Type	Strategy	TSLA			AMZN			NIO			MSFT		
		SPR	CR	MDD	SPR	CR	MDD	SPR	CR	MDD	SPR	CR	MDD
FinCon Selection (2022-10-05 to 2023-06-10)													
Rule Based	Buy And Hold	0.247	2.056	-54.508	0.150	2.193	-32.177	-0.858	-51.569	-53.563	1.071	32.629	-14.452
	SMA Cross	-0.151	-3.973	-23.173	**0.599**	**13.731**	-18.910	0.810	22.047	**-17.976**	**1.641**	**32.057**	-8.746
	WMA Cross	1.104	32.058	-18.492	0.513	11.765	-21.030	-0.771	-9.412	-18.732	1.526	30.344	-8.883
	ATR Band	-0.554	-22.136	-39.599	0.494	11.007	**-15.842**	0.681	24.684	-21.229	0.827	12.979	**-7.709**
	Bollinger Bands	-0.249	-12.756	-44.655	-0.381	-7.105	-20.615	**0.940**	**25.476**	**-16.623**	**1.759**	31.619	**-3.475**
	Turn of The Month	0.928	27.850	**-11.642**	0.123	3.487	**-14.892**	**0.874**	**31.344**	-17.995	0.407	7.744	-11.955
LLM	FinGPT	0.044	1.549	-42.400	-1.810	-29.811	-29.671	-0.121	-4.959	-37.344	1.315	21.535	-16.503
	FinMem	**1.552**	**34.624**	**-15.674**	-0.773	-18.011	-36.825	-1.180	-48.437	-64.144	-1.247	-22.036	-29.435
	FinAgent	0.271	11.960	-55.734	-1.493	-24.588	-33.074	0.051	0.933	-19.181	-1.247	-27.534	-39.544
	FinCon	**1.972**	**82.871**	-29.727	**0.904**	**24.848**	-25.889	0.335	17.461	-40.647	1.538	31.625	-15.010

Type	Strategy	AAPL			GOOG			NFLX			COIN		
		SPR	CR	MDD	SPR	CR	MDD	SPR	CR	MDD	SPR	CR	MDD
FinCon Selection (2022-10-05 to 2023-06-10)													
Rule Based	Buy And Hold	0.906	24.558	-19.508	**0.683**	**20.884**	-20.278	**1.594**	**77.367**	-20.421	0.024	-23.761	-54.402
	SMA Cross	1.423	21.054	**-6.030**	0.382	8.497	-17.035	-0.855	-8.393	-18.545	0.232	1.286	-35.559
	WMA Cross	**1.648**	**25.257**	-6.114	0.635	13.659	-14.985	-1.009	-9.479	-18.531	0.087	-7.461	-40.883
	ATR Band	0.241	4.522	**-5.159**	0.067	2.616	-13.522	0.522	10.739	**-12.231**	**0.777**	**25.169**	**-22.906**
	Bollinger Bands	-	-	-	0.365	7.526	-13.522	-0.182	-0.710	-13.244	-0.705	-24.371	-40.733
	Turn of The Month	0.098	3.337	-12.498	0.343	7.188	**-13.519**	0.987	18.942	**-10.641**	-0.020	-8.999	**-33.895**
LLM	FinGPT	1.161	20.321	-16.759	0.011	0.242	-26.984	0.472	11.925	-20.201	-1.807	-99.553	-74.967
	FinMem	0.994	12.397	-11.268	0.018	0.311	-21.503	-0.478	-10.306	-27.692	0.017	0.811	-50.390
	FinAgent	1.041	20.757	-19.896	-1.024	-7.440	**-10.360**	1.960	61.303	-20.926	-0.106	-5.971	-56.882
	FinCon	**1.597**	**27.352**	-15.266	**1.052**	**25.077**	-17.530	**2.370**	**69.239**	-20.792	**0.825**	**57.045**	-42.679

Table 7: Backtest performance of traditional rule-based (indicator-based) strategies and *FinCon* over the selective period (2022-10-05 to 2023-06-10), as presented in Yu et al. [52], evaluated using four metrics: cumulative return (CR), Sharpe ratio (SPR), annual volatility (AV), and maximum drawdown (MDD). The best metrics are highlighted in red, while the second best are marked in blue. "-" metrics across the board indicate no trade signals were triggered.

Strategies	Parameters
SMA Cross	short_window=10, long_window=20
WMA Cross	short_window=10, long_window=20
ATR Band	atr_period=14, multiplier=1.5
Bollinger Band	period=20, devfactor=2.0
Trend Following	atr_period=10, period=20
Turn of the Month	before_end_of_month_days=5, after_start_of_month_business_days=3
ARIMA	order=(5,1,0)
XGBoost	num_boost_round=10, n_estimators=1000
RL-A2C	learning_rate=1e-5, ent_coef=0.1, vf_coef=0.5, max_grad_norm=0.5, gae_lambda=0.95, gamma=0.99
RL-PPO	batch_size=64, learning_rate=2.5e-4, ent_coef=0.1, clip_range=0.2, gae_lambda=0.95, gamma=0.99
RL-SAC	learning_rate=2e-2, buffer_size=1000000, batch_size=256, learning_starts=100, ent_coef=0.1, tau=0.005, gamma=0.99, action_noise="normal"
RL-TD3	learning_rate=3e-2, buffer_size=1000000, tau=0.005, gamma=0.99, policy_delay=2, target_policy_noise=0.5, target_noise_clip=0.5, action_noise="normal"
FinMem	model=gpt-4o-mini, top_k=3, embedding_model=text-embedding-ada-002, chunk_size=5000
FinAgent	model=gpt-4o-mini, trader_preference=aggressive_trader, top_k=5, previous_action_look_back_days=14

Table 8: Default parameter settings for benchmark strategies.

F Comparative Drawdown Analysis via Underwater Plots

This appendix provides a visual analysis of strategy risk profiles through **underwater plots**. An underwater plot visualises the drawdown of a portfolio over time, offering an intuitive way to assess the depth, duration, and frequency of its losses.

The plots are derived by calculating the percentage loss of a portfolio's equity curve from its running maximum value (its previous peak). At any given point in time, the drawdown D_t is calculated as: $D_t = (\text{Current Value}_t - \text{Previous Peak}_t)/\text{Previous Peak}_t$.

A value of 0% indicates the portfolio is at a new all-time high, while a negative value shows how far it is "underwater". When interpreting the plots, two key features should be considered:

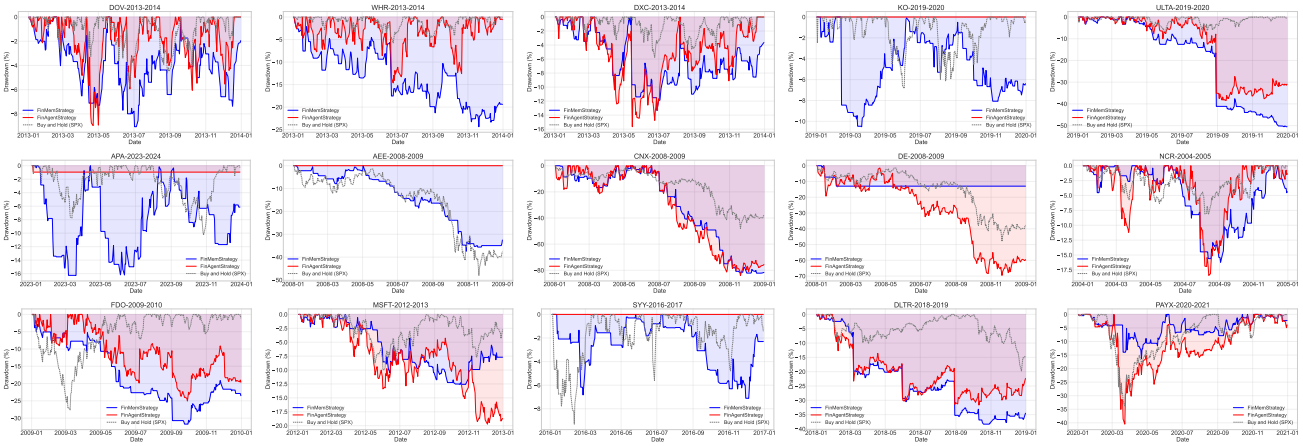

Figure 3: Comparative underwater plots for the FinMem (blue) and FinAgent (red) strategies against the Buy and Hold (SPX) benchmark across individual stocks selected in the Composite setup. The plots are grouped by the market regime of the period shown: bull markets (top two rows), bear market (third row), and sideways markets (bottom two rows).

- **Depth:** The magnitude of the drawdown, indicated by how low the line drops on the y-axis. Deeper drawdowns represent larger losses and greater risk.
- **Duration:** The length of time the line stays below the 0% axis. Longer durations represent slower recoveries and more prolonged periods of underperformance for the investor.

A superior strategy will exhibit shallower and briefer drawdowns compared to its benchmark.

The visual case studies shown in Figure 3 complement the aggregated quantitative results in the main paper, offering a granular perspective on the agents' behavioural patterns under different market conditions.

Bull Markets. The top two rows of the figure display strategy performance during bull market years, revealing a stark divergence in the agents' approaches. The *FinAgent* strategy (red) sometimes exhibits an **overly conservative posture**, as seen in `KO (2019-2020)` and `APA (2023-2024)`. Its drawdowns are shallower than the benchmark's, or it may not trigger any trading activities. While this appears safe, it visually confirms the low beta values from our quantitative analysis and indicates a missed opportunity to capitalise on market gains. However, this risk-averse behaviour is fragile; in the case of `ULTA (2019-2020)`, *FinAgent* experiences a catastrophic drawdown, revealing its risk model to be unreliable and poorly calibrated.

In contrast, the *FinMem* strategy (blue) consistently **fails to manage single-stock volatility**. In most bull-market cases (`DOV`, `WHR`, `DXC`), its drawdowns are significantly deeper and more prolonged than *FinAgent*'s. This demonstrates an inability to handle the inherent risk of the underlying asset, leading to the significant underperformance identified in the main paper.

Bear Markets. The third row, depicting the 2008 Global Financial Crisis, provides the most critical insight into the agents' flaws. While a single stock is expected to be more volatile than the index during a crash, the LLM strategies, particularly *FinMem*, **catastrophically amplify this downside risk**. For `DE`, the *FinMem* strategy's drawdown approaches -75%, a far more severe loss than the SPX benchmark's -50%. Rather than providing any form of risk mitigation, the agents appear to make pro-cyclical decisions that accelerate losses. The *FinAgent* strategy, true to its more conservative nature, often mitigates some of these losses relative to *FinMem*, yet it still fails to generate a positive outcome. For instance, while its drawdown for `CNX` is shallower than FinMem's, it remains severe and prolonged. This relative outperformance is insufficient and aligns with our market regime analysis (§7), which finds that both agents are poorly calibrated for bear markets and ultimately succumb to losses [9].

Sideways Markets. The final two rows illustrate performance in sideways, where the primary challenge is managing idiosyncratic stock risk without a clear market tailwind. Generally (but not consistently), the *FinAgent* strategy (red) exhibits shallower and less severe drawdowns than *FinMem* (blue), as seen in cases like `NCR (2004-2005)`, `FDO (2009-2010)`, and `SYY (2016-2017)`. However, *FinAgent*'s conservative nature can also lead to periods of complete inactivity where no trades are triggered (observed before in bull market and bear market), causing it to miss minor recovery opportunities that the benchmark captures, as seen in `FDO (2009-2010)`.

In summary, these visual case studies reinforce the quantitative conclusions in §6.3 and §7. LLM agents are poorly calibrated to distinct market regimes, behaving too timidly in uptrends and too recklessly in downturns, ultimately failing to provide the adaptive risk management necessary for consistent performance.

G LLM Strategies Cost Analysis

To better understand the practical deployment of LLM-based investing strategies, we monitor the API costs associated with running backtests on the **Composite** experiment with Volatility Effect selection as a representative example. The cost for backtesting *FinAgent* was $198.24, while *FinMem* incurred a significantly lower cost

of $31.79 using GPT-4o mini. This reflects the higher prompt complexity and more frequent calls involved in FinAgent's multi-agent decision-making process.

Extrapolating from these numbers, we estimate that completing all **Composite** experiments required approximately $700 in LLM API costs. The **Selected 4** setup likely incurred even greater cost, given its larger rolling window size and the increased volume of financial news associated with these selectively popular symbols.

FinAgent was roughly 6 times more expensive than FinMem in our tests. Importantly, these figures only account for LLM generation costs (i.e., `chat/completions` endpoints), and do not include the cost of generating embeddings (e.g., via *text-embedding-ada-002*[11]), which would further increase the total budget.

This observation raises a practical consideration for future research: when evaluating LLM-driven strategies, computational cost should be factored into the financial metrics, particularly for real-world deployment scenarios. Incorporating API usage cost into risk-adjusted performance metrics (e.g., Sharpe or Sortino) could provide a more holistic picture of strategy efficiency.

Recommendation. For researchers with limited budget, we recommend adopting open-source LLMs (e.g., `LLaMA`, `Qwen`, `Mistral`) for benchmarking and prototyping. These models can be deployed locally or via cost-effective cloud infrastructure, significantly reducing evaluation costs while enabling reproducible experimentation.

[11] https://platform.openai.com/docs/models/text-embedding-ada-002