al., 1998; Cont & da Fonseca, 2002; Zhang & Xiang, 2008). These option-implied measures are informative

about future outcomes. Implied volatility predicts realized volatility (Christensen & Prabhala, 1998), option trading volume carries information about future stock prices and volatility (Easley et al., 1998; Pan & Poteshman, 2006; Ni et al., 2008), and volatility-related characteristics predict the cross-section of option returns (Goyal & Saretto, 2009; Cao & Han, 2013), a relation that machine-learning models exploit at scale (Bali et al., 2023). Our market state is built from these measures.

2.3 Option Trading Strategies

Option strategies combine contracts to create different payoff exposures. A second-order approximation of option value gives

$$dV \approx \Delta\, dS + \frac{1}{2}\Gamma(dS)^2 + \nu\, d\sigma + \Theta\, dt.$$

Here, delta measures directional exposure, gamma reflects convexity with respect to the underlying price, theta represents time decay, and vega measures sensitivity to implied volatility. Appendix A gives their Black–Scholes–Merton formulas (Black & Scholes, 1973; Merton, 1973). These sensitivities motivate four broad classes of trading strategies in our framework: directional, volatility, skewness, and curvature strategies.

Directional strategies. Directional strategies express a view on whether the underlying asset will rise or fall. Their payoffs are driven mainly by net delta exposure, while different option structures allow the same directional view to be expressed with different degrees of leverage, asymmetry, and downside protection.

Volatility strategies. Volatility strategies express a view on how much the underlying asset will move, rather than on the direction of the move. Long-volatility positions benefit from large price movements, while short-volatility positions benefit when price movements remain limited. Their performance depends on the trade-off between convexity, time decay, and changes in option prices (Bakshi & Kapadia, 2003; Carr & Wu, 2009; Goyal & Saretto, 2009).

Skewness strategies. Skewness strategies express a view on whether downside and upside risks are priced differently. They compare the relative value of put and call options away from the current stock price and can profit when this asymmetry changes (Bakshi et al., 2003; Gârleanu et al., 2009; Kozhan et al., 2013).

Curvature strategies. Curvature strategies express a view on how option prices differ between contracts near the current stock price and contracts further away from it. They profit from changes in the relative pricing of central and more extreme strike prices (Breeden & Litzenberger, 1978; Zhang & Xiang, 2008).

SOTA performs structured strategy selection across nine strategy families that span these four categories. Appendix B illustrates representative payoff structures for each family.

3 SOTA: Structured Strategy Selection for Stock Option Trading Agents

Option trading presents a large and dynamically changing decision problem: for a single underlying, an agent may face thousands of contracts that differ in strike, maturity, and option type. Direct contract-level selection is therefore difficult in a time-sensitive trading setting.

SOTA formulates this problem as structured strategy selection. The language model chooses an option strategy and its parameters, while deterministic resolvers handle contract selection, position sizing, hedging, and portfolio implementation. This separation reduces the contract-level decision problem while preserving flexibility across heterogeneous payoff structures. Figure 1 summarizes the framework.

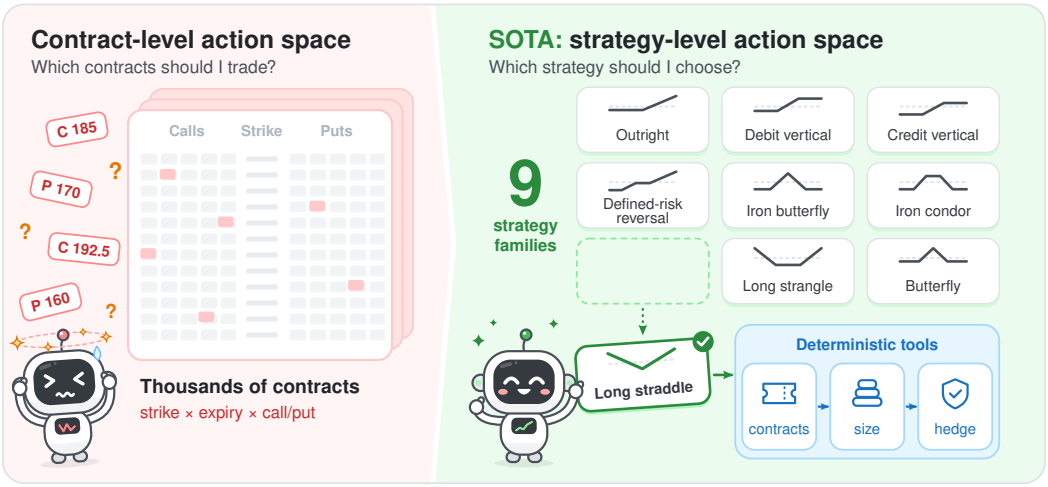


Figure 1: From contract-level to strategy-level decisions. Left: at the contract level, an agent must choose among thousands of listed contracts that differ in strike, expiry, and call/put type. Right: SOTA instead has the agent select one of nine option-strategy families and its parameters (here, a long straddle), and deterministic tools resolve that choice into exact contracts, a position size, and hedge trades.

3.1 Problem Formulation

Trading environment. The agent trades options on a set of underlyings $\mathcal{U}$ over decision dates $t = 0, 1, \ldots, T$. On date t, underlying $u \in \mathcal{U}$ has price S_t^u and a listed option chain $\mathcal{C}_t^u$. Each contract $c \in \mathcal{C}_t^u$ is defined by its strike, expiration date, and type (call or put), and has price p_t^c per contract, taken at the bid–ask midpoint. A chain holds thousands of contracts with different nonlinear exposures, and it changes over time as contracts expire and new ones are listed.

The portfolio entering date t holds q_t^c contracts of each option c, h_t^u shares of each underlying for delta hedging, and cash B_t. Its marked-to-market value is

$$V_t = \underbrace{\sum_{u \in \mathcal{U}} \sum_{c \in \mathcal{C}_t^u} q_t^c \, p_t^c}_{V_t^{\text{options}}} + \underbrace{\sum_{u \in \mathcal{U}} h_t^u S_t^u}_{V_t^{\text{underlying}}} + \underbrace{B_t}_{V_t^{\text{cash}}}. \tag{1}$$

Trading is sequential. At each date t, the agent observes the market and its portfolio and issues a trading decision a_t. The environment converts a_t into trades Δq_t^c and Δh_t^u, executes them at midpoint prices, and charges fees and commissions TC_t. Holdings and cash after trading are

$$q_t^{c,+} = q_t^c + \Delta q_t^c, \qquad h_t^{u,+} = h_t^u + \Delta h_t^u, \qquad B_t^+ = B_t - \sum_{u \in \mathcal{U}} \Big(\sum_{c \in \mathcal{C}_t^u} \Delta q_t^c \, p_t^c + \Delta h_t^u S_t^u \Big) - \mathrm{TC}_t. \tag{2}$$

These holdings carry over to date $t+1$, where they are marked at the new prices and expiring contracts are settled. The change from V_t to V_{t+1} therefore reflects option price moves, hedge positions, transaction costs, and settlements.

Choosing Δq_t^c directly over every listed contract creates a large, dynamically changing decision space. We instead represent each decision a_t as a standardized option combination from an economically meaningful strategy family, and deterministic resolvers translate it into contract-level trades.

Markov decision process. We formulate option-strategy selection as a Markov decision process,

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle, \tag{3}$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ the structured strategy space, $\mathcal{T}$ the transition mechanism, $\mathcal{R}$ the reward function, and γ the discount factor.

State. At each trading date t, the agent observes a point-in-time state

$$s_t = (\mathbf{M}_t, \mathbf{N}_t, \mathbf{P}_t, \mathbf{A}_t)\,, \tag{4}$$

where $\mathbf{M}_t$ contains market features, $\mathbf{N}_t$ point-in-time textual summaries of news and filings, $\mathbf{P}_t$ open option positions, and $\mathbf{A}_t$ portfolio-level information. Portfolio state allows each decision to condition on exposures and outcomes inherited from earlier trades.

We construct the market state from variables identified by the option-pricing and option-return literature as relevant to variation in option returns (Goyal & Saretto, 2009; Cao & Han, 2013; Bakshi et al., 2003; Pan & Poteshman, 2006). For each underlying, the state contains 10 features: underlying return, realized volatility, implied-volatility level and change, the implied-minus-realized volatility wedge, term structure, skewness, curvature, option flow, and changes in open interest. Appendix C.1 gives precise definitions.

Structured strategy selection. Instead of choosing directly among listed contracts, the policy selects structured trading strategies. At each decision time, the policy outputs a set of actions: open a new position (a_t^{open}), close an existing position (a_t^{close}), roll an existing position to a later expiry (a_t^{roll}), or leave the portfolio unchanged (a_t^{hold}).

An action a_t is specified as $(u, f, \omega, \theta, \kappa)$, where u is the underlying, f the strategy family, ω the long/short orientation, θ the tenor bucket, and κ the strategy's coordinates on the delta grid.

To avoid querying prices for the full option chain, the agent first proposes up to six candidate strategies for each underlying. We resolve and price only these candidate packages before the final trading decision. This two-stage mechanism preserves price-aware strategy selection without placing thousands of contracts in the model context.

The learned policy therefore focuses on strategy selection, while deterministic resolvers handle contract selection, portfolio sizing, and delta hedging. Appendix E reports the rules implemented by the latter.

Transition. The transition kernel $\mathcal{T}(s_{t+1} \mid s_t, a_t)$ captures changes in market state, news and filings, existing positions and their risk exposures, and portfolio value net of transaction costs, with holdings and cash evolving as in Equation 2.

Reward. The one-step reward is the change in log portfolio value,

$$r_t = \log V_{t+1} - \log V_t. \tag{5}$$

We train the policy in two stages. Supervised fine-tuning on frontier-model trading trajectories provides an initial strategy-selection policy, after which GRPO (Shao et al., 2024) optimizes the policy for portfolio performance. Appendix F describes the full procedure, and Appendix G.3 reports the hyperparameters.

4 Experiments

4.1 Experiment Setup

Datasets. We evaluate SOTA on 10 actively traded underlyings: SPY and nine large-capitalization U.S. equities (AAPL, AMZN, GOOGL, META, MSFT, MU, NVDA, PLTR, and TSLA). The option data include OPRA trades and quotes and OptionMetrics. Underlying stock prices are CRSP daily open and close prices for the nine equities and SpiderRock underlying marks for SPY. News comes from Massive.com's news API, and regulatory filings come from SEC EDGAR. Table 1 reports the chronological sample splits.

Table 1: Chronological data splits.

Phase	Dates	Role
Supervised fine-tuning	2024-09-03 to 2024-11-29	Frontier-model generation and supervised fine-tuning
Reinforcement learning	2024-12-02 to 2025-02-28	Reinforcement learning to optimize the policy
Test	2025-03-03 to 2025-08-29	Out-of-sample evaluation of the policy and baselines

Table 2: Baseline comparison for option-strategy selection policies

	Profit	Risk-adjusted profit			Risk		Trade	
Policy	TR(%)↑	ASR↑	ACR↑	ASoR↑	AVOL(%)↓	MDD(%)↓	WR(%)↑	PLR↑
Rule-based policies								
Equal-weighted rule	−5.22	−13.46	−1.98	−11.54	**0.81**	**5.22**	31.11	1.08
GARCH rule	−49.24	−6.27	−1.52	−6.25	22.17	49.42	29.85	0.75
Threshold rule	−28.46	−1.95	−1.50	−3.50	35.20	33.16	**41.31**	1.89
Machine-learning policy								
GBDT	−49.78	−8.55	−1.52	−8.18	16.51	49.78	24.07	1.23
Logistic	−55.45	−11.63	−1.46	−9.70	14.24	55.47	20.42	1.00
Language policy								
SOTA	**18.32**	**1.60**	**4.59**	**4.77**	21.52	8.96	29.17	**4.68**
Hindsight reference								
Oracle	608.31	6.28	740.22	35.87	63.91	7.32	–	–

Notes: Upward arrows indicate that higher values are better; downward arrows indicate that lower values are better. Bold marks the best value in each column among the tradable policies; the hindsight oracle is excluded. Appendix E.4 gives detailed definitions of the metrics.

Evaluation metrics. We evaluate portfolio performance using eight financial metrics: total return (TR), annualized Sharpe ratio (ASR), annualized Calmar ratio (ACR), annualized Sortino ratio (ASoR), annualized volatility (AVOL), maximum drawdown (MDD), win rate (WR), and profit/loss ratio (PLR). TR measures cumulative return over the evaluation period. AVOL is the annualized standard deviation of daily returns, and MDD is the largest peak-to-trough decline in portfolio value. ASR measures return relative to total volatility, ACR relative to maximum drawdown, and ASoR relative to downside deviation. WR is the fraction of completed trades with positive profit, and PLR is the average profit of winning trades relative to the average loss of losing trades. Appendix E.4 gives detailed definitions.

Baselines. We compare SOTA with conventional rule-based and supervised machine-learning policies for option-strategy selection. The rule-based baselines include equal-weighted strategy allocation, a GARCH-based rule (Bollerslev, 1986), and a threshold policy based on the structured market features. The GARCH rule combines a volatility forecast with contemporaneous option-market information. The supervised machine-learning baselines predict returns from the numerical market state and map those predictions into strategy choices using predetermined implementation rules. All feasible policies use the same test dates, execution model, transaction-cost assumptions, portfolio constraints, and deterministic resolvers, so performance differences reflect the strategy-selection rule rather than portfolio implementation. Appendix G provides the specifications of the baselines.

4.2 Baseline Comparison

Table 2 compares SOTA with the rule-based and machine-learning baselines over the test period. SOTA earns a total return of 18.32%, an annualized Sharpe ratio of 1.60, and a maximum drawdown of 8.96%. All rule-based and supervised machine-learning baselines produce negative total and risk-adjusted returns in the same evaluation. The equal-weighted rule has relatively low volatility but negative return, while the GARCH and threshold rules exhibit larger losses and drawdowns. The GBDT and logistic policies lose 49.78% and 55.45%, respectively. In this evaluation, SOTA is the only feasible policy with positive total and risk-adjusted returns, indicating that the post-trained language policy captures strategy-selection patterns not reproduced by the fixed rules or supervised mappings considered here.

The final row of Table 2 reports a hindsight oracle. At each decision time, it selects the best-performing strategy from the same candidate set using realized future outcomes. The oracle is therefore not tradable, but it uses the same transaction costs, sizing rule, position limits, and hedging rule as the feasible policies. Its return provides a reference for the profitability available within the structured strategy space.

5 Ablation Studies

We examine two components of the training pipeline: the role of news in the policy information set and the contribution of reinforcement learning after supervised fine-tuning. Table 3 summarizes the results.

5.1 The Role of News in the Information Set

To reduce temporal and identity leakage, we anonymize the information presented to the frontier teacher by masking stock identity, absolute price levels, and calendar time, while retaining market states and contemporaneously available news. We find that including anonymized news improves the quality of the frontier-model trajectories used for supervised fine-tuning. After filtering trajectories by annualized Sharpe ratio, Appendix E.5 reports performance across the retained trajectories by strategy family.

We then ask whether news remains useful during reinforcement learning. Holding the supervised checkpoint fixed, we compare reinforcement learning with and without contemporaneous news in the policy information set. The policy trained without news achieves an 18.32% total return and a Sharpe ratio of 1.60, with a maximum drawdown of 8.96%. Retaining news during reinforcement learning lowers total return to -2.72% and raises maximum drawdown to 31.69%.

These results point to an asymmetric role of textual information: news is useful for constructing frontier-model supervision, but retaining it during policy optimization does not improve out-of-sample performance.

5.2 The Role of Reinforcement Learning

We next evaluate the supervised policy before reinforcement learning. Because the SFT-only policy and the RL-with-news policy use the same information set, their comparison isolates the contribution of reinforcement learning.

Reinforcement learning raises total return from -10.19% to -2.72% and improves the Sharpe ratio from -1.25 to -0.16, although volatility and maximum drawdown also increase. This indicates that reinforcement learning improves the supervised policy under the same information set, while the performance of the full method additionally depends on removing news during policy optimization.

Table 3: Ablation studies over the test period. Bold marks the best value in each column.

	Factor		Profit	Risk-adjusted profit			Risk		Trade	
Variant	RL	News	TR(%)↑	ASR↑	ACR↑	ASoR↑	AVOL(%)↓	MDD(%)↓	WR(%)↑	PLR↑
Full method										
SOTA	✓	×	**18.32**	**1.60**	**4.59**	**4.77**	21.52	**8.96**	29.17	**4.68**
Ablations										
w/ news during RL	✓	✓	−2.72	−0.16	−0.17	−0.29	36.00	31.69	**38.89**	1.36
w/o RL (SFT only)	×	✓	−10.19	−1.25	−1.27	−1.91	**17.59**	15.56	37.50	0.97

6 Conclusion

We introduce SOTA, a language-model post-training framework for structured option-strategy selection. By separating strategy-level reasoning from contract-level implementation, SOTA allows the learned policy to choose among heterogeneous option strategies while deterministic resolvers handle the implementation. Across 10 actively traded option underlyings, the post-trained policy improves out-of-sample portfolio performance relative to the non-LLM strategy-selection baselines and the fine-tuned language policy. We also document an asymmetric role of textual information: news improves frontier-model supervision, but retaining news during reinforcement learning does not improve out-of-sample performance. More broadly, the results suggest that information useful for constructing expert supervision need not remain equally useful during downstream policy optimization. Future work can extend the learned decision space to contract selection, position sizing,

and dynamic hedging, as well as explore alternative representations of structured market and textual information.

ACKNOWLEDGMENTS

We gratefully acknowledge the Pittsburgh Supercomputing Center for providing computational resources that supported this research. The options and market data used in this study were obtained under licenses from commercial data providers. All errors and omissions remain the responsibility of the authors.

REFERENCES

Gurdip Bakshi and Nikunj Kapadia. Delta-hedged gains and the negative market volatility risk premium. *Review of Financial Studies*, 16(2):527–566, 2003. doi: 10.1093/rfs/hhg002.

Gurdip Bakshi, Nikunj Kapadia, and Dilip Madan. Stock return characteristics, skew laws, and the differential pricing of individual equity options. *Review of Financial Studies*, 16(1):101–143, 2003.

Turan G. Bali, Heiner Beckmeyer, Mathis Mörke, and Florian Weigert. Option return predictability with machine learning and big data. *Review of Financial Studies*, 36(9):3548–3602, 2023. doi: 10.1093/rfs/hhad017.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

Douglas T. Breeden and Robert H. Litzenberger. Prices of state-contingent claims implicit in option prices. *Journal of Business*, 51(4):621–651, 1978.

Antonio Briola, Jeremy Turiel, Riccardo Marcaccioli, Alvaro Cauderan, and Tomaso Aste. Deep reinforcement learning for active high frequency trading. *arXiv preprint arXiv:2101.07107*, 2021.

Svetlana Bryzgalova, Anna Pavlova, and Taisiya Sikorskaya. Retail trading in options and the rise of the big three wholesalers. *Journal of Finance*, 78(6):3465–3514, 2023. doi: 10.1111/jofi.13285.

Hans Bühler, Lukas Gonon, Josef Teichmann, and Ben Wood. Deep hedging. *Quantitative Finance*, 19(8):1271–1291, 2019. doi: 10.1080/14697688.2019.1571683.

Jie Cao and Bing Han. Cross section of option returns and idiosyncratic stock volatility. *Journal of Financial Economics*, 108(1):231–249, 2013.

Peter Carr and Liuren Wu. Variance risk premiums. *Review of Financial Studies*, 22(3):1311–1341, 2009.

Yanxu Chen, Zijun Yao, Yantao Liu, Amy Xin, Jin Ye, Jianing Yu, Lei Hou, and Juanzi Li. StockBench: Can LLM agents trade stocks profitably in real-world markets? *arXiv preprint arXiv:2510.02209*, 2025a.

Zeting Chen, Xinyu Cai, Molei Qin, and Bo An. OPHR: Mastering volatility trading with multi-agent deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Bent J. Christensen and Nagpurnanand R. Prabhala. The relation between implied and realized volatility. *Journal of Financial Economics*, 50(2):125–150, 1998. doi: 10.1016/S0304-405X(98)00034-8.

Rama Cont and José da Fonseca. Dynamics of implied volatility surfaces. *Quantitative Finance*, 2 (1):45–60, 2002. doi: 10.1088/1469-7688/2/1/304.

Charles J. Corrado and Tie Su. Skewness and kurtosis in S&P 500 index returns implied by option prices. *Journal of Financial Research*, 19(2):175–192, 1996. doi: 10.1111/j.1475-6803.1996.tb00592.x.

Patrick Dennis and Stewart Mayhew. Risk-neutral skewness: Evidence from stock options. *Journal of Financial and Quantitative Analysis*, 37(3):471–493, 2002. doi: 10.2307/3594989.

Bernard Dumas, Jeff Fleming, and Robert E. Whaley. Implied volatility functions: Empirical tests. *Journal of Finance*, 53(6):2059–2106, 1998. doi: 10.1111/0022-1082.00083.

David Easley, Maureen O'Hara, and P. S. Srinivas. Option volume and stock prices: Evidence on where informed traders trade. *The Journal of Finance*, 53(2):431–465, 1998. doi: 10.1111/0022-1082.194060.

Nicolae Gârleanu, Lasse Heje Pedersen, and Allen M. Poteshman. Demand-based option pricing. *The Review of Financial Studies*, 22(10):4259–4299, 2009. doi: 10.1093/rfs/hhp005.

Amit Goyal and Alessio Saretto. Cross-section of option returns and volatility. *Journal of Financial Economics*, 94(2):310–326, 2009.

Zhengyao Jiang and Jinjun Liang. Cryptocurrency portfolio management with deep reinforcement learning. In *2017 Intelligent Systems Conference (IntelliSys)*, pp. 905–913, 2017. doi: 10.1109/IntelliSys.2017.8324237.

Roman Kozhan, Anthony Neuberger, and Paul Schneider. The skew risk premium in the equity index market. *Review of Financial Studies*, 26(9):2174–2203, 2013. doi: 10.1093/rfs/hht039.

Charles M. C. Lee and Mark J. Ready. Inferring trade direction from intraday data. *The Journal of Finance*, 46(2):733–746, 1991. doi: 10.1111/j.1540-6261.1991.tb02683.x.

Xiao-Yang Liu, Ziyi Xia, Jingyang Rui, Jiechao Gao, Hongyang Yang, Ming Zhu, Christina Dan Wang, Zhaoran Wang, and Jian Guo. FinRL-Meta: Market environments and benchmarks for data-driven financial reinforcement learning. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 35, 2022.

Robert C. Merton. Theory of rational option pricing. *Bell Journal of Economics and Management Science*, 4(1):141–183, 1973.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. doi: 10.1038/nature14236.

Sophie X. Ni, Jun Pan, and Allen M. Poteshman. Volatility information trading in the option market. *Journal of Finance*, 63(3):1059–1091, 2008. doi: 10.1111/j.1540-6261.2008.01352.x.

Jun Pan and Allen M. Poteshman. The information in option volume for future stock prices. *The Review of Financial Studies*, 19(3):871–908, 2006. doi: 10.1093/rfs/hhj024.

Molei Qin, Shuo Sun, Wentao Zhang, Haochong Xia, Xinrun Wang, and Bo An. EarnHFT: Efficient hierarchical reinforcement learning for high frequency trading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14669–14676, 2024. doi: 10.1609/aaai.v38i13.29384.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Her-Jiun Sheu and Yu-Chen Wei. Effective options trading strategies based on volatility forecasting recruiting investor sentiment. *Expert Systems with Applications*, 38(1):585–596, 2011. doi: 10.1016/j.eswa.2010.07.007.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, second edition, 2018.

Wee Ling Tan, Stephen Roberts, and Stefan Zohren. Deep learning for options trading: An end-to-end approach. In *Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF '24)*, pp. 487–495. Association for Computing Machinery, 2024.

A. Elizabeth Whalley and Paul Wilmott. An asymptotic analysis of an optimal hedging model for option pricing with transaction costs. *Mathematical Finance*, 7(3):307–324, 1997.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. TradingAgents: Multi-agents LLM financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024.

Hongyang Yang, Xiao-Yang Liu, Shan Zhong, and Anwar Walid. Deep reinforcement learning for automated stock trading: An ensemble strategy. In *Proceedings of the First ACM International Conference on AI in Finance (ICAIF '20)*, pp. 1–8, 2020. doi: 10.1145/3383455.3422540.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*, 2023a.

Hongyang Yang, Boyu Zhang, Neng Wang, Cheng Guo, Xiaoli Zhang, Likun Lin, Junlin Wang, Tianyu Zhou, Mao Guan, Runjia Zhang, and Christina Dan Wang. FinRobot: An open-source AI agent platform for financial applications using large language models. *arXiv preprint arXiv:2405.14767*, 2024.

Yi Yang, Yixuan Tang, and Kar Yan Tam. InvestLM: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*, 2023b.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Zhenyu Cui, Rong Liu, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. FinCon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In *Advances in Neural Information Processing Systems*, volume 37, pp. 137010–137045, 2024.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Jordan W. Suchow, Denghui Zhang, and Khaldoun Khashanah. FinMem: A performance-enhanced LLM trading agent with layered memory and character design. *IEEE Transactions on Big Data*, 11(6):3443–3459, 2025. doi: 10.1109/TBDATA.2025.3593370.

Chong Zhang, Xinyi Liu, Zhongmou Zhang, Mingyu Jin, Lingyao Li, Zhenting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, Sujian Li, Mengnan Du, and Yongfeng Zhang. When AI meets finance (StockAgent): A benchmark for simulating large language model behaviors in controlled trading environments. *ACM Transactions on Intelligent Systems and Technology*, 2026. doi: 10.1145/3844605.

Jin E. Zhang and Yi Xiang. The implied volatility smirk. *Quantitative Finance*, 8(3):263–284, 2008. doi: 10.1080/14697680601173444.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4314–4325, 2024. doi: 10.1145/3637528.3671801.

Chuqiao Zong, Chaojie Wang, Molei Qin, Lei Feng, Xinrun Wang, and Bo An. MacroHFT: Memory augmented context-aware reinforcement learning on high frequency trading. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4712–4721, 2024. doi: 10.1145/3637528.3672064.

A Option Pricing and Greeks

The Black-Scholes-Merton (BSM) Model is a seminal framework for pricing options. Write S for the underlying price, K the strike, T the time to expiry in years, r the risk-free rate, q the continuous dividend yield, and σ the implied volatility.

A.1 Black–Scholes-Merton Pricing Formula

With

$$d_1 = \frac{\ln(S/K) + \left(r - q + \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}, \qquad d_2 = d_1 - \sigma\sqrt{T}, \tag{6}$$

the European call and put values are

$$C = Se^{-qT}\Phi(d_1) \; - \; Ke^{-rT}\Phi(d_2), \tag{7}$$

$$P = Ke^{-rT}\Phi(-d_2) \; - \; Se^{-qT}\Phi(-d_1). \tag{8}$$

where Φ is the standard normal distribution function. Implied volatility is defined as the volatility σ that equates an option's market price with the theoretical value.

A.2 Option Greeks

$$\Delta = \frac{\partial V}{\partial S} = e^{-qT}N(d_1) \text{ (call)}, \qquad -e^{-qT}N(-d_1) \text{ (put)}, \tag{9}$$

$$\Gamma = \frac{\partial^2 v}{\partial S^2} = \frac{e^{-qT}\,\phi(d_1)}{S\,\sigma\sqrt{T}}, \tag{10}$$

$$\nu = \frac{\partial V}{\partial \sigma} = Se^{-qT}\,\phi(d_1)\,\sqrt{T}, \tag{11}$$

$$\Theta = \frac{\partial V}{\partial t} = -\frac{Se^{-qT}\phi(d_1)\sigma}{2\sqrt{T}} - rKe^{-rT}N(d_2) + qSe^{-qT}N(d_1) \text{ (call)}. \tag{12}$$

where ϕ is the standard normal density.

B Strategy payoffs illustrated

This appendix illustrates the net payoff structure of each of the nine strategy families considered.

C Point-in-time data pipeline

C.1 Market features

Write S_t for the underlying price at the decision, r_i for a daily close-to-close log return, and $\sigma^{\mathrm{iv}}(\delta, \tau)$ for the implied volatility at delta coordinate δ and tenor τ calendar days. Unless stated otherwise volatility is annualized for comparability.

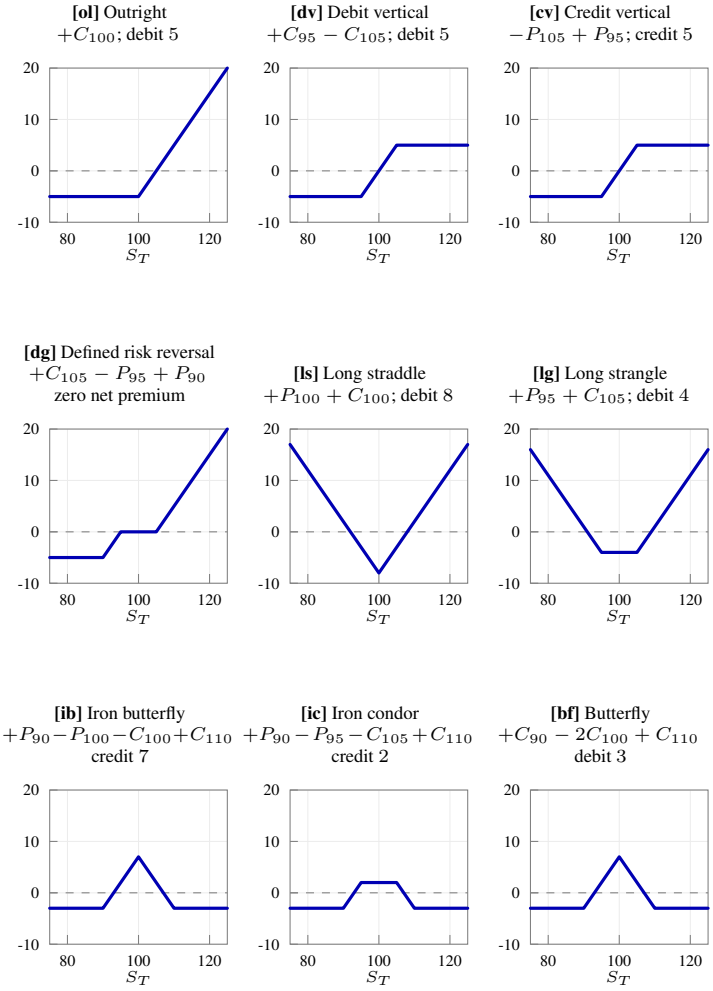
Return and realized volatility. The step return is the log return over the session,

$$\mathrm{ret}_t \; = \; \log\!\big(S_t/S_{t-1}\big), \tag{13}$$

Realized volatility is the zero-mean estimator over a trailing $n = 21$ trading days,

$$\mathrm{rv}_t \; = \; \sqrt{\frac{252}{n}\sum_{i=1}^{n} r_{t-i}^2}. \tag{14}$$

Table 4: Expiration profit and loss of the nine option-strategy families, net of the initial cost of the package [a]



[a] C_K and P_K denote a call and a put with strike K and a common expiration T; the subscript denotes the strike. Their terminal payoffs are $(S_T - K)^+$ and $(K - S_T)^+$, where $z^+ = \max(z, 0)$. A plus sign denotes a long position and a minus sign a short position; coefficients indicate contract ratios. A *debit* is the net premium paid at entry and is subtracted from terminal payoff, while a *credit* is the net premium received and is added. All values are per share, with illustrative strikes and premiums. Vertical and butterfly families admit the corresponding call/put and directional variants; one representative payoff is shown for each family.

Implied volatility and its change. The implied volatility is measured using the at-the-money call (with Δ=0.5) with 30 days to expiry. The feature dv is the intraday change of the iv from market open to market close.

$$\mathrm{iv}_t \;=\; \sigma_t^{\mathrm{iv}}(0.50\,\mathrm{C},\,30), \qquad \mathrm{dv}_t \;=\; \mathrm{iv}_t - \mathrm{iv}_t^{\mathrm{open}}. \tag{15}$$

The iv wedge, term structure, skewness, and curvature.

$$\mathrm{w}_t \;=\; \mathrm{iv}_t - \mathrm{rv}_t, \tag{16}$$

$$\mathrm{ts}_t \;=\; \sigma_t^{\mathrm{iv}}(0.50\,\mathrm{C}, 90) \;-\; \sigma_t^{\mathrm{iv}}(0.50\,\mathrm{C}, 30), \tag{17}$$

$$\mathrm{sk}_t \;=\; \sigma_t^{\mathrm{iv}}(0.25\,\mathrm{P}, 30) \;-\; \sigma_t^{\mathrm{iv}}(0.25\,\mathrm{C}, 30), \tag{18}$$

$$\mathrm{bf}_t \;=\; \tfrac{1}{2}\big[\sigma_t^{\mathrm{iv}}(0.25\,\mathrm{P}, 30) + \sigma_t^{\mathrm{iv}}(0.25\,\mathrm{C}, 30)\big] \;-\; \mathrm{iv}_t. \tag{19}$$

The wedge w_t measures the difference between implied and realized volatility. The term structure ts_t measures the slope of the implied-volatility term structure as the difference between 90-day and 30-day at-the-money call implied volatility. The skewness measure sk_t captures the asymmetry of the implied-volatility surface as the difference between 30-day 25-delta put and call implied volatility. Finally, the curvature measure bf_t is the 30-day 25-delta butterfly, defined as the average implied volatility of the 25-delta put and call relative to at-the-money implied volatility. It captures the curvature of the implied-volatility smile.

Flow. The flow imbalance is the delta-weighted signed demand over the session, divided by the total delta exposure transacted,

$$\mathrm{fi}_t \;=\; \frac{\sum_k q_k \,\mathrm{sign}_k\,|\Delta_k|}{\sum_k q_k\,|\Delta_k|}, \qquad \mathrm{doi}_t \;=\; \frac{\mathrm{OI}_t - \mathrm{OI}_{t-1}}{\mathrm{OI}_{t-1}}, \tag{20}$$

the sums running over option trades k printed between the session open and the decision time, with $\mathrm{sign}_k \in \{+1, -1\}$ the inferred trade direction. Trade direction is inferred from the prevailing quotes: trades above (below) the midpoint are classified as buyer- (seller-) initiated, with midpoint trades assigned using the most recent price change (Lee & Ready, 1991).

D TRADING AND EXECUTION ASSUMPTIONS

Table 5: Trading and execution assumptions.

Assumption	Value
Traded price	Midpoint of the best bid and best ask
Option fee	\$0.65/contract/leg
Assignment fee	\$5/event
Share commission	\$0.005/share
Stock borrow	0.5% annually
Stop-loss	60% of entry premium
Ruin floor	20% of initial NAV

E PORTFOLIO IMPLEMENTATION

This appendix states the three resolvers to determine contract, size, and hedge.

E.1 CONTRACT RESOLVER

The contract resolver maps a strategy into an exact set of contracts. The expiry is chosen by the closest available expiry to the anchor of each tenor bucket $a \in \{5, 14, 45, 120\}$ days. The strike price is selected by the contract with Δ exposure closest to the requested Δ^*.

E.2 SIZE RESOLVER

The packages in a portfolio are traded in integer units based on equal notional value, and equal risk.

The risk is estimated by a second-order Taylor expansion of package value. The shocks are a one-standard-deviation move in the underlying taken in whichever direction hurts, a 10% move in implied volatility, and one day of carry:

$$R = \underbrace{|\Delta S|\,\hat{\sigma}}_{\text{delta}} - \underbrace{\frac{1}{2}\Gamma(S\hat{\sigma})^2}_{\text{gamma}} + \underbrace{|\nu|\,\sigma\,\kappa}_{\text{vega}} - \underbrace{\Theta}_{\text{carry}}, \qquad \hat{\sigma} = \frac{\sigma^{\text{atm}}}{\sqrt{252}}. \tag{21}$$

with κ the relative volatility shock.

E.3 HEDGE RESOLVER

Hedging is applied to non-directional strategies to offset the Δ exposure of the portfolio. To reduce transaction cost, there is a no-hedge band for small Δ, which is given by the formula that Whalley & Wilmott (1997) derive as

$$H \;=\; \left(\frac{3}{2}\,\frac{k\,S\,\Gamma^2\,e^{-r(T-t)}}{\lambda}\right)^{1/3}, \tag{22}$$

where the hedging neutralizes the aggregated option positions' delta exposure.

E.4 EVALUATION METRICS

Let N_t denote portfolio net asset value at the close of date t, r_t the daily portfolio return, $\bar{r}$ its sample mean, $m = 252$ the annualization factor, and r_f the annualized risk-free rate.

Total return (TR). TR measures the cumulative portfolio return over the evaluation period: $\mathrm{TR} = N_T/N_0 - 1$.

Annualized volatility (AVOL). AVOL measures the annualized standard deviation of daily portfolio returns: $\mathrm{AVOL} = \sqrt{m}\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(r_t - \bar{r})^2}$.

Maximum drawdown (MDD). MDD measures the largest peak-to-trough decline in portfolio net asset value: $\mathrm{MDD} = \max_{0\le u\le t\le T}(N_u - N_t)/N_u$.

Annualized Sharpe ratio (ASR). ASR measures annualized excess return relative to total volatility: $\mathrm{ASR} = (m\bar{r} - r_f)/\mathrm{AVOL}$.

Annualized Calmar ratio (ACR). ACR measures annualized excess return relative to maximum drawdown: $\mathrm{ACR} = (m\bar{r} - r_f)/\mathrm{MDD}$.

Annualized Sortino ratio (ASoR). ASoR measures annualized excess return relative to annualized downside deviation: $\mathrm{ASoR} = (m\bar{r} - r_f)/\mathrm{DD}$, where $\mathrm{DD} = \sqrt{m}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\min(r_t, 0)^2}$.

Win rate (WR). WR is the fraction of completed trades with positive realized profit: $\mathrm{WR} = N_{\text{win}}/N_{\text{total}}$.

Profit/loss ratio (PLR). PLR is the average profit of winning trades relative to the average absolute loss of losing trades: $\mathrm{PLR} = |\mathbb{E}[r_i \mid r_i > 0]| \,/\, |\mathbb{E}[r_i \mid r_i < 0]|$.

E.5 FRONTIER TRAJECTORIES PERFORMANCE BY STRATEGY FAMILY

Table 6: Frontier teacher performance by strategy families.

			Profit	Risk-adjusted			Risk		
Code	Family	Trades	TR(%)	ASR	ACR	ASoR	AVOL(%)	MDD(%)	WR(%)
Directional									
ol	Outright	4.29	+3.16	1.43	5.08	2.93	6.91	2.33	46.3
dv	Debit vertical	2.84	+2.86	2.10	8.24	4.60	5.51	1.68	75.3
cv	Credit vertical	1.46	+0.73	1.97	14.88	7.28	2.54	0.70	82.4
dg	Defined risk reversal	0.28	+0.22	1.31	4.48	1.99	4.87	1.51	60.9
Long gamma									
ls	Long straddle	15.00	−0.57	−0.20	−0.29	−0.21	20.38	9.75	30.8
lg	Long strangle	0.69	+0.68	0.25	2.46	1.72	6.20	1.91	38.3
Short gamma									
ib	Iron butterfly	16.23	+1.54	0.34	0.73	0.34	10.44	4.89	47.4
ic	Iron condor	5.49	+2.21	1.47	4.23	2.44	7.93	3.20	62.8
bf	Butterfly	0.02	−0.01	−0.29	3.09	0.20	1.43	1.02	33.3
Delta hedge		–	−2.83	−0.82	−1.06	−1.16	15.55	12.11	–
Total		–	**+7.99**	–	–	–	–	–	–

We generate 1,000 trajectories using the frontier model and retain those with an annualized Sharpe ratio above 0.75, resulting in 144 trajectories used for SFT.

F TRAINING ALGORITHM

Algorithm 1 states the two training phases: frontier-teacher strategy generation followed by supervised fine-tuning, and reinforcement learning in the portfolio environment.

Algorithm 1 SOTA Training

Require: Trading environment ENV; frontier teacher π_T; base policy π_{θ_0}; deterministic resolver $\mathcal{R}$; SFT episodes $\mathcal{D}_{\mathrm{SFT}}$; RL episodes $\mathcal{D}_{\mathrm{RL}}$; discount factor γ

Phase I: Frontier-Teacher Strategy Generation and SFT
Initialize supervised trajectory set $\mathcal{D}_T \leftarrow \emptyset$
for each teacher run over the supervised-training period **do**
  Initialize trading environment ENV with portfolio NAV $V_0 = \$1{,}000{,}000$
  for each episode $e \in \mathcal{D}_{\mathrm{SFT}}$, in calendar order **do**
    Carry the portfolio (cash, open positions, NAV) over from the previous episode, if any
    for each decision time t in episode e **do**
      Observe market state s_t and contemporaneous news n_t
      Generate frontier-teacher strategy decision

$$a_t^T \sim \pi_T(\cdot \mid s_t, n_t)$$

      Resolve strategy decision into executable positions

$$z_t^T = \mathcal{R}(a_t^T, s_t)$$

      Execute option and hedge trades in ENV
      Advance ENV and observe updated portfolio state
      Store (s_t, n_t, a_t^T) in $\mathcal{D}_T$
    end for
  end for
end for
Supervised fine-tune the base policy on teacher-generated trajectories

$$\theta_{\mathrm{SFT}} = \arg\min_{\theta} \left[- \sum_{(x_t, a_t^T) \in \mathcal{D}_T} \log \pi_\theta(a_t^T \mid x_t) \right]$$

Phase II: Reinforcement Learning
Initialize policy $\pi_\theta \leftarrow \pi_{\theta_{\mathrm{SFT}}}$
for each episode $e \in \mathcal{D}_{\mathrm{RL}}$ **do**
  Reinitialize the same trading environment ENV
  Set initial portfolio NAV $V_0 = \$1{,}000{,}000$
  for each decision time t in episode e **do**
    Observe policy information x_t
    Generate strategy decision

$$a_t \sim \pi_\theta(\cdot \mid x_t)$$

    Resolve strategy decision into executable positions

$$z_t = \mathcal{R}(a_t, s_t)$$

    Execute option and hedge trades in ENV
    Advance ENV and observe V_{t+1} and the next state
    Compute portfolio reward

$$r_t = \log V_{t+1} - \log V_t$$

    Store (x_t, a_t, r_t) in the rollout
  end for
  Compute discounted portfolio rewards

$$G_t = \sum_{k \geq 0} \gamma^k r_{t+k}$$

  Update π_θ using the RL objective
end for
return trained option-strategy policy π_θ

Table 7: GARCH rule.

Parameter	Symbol	Value	Selection method
ARCH coefficient	α	0.060	QML on estimation sample
GARCH coefficient	β	0.762	QML on estimation sample
Persistence	$\alpha + \beta$	0.822	implied (half-life 3.5 d)
Winsorisation	—	$4\bar{\sigma}_i$	fixed a priori
Forecast horizon	H	10 trading days	matches the 14-day tenor anchor
Estimation sample	—	3,710 obs, 2023-09-01 to 2025-02-28	strictly pre-test
Long-vol entry	—	$R \leq 0.90$	fixed; 10% edge vs 3.2% friction
Short-vol entry	—	$R \geq 1.25$	fixed; 25% edge vs 10.6% friction
Long-vol structure	—	long straddle	defined-risk long gamma
Short-vol structure	—	iron condor	defined-risk short gamma

Table 8: Threshold rule.

Family	Code	Feature	Entry	Edge	Friction
Participation gate, applied to every family					
all	—	ts	< 0	—	—
Volatility: iv/rv, *fair value* 1.0					
Long straddle	`ls`	iv/rv	≤ 0.90	10%	3.2%
Iron butterfly	`ib`	iv/rv	≥ 1.20	20%	9.1%
Iron condor	`ic`	iv/rv	≥ 1.25	25%	10.6%
Curvature: bf/iv, *fair value* 0.0					
Long strangle	`lg`	bf/iv	≤ 0	—	4.8%
Butterfly	`bf`	bf/iv	≤ 0	—	n/a
Skew: sk/iv, *fair value* 0.0					
Defined-risk reversal	`dg`	sk/iv	≥ 0.15	—	32.9[a]
Credit vertical	`cv`	sk/iv	≤ 0	—	9.7%
Momentum: $\lvert\text{ret}\rvert$, *orientation from the sign*					
Outright	`ol`	$\lvert\text{ret}\rvert$	≥ 0.03	—	3.2%
Debit vertical	`dv`	$\lvert\text{ret}\rvert$	≥ 0.02	—	11.3%

G BASELINE AND TRAINING CONFIGURATIONS

This section provides the specifications for the non-LLM baselines, and hyperparameters for the LLM training.

G.1 GARCH RULE

We fit GARCH(1,1) on daily close-to-close log returns, estimated by Gaussian quasi-maximum likelihood with variance targeting, $h_{i,t} = \bar{\sigma}_i^2(1 - \alpha - \beta) + \alpha\varepsilon_{i,t-1}^2 + \beta h_{i,t-1}$.

G.2 THRESHOLD RULE

We consider one fixed threshold on one feature per family, covering all nine admitted families, and construct an equal-weighted portfolio across strategy families and underlying.

G.3 TRAINING HYPERPARAMETERS

Table 9: Training hyperparameters for Qwen3.8-27B.

Parameter	Qwen3.8-27B
Model and supervised fine-tuning	
Hugging Face revision	`1d4bf0f2`
Precision	bf16
SFT episodes	414
Maximum sequence length	33,280
Train batch size	16
Learning rate	1×10^{-5}
Adam (β_1, β_2)	$(0.9, 0.999)$
Weight decay	0.01
Epochs	3
Optimizer steps	75
Reinforcement learning (GRPO)	
Group size	8
Train batch size	3
Actor learning rate	5×10^{-7}
KL coefficient	0.01
PPO clip ratio	0.2
Maximum prompt / response tokens	8,192/28,672
Rollout temperature (train / eval)	1.0/0
Optimizer steps	40
Checkpoint interval	10
Compute	
Tensor parallelism	2
GPUs $\times$ nodes	8×1
Seed	0

Note: The maximum sequence length of 33,280 includes chat-template control tokens in addition to the 32,768-token content budget. Episodes exceeding the content budget are excluded during corpus construction. During RL, discounting is applied within the trajectory-level reward. Evaluation returns are undiscounted. Checkpoints are written every 10 optimizer steps, and each was evaluated on the test window. The results reported throughout use the checkpoint at step 20.

G.4 ACTION SPACE AND PORTFOLIO CONSTRUCTION

Table 10 gives the parameters of the action space and of the three deterministic resolvers. These values are tuned on the supervised window and then held fixed for the RL training window and test window.

G.5 DEFAULT STRATEGY COORDINATES

A family name alone does not determine a package. Where the policy omits a delta coordinate, the contract resolver substitutes the family default in Table 11. Any coordinate the policy does supply is clipped to the stated range and snapped to the Δ_{step} grid. The resolver selects the contract whose delta is nearest the requested value inside that range. The realized coordinate is returned to the policy on the receipt line, so a request and a fill are never assumed equal.

Table 10: Action space, sizing and hedging parameters.

Parameter	Symbol	Value
Universe and action space		
Underlyings	N	10 (nine single names, one index slot)
Initial net asset value	V_0	\$1,000,000
Cash accrual	r_f	0
Registered families	$\lvert\mathcal{F}\rvert$	9
Tenor buckets (days to expiry)	$\mathcal{K}$	0–7, 8–30, 31–90, 91–180
Order lines per decision	K	≤ 8
Open positions, portfolio	$P_{\max}$	12
Open positions, per underlying	P_u	3
Quote-round candidates per underlying	Q	6
Delta coordinate grid	Δ_{step}	0.05
Size resolver		
Target scenario risk, fraction of NAV	ϕ	0.005
Relative implied-volatility shock	ν	0.10
NAV fraction cap	f_{NAV}	0.10
Scenario risk cap per underlying	ρ_u	0.05
Total open scenario risk cap	ρ_{tot}	0.20
Net dollar delta cap, fraction of NAV	$\delta_{\max}$	0.50
Collateral buffer	b	0.05
Hedge resolver		
Band half-width, fraction of NAV	Δ_{band}	0.005
Hedged families	—	the five volatility families
Band rule; hedge target	—	fixed; band edge
Marking and forced exit		
Half-spread multiplier	m_{hs}	1.0 (full quoted half-spread)
Share half-spread	s_{shr}	1 bp
Force-close threshold, days to expiry	$d_{\min}$	1
Maximum tolerated stale-mark share	$\bar{s}$	0.25
Binomial steps, American valuation	n_{tree}	128
Maximum mark quote age	τ_q	6 h

Table 11: Family default coordinates and admitted ranges.

Family	Default coordinates	Admitted range
Outright	long delta 0.55	$[0.20, 0.80]$
Debit vertical	long delta 0.55, width 0.10	long $[0.20, 0.80]$, width $[0.05, 0.30]$
Credit vertical	short delta 0.35, width 0.10	short $[0.10, 0.55]$, width $[0.05, 0.30]$
Butterfly	centre 0.45, widths 0.10/0.10	centre $[0.25, 0.70]$, widths $[0.05, 0.30]$
Defined-risk reversal	directional 0.35, tail wing 0.20	directional $[0.15, 0.60]$, wing $[0.05, 0.35]$
Long straddle	call 0.50, put 0.50	$[0.10, 0.55]$
Long strangle	call 0.25, put 0.25	$[0.10, 0.55]$
Iron butterfly	short delta 0.50, wing 0.25	short $[0.10, 0.55]$, wing $[0.05, 0.40]$
Iron condor	short delta 0.30, width 0.10	short $[0.10, 0.55]$, width $[0.05, 0.30]$

The coordinates are given by the Δ exposure of the contract.