Type	Strategy	TSLA				NFLX				AMZN				MSFT			
		SPR↑	CR↑	MDD↑	AV↓	SPR↑	CR↑	MDD↑	AV↓	SPR↑	CR↑	MDD↑	AV↓	SPR↑	CR↑	MDD↑	AV↓
FinMem Selection (2022-10-06 to 2023-04-10)																	
Rule Based	Buy and Hold	-0.342	-20.483	-52.729	55.910	1.326	**43.079**	-20.184	41.523	-0.460	-13.250	-31.546	35.624	0.974	21.171	-14.192	28.327
	SMA Cross	-0.293	-5.540	-18.517	38.602	-1.020	-8.285	-15.942	20.477	-0.420	-4.433	-18.910	27.084	1.515	18.289	-8.746	20.821
	WMA Cross	0.215	3.741	-18.492	42.062	-0.803	-6.004	-14.290	19.826	-0.563	-6.121	-21.030	26.831	1.334	16.576	-8.883	21.503
	ATR Band	-0.595	-19.142	-39.599	42.161	0.150	2.992	-12.231	19.314	**0.622**	**11.007**	-15.842	23.272	1.036	12.979	-7.709	**15.005**
	Bollinger Bands	-0.769	-24.747	-44.655	45.366	-0.558	-4.996	-13.244	**16.754**	-0.402	-7.105	-20.615	26.559	**2.115**	**31.619**	**-3.475**	18.243
	Turn of The Month	0.219	3.639	**-11.642**	**31.042**	0.559	8.383	**-10.641**	17.194	-0.037	0.039	-14.892	20.722	-0.034	0.970	-11.955	**15.097**
Predictor	ARIMA	0.601	15.007	-24.446	41.402	1.159	23.783	-15.043	25.749	-0.225	-4.752	-20.046	26.899	**2.245**	**44.777**	**-7.121**	22.636
	XGBoost	0.331	6.213	-35.374	**37.729**	0.770	10.134	**-11.246**	**14.928**	**1.955**	**42.468**	**-8.816**	25.135	0.895	12.678	-10.734	16.721
RL	A2C	-0.201	-15.876	-52.642	56.172	1.262	36.760	-20.436	37.542	-0.093	-3.253	-24.042	30.903	1.166	24.804	-13.437	26.743
	PPO	-0.254	-18.223	-52.609	57.301	1.420	40.181	-18.036	35.170	-0.576	-9.485	-22.761	24.169	1.149	25.752	-14.444	28.503
	SAC	-0.320	-20.598	-53.614	57.665	1.325	**42.872**	-20.121	41.448	-0.440	-13.215	-32.145	36.533	1.004	22.304	-14.522	28.904
	TD3	-0.343	-20.423	-52.592	55.859	1.325	**42.872**	-20.121	41.448	-0.440	-13.215	-32.145	36.533	0.973	21.026	-14.099	28.073
LLM	FinMem (GPT-4o-mini)	**0.927**	**19.940**	-30.144	48.638	**1.704**	32.549	-13.018	34.766	0.297	2.800	**-2.744**	**10.247**	-0.554	-7.104	-14.588	25.969
	FinMem (GPT-4o)	0.404	5.312	-36.351	54.434	0.896	16.244	-15.234	38.209	-0.968	-20.091	-31.164	40.896	0.792	12.834	-13.555	33.884
	FinMem (reported)	**2.679**	**61.776**	**-10.800**	46.865	**2.017**	36.449	-15.850	36.434	0.233	4.885	-22.929	42.658	1.440	23.261	-14.989	32.562
	FinAgent	-	-	-	-	1.543	41.167	-20.417	51.030	-1.108	-6.113	-9.317	**13.257**	1.252	21.438	-14.502	32.952

Table 2: Backtest performance over the previously reported period (2022-10-06 to 2023-04-10) where LLM investing strategies were shown to be effective. "-" metrics indicate no trading activities were triggered. Top in red and second-best in blue.

6 Experiments

Our experiments address methodological flaws in prior LLM-based investing evaluations identified in §4, specifically survivorship and data-snooping biases from selective stock choices and short evaluation periods. We demonstrate how these practices inflate results and illustrate how FINSABER enables fairer assessments.

Specifically, our experiments include two parts: (1) **Pitfalls of selective evaluation**: Replicating previously reported results on select periods and symbols, then extending this evaluation period to demonstrate performance deterioration. (2) **Fair and robust comparisons**: Implementing systematic stock-selection methods to explicitly mitigate survivorship and data-snooping biases for fairer LLM assessments. We only consider go-long positions, aligning with current LLM strategies. Technical details, including hyperparameter configurations, are provided in Appendix E.

6.1 Pitfalls of Selective Evaluation

Revisiting Reported Claims. We begin by replicating earlier evaluation setups that demonstrated the effectiveness of LLM investing strategies on TSLA, NFLX, AMZN, and MSFT during the previously reported period (6 October 2022 to 10 April 2023). Additionally, we incorporate broader benchmarks, including traditional rule-based, ML, and DL methods. Previous studies omit key details such as exact risk-free rates and transaction costs. Thus, we set a historical average risk-free rate of 0.03 and use Moomoo's[3] standard US commission fee ($0.0049/share, minimum $0.99/order), comparable to HSBC and TradeUp[4].

Table 2 summarises these results. Our analysis indicates that **LLM investors are not universally superior, even in their preferred setups**. Specifically, *FinMem* only consistently outperforms for TSLA, while traditional benchmarks remain competitive or superior for other symbols. These results caution against overly optimistic interpretations from selective evaluations. *FinAgent*, the other LLM-based method, performs similarly to *FinMem* on NFLX and MSFT but generally lacks consistent improvements across the set. Furthermore, **LLM-based strategies exhibit high annual volatility and significant maximum drawdowns**, indicating a high-risk profile. This highlights the necessity of explicit risk assessments when evaluating such strategies.

Further evidence in Appendix D supports the instability of short-period evaluations, where even a slight two-month extension of the evaluation period results in substantial variation for LLM-based strategies.

Extending the Evaluation Period. To further illustrate the limitations of short evaluation horizons, We extend the evaluation period (2004–2024) using the same four symbols (TSLA, NFLX, AMZN, MSFT) to assess LLM performance robustness over the long term.

Table 3 summarises these extended period results. Crucially, extending the evaluation horizon significantly diminishes the perceived superiority of LLM investors. Over two decades, traditional strategies like *Buy and Hold* consistently rank among the top performers across most symbols. TSLA is the only case where LLM investors (*FinMem*, *FinAgent*) clearly lead in AR, while for NFLX, AMZN, and MSFT, *Buy and Hold* or other strategies match or outperform them. This further supports that **previously reported LLM advantages are likely short-lived, potentially hand-picked, and highly sensitive to the evaluation period.**

It is crucial to note that we cannot yet conclude that benchmark strategies cannot outperform the market. As mentioned, backtesting only on popular stocks may inadvertently introduce survivorship bias, as these stocks have gained popularity due to past success during prolonged bull markets. Thus, expanding the range of symbols is essential to ensure a more systematic and unbiased evaluation.

6.2 Fair Comparisons with the Composite Approach

To overcome the aforementioned biases, we introduce the **Composite** evaluation setup within FINSABER. This setup integrates

[3] https://www.moomoo.com/ca/support/topic10_122

[4] https://www.tradeup.com/pricing/detail

Type	Strategy	TSLA					NFLX				
		SPR↑	STR↑	AR↑	MDD↑	AV↓	SPR↑	STR↑	AR↑	MDD↑	AV↓
Rule Based	Buy and Hold	0.630	0.915	37.767	-50.839	45.243	**0.622**	**0.952**	**23.919**	-48.119	41.703
	SMA Cross	0.680	1.013	23.681	**-23.707**	24.680	0.087	0.160	5.514	-28.689	21.836
	WMA Cross	0.664	0.955	21.158	-25.135	24.087	0.004	0.071	1.447	-32.409	23.074
	ATR Band	0.022	0.066	-0.005	-38.536	26.609	0.186	0.377	2.202	-35.603	23.922
	Bollinger Bands	0.193	0.294	4.282	-37.157	26.267	0.075	0.381	0.286	-34.002	23.088
	Trend Following	**0.815**	**1.356**	36.289	-28.113	28.628	0.403	0.646	11.868	-29.179	25.368
	Turn of The Month	0.207	0.353	7.872	-27.902	23.595	0.287	0.487	7.097	-21.646	17.166
Predictor	ARIMA	**0.681**	1.003	24.138	-30.450	27.612	**0.659**	**1.035**	19.022	-27.567	25.514
	XGBoost	0.142	0.370	10.877	-22.901	**19.537**	0.202	0.355	4.957	-21.301	17.302
RL	A2C	0.172	0.249	3.875	-27.367	22.890	0.171	0.243	4.359	-20.960	**16.129**
	PPO	0.469	0.663	28.189	-46.810	40.156	0.541	0.814	19.279	-39.615	33.630
	SAC	0.119	0.190	6.654	**-11.042**	**9.902**	0.186	0.285	8.397	**-9.545**	**9.216**
	TD3	0.417	0.604	23.336	-33.725	30.233	0.291	0.431	10.900	-21.451	19.304
LLM	FinMem	0.641	**1.069**	**42.153**	-34.234	35.030	0.293	0.622	12.566	-27.721	26.876
	FinAgent	0.206	0.649	**38.591**	-36.930	38.302	-0.419	0.621	**22.543**	**-20.466**	22.838

Type	Strategy	AMZN					MSFT				
		SPR↑	STR↑	AR↑	MDD↑	AV↓	SPR↑	STR↑	AR↑	MDD↑	AV↓
Rule Based	Buy and Hold	**0.551**	0.829	**15.997**	-36.842	30.860	**0.461**	**0.620**	**11.238**	-25.463	21.791
	SMA Cross	0.057	0.205	3.896	-22.096	17.520	-0.263	-0.314	0.192	-17.656	11.840
	WMA Cross	0.175	0.300	5.702	-19.309	17.178	-0.363	-0.437	-1.664	-19.075	11.932
	ATR Band	0.443	**0.998**	5.452	-19.990	15.130	0.317	**0.637**	5.725	-11.893	10.885
	Bollinger Bands	0.019	0.125	0.895	-23.757	15.763	-0.054	-0.029	1.578	-16.101	11.931
	Trend Following	**0.649**	**1.111**	**16.018**	-19.120	20.130	0.205	0.321	5.438	-17.515	13.419
	Turn of The Month	-0.029	-0.009	1.534	-20.422	15.728	-0.263	-0.343	-0.177	-14.308	**10.438**
Predictor	ARIMA	0.339	0.504	7.523	-20.612	19.115	0.304	0.466	8.207	-15.227	13.819
	XGBoost	-0.587	-0.366	1.200	**-13.659**	**11.106**	0.171	0.322	5.890	**-10.523**	10.335
RL	A2C	0.165	0.247	3.925	-14.841	**11.654**	0.279	0.380	7.478	-13.447	11.933
	PPO	0.505	0.767	13.831	-29.128	24.392	**0.344**	0.463	8.589	-16.697	14.410
	SAC	0.179	0.257	4.438	**-14.093**	11.665	0.216	0.288	5.329	-14.866	11.835
	TD3	0.382	0.597	11.738	-21.942	19.149	0.050	0.070	1.405	**-9.491**	**6.648**
LLM	FinMem	0.188	0.340	5.695	-28.296	24.786	0.203	0.293	4.567	-19.270	17.891
	FinAgent	0.364	0.663	12.699	-25.516	25.390	0.285	0.432	**11.123**	-18.596	18.863

Table 3: Backtest performance for previously reported LLM-selected symbols over an extended period (2004-01-01 or earliest available to 2024-01-01). Top in red and second-best in blue.

systematic *selection-based strategies* to expand and diversify the stock universe, explicitly addressing survivorship and data-snooping biases. Specifically, we use four unbiased stock selection approaches from the strategies base (details in Appendix B): Random Five, Momentum Factor [40], Volatility Effect [3], and the FinCon Selection Agent in the FinCon [54] framework.

For each rolling window, the selection strategy identifies a set of K symbols. Each *timing-based strategy* is then applied independently to each selected symbol, generating separate trades and performance records. The reported results for each timing strategy reflect the average performance across all selected symbols within the window, as these models operate on individual stocks and do not construct or manage a coordinated portfolio across symbols.

To mitigate survivorship bias, we use historical constituent lists, specifically S&P 500 for US market, at each evaluation period's start and explicitly include delisted symbols. To address data-snooping bias, we evaluate a large and diversified symbol universe: 91, 84, and 63 total distinct symbols for Random Five, Momentum-based, and Volatility-based selection, respectively. These counts reflect all unique symbols encountered across rolling windows, where stocks are reselected in each window, preventing cherry-picking and short-horizon bias.

Table 4 summarises these comprehensive evaluations. Results obtained through this unbiased and systematic approach **further validate our previous findings from the selected-four evaluation**. Specifically, both the Random Five and Momentum-based selections reinforce the conclusion that the previously claimed superiority of LLM investors is largely driven by selective evaluation setups. For instance, in the Random Five setup, *Buy and Hold*, *ATR Band* and *ARIMA* outperform *FinMem* and *FinAgent* in terms of risk-adjusted metrics. Similarly, *ARIMA* and simple rule-based strategies often perform better than LLM-based methods under the Momentum-based selection. In the Volatility-based selection, traditional methods dominate even more clearly: *Buy and Hold* achieves the highest Sharpe (0.703), Sortino (1.291), and AR (7.898%), while *PPO* and

Type	Timing Strategy	RANDOM 5 (91 symbols)					MOMENTUM FACTOR (84 symbols)				
		SPR ↑	STR ↑	AR ↑	MDD ↑	AV ↓	SPR ↑	STR ↑	AR ↑	MDD ↑	AV ↓
Rule Based	Buy and Hold	**0.315**	**0.456**	**6.694**	-35.130	27.410	**0.384**	0.694	9.916	-32.596	37.421
	SMA Cross	-0.298	-0.290	0.446	-22.292	15.774	-0.251	0.008	2.109	-19.438	20.050
	WMA Cross	-0.299	-0.305	0.232	-22.754	15.528	-0.169	0.051	3.674	-18.651	20.330
	ATR Band	0.232	0.425	5.119	-21.535	16.113	0.197	0.595	4.314	-19.407	20.038
	Bollinger Bands	0.129	0.288	3.521	-22.487	16.290	0.114	0.702	1.881	-19.451	21.555
	Trend Following	-0.389	-0.198	2.525	**-8.587**	**8.223**	0.119	0.531	6.380	-15.726	18.696
	Turn of The Month	0.015	0.072	2.870	-18.582	13.542	0.056	0.662	3.197	-18.108	18.055
Predictor	ARIMA	**0.255**	**0.434**	**6.928**	-21.691	17.504	**0.542**	**1.043**	**13.257**	-18.277	22.892
	XGBoost	-0.055	0.028	3.089	-17.160	13.075	0.094	**1.525**	6.131	**-12.754**	17.238
RL	A2C	0.086	0.122	1.902	**-9.220**	**6.887**	0.105	0.171	2.488	-14.452	**14.815**
	PPO	0.179	0.256	3.282	-18.395	13.783	0.185	0.308	1.939	-23.177	25.527
	SAC	0.097	0.142	1.389	-16.058	12.375	0.195	0.321	5.591	**-12.235**	16.144
	TD3	0.173	0.248	3.682	-14.471	11.565	0.186	0.293	3.464	-14.593	**14.953**
LLM	FinMem	-0.253	0.114	-0.094	-24.243	21.214	0.025	0.170	3.649	-23.335	28.078
	FinAgent	0.094	0.323	4.477	-28.059	26.387	0.104	0.534	**13.950**	-20.675	30.635

Type	Timing Strategy	VOLATILITY EFFECT (63 symbols)					FINCON SELECTION AGENT (80 symbols)				
		SPR ↑	STR ↑	AR ↑	MDD ↑	AV ↓	SPR ↑	STR ↑	AR ↑	MDD ↑	AV ↓
Rule Based	Buy and Hold	**0.703**	**1.291**	**7.898**	-14.146	14.720	**0.389**	**0.671**	6.940	-30.943	41.710
	SMA Cross	-0.568	-0.544	0.781	-9.296	8.665	-0.346	-0.351	-4.187	-21.095	20.765
	WMA Cross	-0.665	-0.348	1.908	-8.481	8.573	-0.176	-0.129	-1.683	-19.432	21.141
	ATR Band	-0.026	0.120	2.798	-8.032	7.951	0.181	0.539	4.469	-18.827	24.820
	Bollinger Bands	-0.077	0.029	2.503	-7.618	7.774	0.116	0.333	7.155	-19.145	27.250
	Trend Following	0.230	0.619	5.503	-8.115	9.297	-0.008	0.189	1.358	-19.500	20.400
	Turn of The Month	-0.156	-0.095	2.881	-6.889	7.233	0.013	0.141	2.020	-15.871	16.862
Predictor	ARIMA	0.325	0.838	4.898	-9.111	9.807	**0.532**	**0.841**	**10.662**	-16.018	19.181
	XGBoost	-0.108	-0.055	2.775	-6.676	7.077	0.116	0.325	**8.057**	-15.320	18.078
RL	A2C	0.421	0.795	4.620	**-4.428**	**5.149**	-0.004	-0.061	0.823	-12.557	**11.767**
	PPO	**0.514**	**0.972**	**5.805**	-8.757	9.461	0.132	0.147	2.327	**-9.744**	**10.257**
	SAC	0.402	0.810	3.527	**-4.821**	**5.030**	0.180	0.279	2.661	**-11.979**	14.210
	TD3	0.269	0.394	4.610	-5.442	5.992	0.130	0.334	0.695	-14.621	21.693
LLM	FinMem	-0.228	0.483	4.061	-10.860	11.641	-0.292	0.135	-1.686	-20.809	24.948
	FinAgent	0.241	0.527	4.954	-10.268	11.502	-0.076	0.381	5.168	-15.563	22.565

Table 4: Backtest performance under the Composite setup, using three different selection strategies across historical S&P 500 constituents (2004–2024), including delisted symbols. Top in red and second-best in blue.

ARIMA again show strong all-round performance. LLM-based methods lag behind, with *FinAgent* offering moderate returns but lower Sharpe (0.241) and larger drawdowns. Notably, our reported LLM performances do not adjust for potential data leakage: given the use of pretrained models like GPT-4o, the LLMs may have seen parts of the data during training—yet they still fail to outperform traditional strategies under fair evaluation, casting further doubt on their real-world advantage.

Nevertheless, it is important to acknowledge **LLM-based strategies still show potential regarding absolute annual returns**. For instance, *FinAgent* achieves the highest AR (13.950%) in the MOMENTUM-based selection setup. However, the relatively weaker performance observed in SPR (0.104) and MMD metrics suggests a clear need for improved risk management within LLM-driven approaches before they can be reliably adopted in practice.

Moreover, by comparing *Buy and Hold* with different selection strategies, we clearly identify the relative effectiveness of each selection strategy: VOLATILITY EFFECT selection (Sharpe 0.703) outperforms FINCON SELECTION AGENT (0.389) and MOMENTUM FACTOR (0.384), which in turn surpass RANDOM FIVE (0.315). RL-based methods exhibit the clearest alignment with selection quality. Strategies like *PPO*, *SAC*, and *TD3* systematically achieve their best performance under the VOLATILITY selection and degrade under the other three. This suggests **RL methods are more dependent on the quality of the stock candidates.** Among LLM strategies, *FinAgent* exhibits a greater dependency on selection quality than *FinMem*.

Overall, these results not only confirm our earlier insights but also underscore the critical importance of unbiased, systematic stock-selection methodologies for accurately assessing the true capabilities of LLM-based investing strategies.

6.3 Statistical Validation and Behavioural Diagnostics of LLM Agents

To validate our findings from the composite backtests and diagnose the underlying drivers of LLM agent performance, we conduct a unified statistical and behavioural analysis. First, we conduct paired t-tests comparing *Buy and Hold*, *FinMem*, and *FinAgent* across both **Selected 4** (Table 3) and **Composite** (Table 4) setups. Second, we