Improving the Performance of Long-Term Stock Investments with a Two-Stage LLM Filtering Framework - Supplementary Information

This document accompanies the manuscript and provides the reproducibility material which is requested in the review. This document contains the evaluation window list, per-cell run counts, different configurations' performance, risk statistics, prompts, and more.

S1. Evaluation windows

Each window consists of a 12-month holding period starting from the 1st day of the indicated month. The windows overlap for one month each which means that they are not independent. All of the inference tests in the paper are paired window tests. GPT-5.1 is limited to windows starting after its training cut-off period.

Table S1. *Evaluation windows and analyzed runs per model.*

Window start	Window end	GPT-4o	GPT-4o-mini	GPT-5.1
2024-06-01	2025-06-01	78	82	0
2024-07-01	2025-07-01	78	88	0
2024-08-01	2025-08-01	87	92	0
2024-09-01	2025-09-01	83	309	0
2024-10-01	2025-10-01	75	76	75
2024-11-01	2025-11-01	72	75	74
2024-12-01	2025-12-01	71	76	73
2025-01-01	2026-01-01	72	76	74
2025-02-01	2026-02-01	66	76	74
2025-03-01	2026-03-01	70	74	74
2025-04-01	2026-04-01	63	75	70
2025-05-01	2026-05-01	55	59	58
2025-06-01	2026-06-01	51	54	54

S2. Run accounting

All run files created are listed below. The run is said to be "analyzed" if it was parsed, had all the metadata associated with it, created a portfolio, and provided an estimated return on investment (ROI) over 12 months based on one of the three models. Depth of the runs varies per window; for GPT-4o-mini, the September 2024 window was generated at deeper depth than other windows, making up 309 out of 1,212 analyzed runs of this model (Table S3-GPT-4o-mini).

Table S2a. *Run accounting from generated files to the analyzed set.*

Stage	Count
Run files generated	3876
Parsed as valid JSON	3875
Carried complete run metadata	2992
Produced a portfolio and ROI	2958
Excluded from the 2,958: holding period other than 12 months	113
Excluded from the 2,958: exploratory model tag outside the three reported	86
ANALYZED (reported in the manuscript)	2759

Counting all 3,876 generated files regardless of completion, 119 carried a holding period other than 12 months and 621 an exploratory model tag; of these, 6 and 535 respectively never produced a portfolio and are already counted among the incomplete runs in Table S2b. One further file failed to parse.

Table S2b. *Incomplete-run rate by model. Incomplete runs are those whose strategy generation did not complete, and the table accounts for every generated file, including the exploratory model tags excluded from the analyzed set.*

Model tag	Generated	Incomplete	Rate %
gpt-4o	1191	219	18.40
gpt-4o-mini	1371	102	7.40
gpt-5.1	631	0	0.00
gpt-51	62	62	100.00
gpt-5 (exploratory)	615	529	86.00
gpt-50 (exploratory)	6	6	100.00

Table S2c. *Incomplete-run rate by window, over the three reported models (3,193 generated files).*

Window	Generated	Incomplete	Rate %
202406	283	37	13.10
202407	206	35	17.00
202408	213	31	14.60
202409	425	27	6.40
202410	254	19	7.50
202411	245	20	8.20
202412	230	10	4.30
202501	259	37	14.30
202502	229	13	5.70
202503	229	11	4.80
202504	218	10	4.60
202505	205	33	16.10
202506	197	38	19.30

S3. Per-cell run counts

A matched design requires each configuration to have at least one analyzed run in every window. None of the cells are empty for each model.

Table S3-GPT-4o. Analyzed runs per window and configuration, GPT-4o. Empty cells: 0.

Window	unfiltered	Ranked	5→3	10→3	10→5	20→5	20→10	30→5	30→10	30→15
202406	3	9	9	9	9	9	7	7	9	7
202407	3	8	9	8	9	9	9	8	9	6
202408	3	10	9	9	9	9	9	8	10	11
202409	3	10	9	9	9	9	9	7	8	10
202410	3	7	9	9	9	9	8	7	7	7
202411	3	7	9	8	9	8	7	7	7	7
202412	3	7	9	7	8	8	9	6	7	7
202501	3	6	9	9	9	9	7	6	7	7
202502	3	5	9	7	9	9	8	5	5	6
202503	3	7	9	9	9	9	8	5	6	5
202504	3	5	7	9	7	9	7	6	4	6
202505	3	6	6	6	6	6	6	7	3	6
202506	3	7	5	5	5	5	5	6	5	5

Table S3-GPT-4o-mini. *Analyzed runs per window and configuration, GPT-4o-mini. 0 Empty Cells*

Window	unfiltered	Ranked	5→3	10→3	10→5	20→5	20→10	30→5	30→10	30→15
202406	3	9	9	8	9	9	9	8	9	9
202407	3	9	9	9	9	9	9	8	12	11
202408	3	12	9	9	9	9	8	9	12	12
202409	6	45	29	23	31	31	24	32	44	44
202410	3	7	9	9	9	9	9	7	7	7
202411	3	7	9	8	9	9	9	7	7	7

202412	**3**	**7**	**9**	**9**	**9**	**9**	**9**	**7**	**7**	**7**
202501	**3**	**7**	**9**	**9**	**9**	**9**	**9**	**7**	**7**	**7**
202502	**3**	**7**	**9**	**9**	**9**	**9**	**9**	**7**	**7**	**7**
202503	**3**	**7**	**8**	**8**	**9**	**9**	**9**	**7**	**7**	**7**
202504	**3**	**7**	**9**	**9**	**8**	**9**	**9**	**7**	**7**	**7**
202505	**3**	**7**	**6**	**6**	**6**	**6**	**6**	**7**	**5**	**7**
202506	**3**	**7**	**5**	**5**	**5**	**5**	**5**	**6**	**7**	**6**

Table S3-GPT-5.1. Analyzed runs per window and configuration, GPT-5.1. Empty cells: 0.

Window	unfiltered	Ranked	5→3	10→3	10→5	20→5	20→10	30→5	30→10	30→15
202410	**3**	**7**	**9**	**9**	**9**	**9**	**8**	**7**	**7**	**7**
202411	**3**	**6**	**9**	**9**	**9**	**9**	**8**	**7**	**7**	**7**
202412	**3**	**6**	**9**	**8**	**9**	**9**	**8**	**7**	**7**	**7**
202501	**3**	**6**	**9**	**9**	**9**	**9**	**9**	**7**	**6**	**7**
202502	**3**	**6**	**9**	**9**	**9**	**9**	**8**	**7**	**7**	**7**
202503	**3**	**7**	**9**	**8**	**9**	**8**	**9**	**7**	**7**	**7**
202504	**3**	**5**	**7**	**9**	**7**	**9**	**9**	**7**	**7**	**7**
202505	**3**	**7**	**5**	**5**	**5**	**6**	**6**	**7**	**7**	**7**
202506	**3**	**6**	**5**	**5**	**5**	**5**	**5**	**6**	**7**	**7**

S4. Per-configuration performance and risk

Risk statistics are computed from the daily adjusted closing prices of each portfolio's stock tickers, weights and dates for runs in Table S3, keeping share counts constant at the beginning of the window and annualizing over 252 trading days against a potential zero risk-free rate, with Sharpe, Sortino, and Calmar metrics included as medians and volatility and maximum drawdown as means.

Table S4-GPT-5.1. Per-configuration performance and risk, GPT-5.1. Median values shown for risk ratios.

Configuration	n	Mean ROI %	Median ROI %	SD ROI	Sharpe	Sortino	Volatility %	Max DD %	Calmar	Excess return vs S&P 500 (pp)
unfiltered	**27**	**45.68**	**41.89**	**12.38**	**1.34**	**2.04**	**27.76**	**-22.37**	**1.58**	**25.03**
Ranked	**56**	**38.87**	**38.53**	**17.74**	**1.40**	**2.14**	**23.93**	**-21.74**	**1.66**	**21.48**
5→3	**71**	**42.16**	**40.67**	**11.07**	**1.43**	**2.31**	**24.87**	**-22.59**	**1.70**	**22.74**
10→3	**71**	**54.05**	**42.93**	**27.78**	**1.54**	**2.45**	**25.98**	**-23.00**	**1.92**	**35.00**
10→5	**71**	**56.22**	**52.43**	**23.82**	**1.61**	**2.52**	**26.62**	**-23.63**	**2.02**	**37.21**
20→5	**73**	**50.38**	**46.87**	**26.49**	**1.55**	**2.40**	**26.97**	**-24.24**	**1.81**	**30.67**
20→10	**70**	**60.57**	**56.96**	**18.46**	**1.70**	**2.66**	**28.07**	**-23.24**	**2.26**	**41.53**
30→5	**62**	**32.32**	**31.72**	**14.11**	**1.30**	**2.04**	**22.44**	**-21.62**	**1.40**	**12.19**
30→10	**62**	**40.14**	**37.69**	**17.38**	**1.35**	**2.08**	**24.38**	**-21.82**	**1.75**	**19.95**
30→15	**63**	**44.20**	**41.50**	**19.08**	**1.43**	**2.16**	**25.59**	**-22.33**	**1.76**	**23.37**

Table S4-GPT-4o. Per-configuration performance and risk, GPT-4o. Median values shown for risk ratios.

Configuration	n	Mean ROI %	Median ROI %	SD ROI	Sharpe	Sortino	Volatility %	Max DD %	Calmar	Excess return vs S&P 500 (pp)
unfiltered	**39**	**34.87**	**33.33**	**15.82**	**1.17**	**1.75**	**27.34**	**-23.49**	**1.32**	**16.13**
Ranked	**94**	**30.50**	**32.64**	**11.46**	**1.16**	**1.81**	**26.76**	**-23.90**	**1.31**	**11.64**
5→3	**108**	**39.51**	**42.10**	**24.33**	**1.38**	**2.13**	**28.41**	**-25.21**	**1.58**	**20.96**

10→3	104	40.22	41.99	18.92	1.32	1.99	29.10	-25.06	1.52	21.59
10→5	107	39.98	41.12	24.36	1.31	2.00	29.93	-26.70	1.49	21.46
20→5	108	39.95	39.46	24.34	1.33	2.01	30.10	-26.37	1.49	21.43
20→10	99	35.13	32.65	18.57	1.14	1.74	27.44	-24.36	1.25	16.82
30→5	85	30.47	30.64	11.81	1.24	1.94	23.08	-21.21	1.38	11.49
30→10	87	26.74	25.66	10.48	1.06	1.60	25.42	-23.50	1.08	8.92
30→15	90	28.88	26.41	11.84	1.04	1.56	25.95	-24.43	1.09	11.88

Table S4-GPT-4o-mini. Per-configuration performance and risk, GPT-4o-mini. Median values shown for risk ratios.

Configuration	n	Mean ROI %	Median ROI %	SD ROI	Sharpe	Sortino	Volatility %	Max DD %	Calmar	Excess return vs S&P 500 (pp)
unfiltered	42	38.41	37.61	14.93	1.22	1.84	28.37	-24.86	1.35	19.39
Ranked	138	32.42	34.88	9.36	1.17	1.78	29.17	-26.37	1.29	15.10
5→3	129	37.50	40.48	18.48	1.26	1.89	30.37	-26.55	1.44	18.93
10→3	121	33.76	35.23	10.17	1.15	1.73	29.90	-26.01	1.28	15.14
10→5	131	36.27	37.35	16.22	1.19	1.83	29.55	-26.93	1.35	17.84
20→5	132	37.68	34.47	19.33	1.18	1.79	28.35	-25.56	1.39	19.11
20→10	124	28.88	28.25	10.21	1.08	1.65	26.01	-23.92	1.16	10.66
30→5	119	25.92	25.85	11.10	1.00	1.53	25.69	-25.39	0.94	7.29
30→10	138	25.94	24.65	9.55	0.98	1.51	25.22	-23.92	1.00	7.93
30→15	138	33.67	31.28	12.64	1.11	1.69	25.23	-23.81	1.19	12.76

Table S4 - Windows. Window-level matched comparisons for GPT-5.1 across nine windows(n = 9). Replicate generations within each window are reduced to a single window mean. A win is counted when the configuration's mean ROI exceeds the matched unfiltered baseline. Sign-test and Wilcoxon signed-rank p-values are exact and two-sided. These window-level tests avoid the pseudo-replication of pairing replicate-group means and provide the conservative basis for significance claims in the manuscript.

Configuration	Window wins	Sign-test p	Wilcoxon p
20→10	8/9	0.039	0.027
10→5	7/9	0.180	0.098
10→3	7/9	0.180	0.098
20→5	8/9	0.039	0.129
30→15	6/9	0.508	0.910
5→3	3/9	0.508	0.359
30→10	4/9	1.000	0.250
Ranked	4/9	1.000	0.250
30→5	0/9	0.004	0.004

S5. Component ablation

Table S5a. *Component ablation on GPT-5.1, configuration 20→10. All arms hold the candidate documents fixed.*

Arm	Runs	Windows	Mean_ROI	Median_ROI	Mean arm tokens
Full two-stage pipeline	35	9	52.42	47.93	9,620
Sequential, reports shown, no elimination	35	9	49.33	47.96	9,615
Single prompt, full candidate pool	3	3	49.24	55.65	17,315
Single prompt over the candidate pool	32	9	49.03	47.30	21,401
Single prompt, budget-matched documents	32	9	47.18	45.86	19,198
Hold the entire candidate pool	32	9	46.32	42.64	9,792

Sequential, financial reports shuffled	**35**	**9**	**43.12**	**39.91**	**9,609**

Tokens are the mean arm-specific prompt tokens per run, excluding the Stage-One and report-retrieval calls shared by every arm, and the three sequential arms are matched to within 0.1%.

Table S5b. *Full pipeline versus each ablation arm, paired by window.*

Comparison arm	Paired windows	Mean ROI difference (pp)	Windows full pipeline wins	Wilcoxon p
Single prompt over the candidate pool	**9**	**3.31**	**6/9**	**0.25**
Single prompt, full candidate pool	**3**	**14.19**	**2/3**	**0.50**
Sequential, reports shown, no elimination	**9**	**3.70**	**5/9**	**0.57**
Sequential, financial reports shuffled	**9**	**9.33**	**9/9**	**0.00**
Single prompt, budget-matched documents	**9**	**5.01**	**6/9**	**0.05**
Hold the entire candidate pool	**9**	**5.61**	**7/9**	**0.05**

S6. Prompts

The prompts are provided as templates, with values in braces substituted at run time.

Stage One - news-based candidate selection

```
You are an expert investment analyst.

Using ONLY the data provided below, clipped to information available on or before {news_end_date},
build the best long-term investment portfolio for the
'{keyword}' sector.

INVESTMENT AMOUNT: ${investment_amount:,.2f}
REINVESTMENT AMOUNT: ${reinvestment_amount:,.2f}

{macro_context or ''}

{'=' * 60}
NEWS DATA ({len(news_df)} articles in the analysis window)
{'=' * 60}
{articles_text}

{'=' * 60}
FINANCIAL & PERFORMANCE DATA (full {len(comprehensive_data)}-ticker universe)
{'=' * 60}
{data_block}

{extra_data_text or ''}
{additional_context or ''}

INSTRUCTIONS:
- Treat {news_end_date} as the hard as-of date.
- Do not infer, cite, or rely on events, prices, financial statements, analyst views, or company outcomes after {news_end_date}.
- You have visibility into the provided temporally clipped news and financial data only.
- Do not use your training memory, general market memory, or known later stock performance to fill gaps.
- If you cannot justify a company from the supplied as-of evidence, do not select it.
- Select the companies with the strongest investment case based on every data point above.
- You may include between 3 and 15 companies.
- Use only tickers from this universe: {', '.join(tickers)}.
```

- Allocations must sum to exactly 100 %.

Return ONLY a JSON object in this exact format:

```
{{
  "investment_strategy": "Overall strategy description",
  "final_recommendations": [
    {{
      "ticker": "TICKER",
      "company_name": "Company Name",
      "final_allocation": 15.5,
      "volatility_assessment": "LOW/MODERATE/HIGH",
      "growth_assessment": "STRONG/GOOD/MODEST/DECLINING",
      "speed_assessment": "FAST/STEADY/SLOW/NEGATIVE",
      "news_factor": "How news influenced selection",
      "allocation_rationale": "Why this allocation percentage",
      "overall_score": 8.2
    }}
  ],
  "portfolio_summary": {{
    "total_allocation": 100.0,
    "expected_return": "8-12%",
    "risk_level": "Moderate",
    "strategy_focus": "Main investment thesis"
  }},
  "analysis_summary": {{
    "companies_analyzed": {len(comprehensive_data)},
```

Stage Two / final portfolio selection

Create a complete investment strategy using all the analyzed data.

```
    {macro_context if macro_context else ""}

    MARKET ANALYSIS FROM NEWS:
    - Key Themes: {market_analysis.get('key_themes', [])}
    - Investment Outlook: {market_analysis.get('investment_outlook', 'N/A')}
    - Major Trends: {market_analysis.get('major_trends', [])}

    COMPREHENSIVE COMPANY DATA:
    {data_summary}

    {prior_perf_block}

    INVESTMENT PARAMETERS:
    - Total Investment: ${investment_amount:,.2f}
    - Reinvestment Amount: ${reinvestment_amount:,.2f}

    {final_selection_instructions}

    Return complete strategy as JSON:
    {{
      "investment_strategy": "Overall strategy description",
      {ranked_schema}
      "final_recommendations": [
        {{
          "ticker": "TICKER",
          "company_name": "Company Name",
          "final_allocation": 15.5,
          "volatility_assessment": "LOW/MODERATE/HIGH",
          "growth_assessment": "STRONG/GOOD/MODEST/DECLINING",
          "speed_assessment": "FAST/STEADY/SLOW/NEGATIVE",
          "news_factor": "How news influenced selection",
          "allocation_rationale": "Why this allocation percentage",
          "overall_score": 8.2
        }}
```

```
  ],
  "portfolio_summary": {{
    "total_allocation": 100.0,
    "expected_return": "8-12%",
    "risk_level": "Moderate",
    "strategy_focus": "Main investment thesis"
  }},
  "analysis_summary": {{
    "companies_analyzed": {len(comprehensive_data)},
    "selection_criteria": "How companies were selected and weighted",
    "market_outlook": "Overall sector outlook"
  }}
}}
```

Unfiltered single-pass baseline

```
You are an expert investment analyst creating a complete long-term investment strategy.

    TASK: Analyze ALL provided data sources simultaneously to generate a final investment portfolio.

    SECTOR FOCUS: {keyword}
    INVESTMENT AMOUNT: ${investment_amount:,.2f}
    REINVESTMENT AMOUNT: ${reinvestment_amount:,.2f}

    {macro_context if macro_context else ""}

    NEWS DATA FOR ANALYSIS:
    {articles_text}

    {extra_data_text if extra_data_text else ""}

    {additional_context if additional_context else ""}

    INSTRUCTIONS:
    {first_step_instructions}

    First, provide a list of companies to analyze in this JSON format:
    {{
      "companies_for_analysis": [
        {{{first_step_schema_extra}"ticker": "AAPL", "company_name": "Apple Inc.", "rationale": "Why selected from news"}},
        {{{first_step_schema_extra}"ticker": "MSFT", "company_name": "Microsoft Corp.", "rationale": "Why selected from news"}}
      ],
      "market_analysis": {{
        "key_themes": ["theme1", "theme2", "theme3"],
        "investment_outlook": "Overall market outlook for the sector",
        "major_trends": ["trend1", "trend2", "trend3"]
      }}
    }}

    {first_step_return_note}
```

Integrated strategy generation

```
You are an expert financial analyst creating a final investment strategy by combining:
    1. News-based market analysis and company recommendations
    2. Detailed financial statement data

    NEWS ANALYSIS SUMMARY:
    - Market Themes: {news_analysis.get('analysis_summary', {}).get('market_themes', [])}
    - Key Trends: {news_analysis.get('analysis_summary', {}).get('key_trends', [])}
    - Companies Identified: {len(news_analysis.get('recommended_companies', []))}

    COMPREHENSIVE FINANCIAL & PERFORMANCE ANALYSIS:
    {combined_analysis}
```

INVESTMENT PARAMETERS:
- Initial Investment: ${investment_amount:,.2f}
- Reinvestment Amount: ${reinvestment_amount:,.2f}

Create a refined investment strategy that:
1. Uses the provided VOLATILITY, GROWTH, and SPEED summaries as primary decision factors
2. Allocates MORE to stocks with: Low volatility + Strong growth + Fast/steady speed
3. Allocates LESS to stocks with: High volatility + Declining growth + Negative momentum
4. Validates recommendations using the comprehensive financial & performance analysis
5. Provides final portfolio allocation percentages (must sum to 100%)
6. Base allocation decisions primarily on the summarized assessments provided

Return a JSON object with this structure:

```
{{
  "final_strategy": {{
    "allocation_methodology": "How volatility, growth, and speed summaries influenced allocations",
    "key_insights": ["insight based on comprehensive analysis"],
    "risk_assessment": "Overall portfolio risk based on combined analysis",
    "implementation_approach": "How to implement this strategy"
  }},
  "final_recommendations": [
    {{
      "ticker": "TICKER",
      "company_name": "Company Name",
      "final_allocation": 15.5,
      "volatility_assessment": "LOW/MODERATE/HIGH based on summary",
      "growth_assessment": "STRONG/GOOD/MODEST/DECLINING based on summary",
      "speed_assessment": "FAST/STEADY/SLOW/NEGATIVE based on summary",
      "allocation_rationale": "Why this allocation based on volatility + growth + speed combination",
      "overall_score": 8.2,
      "implementation_notes": "Specific notes for this holding"
    }}
  ],
  "portfolio_summary": {{
```

News relevance screening (system message)

You are an EXTREMELY STRICT financial news curator for SERIOUS STOCK INVESTMENT ANALYSIS. Only include articles that directly impact investment decisions.

STRICT CRITERIA - INCLUDE ONLY IF:
1. DIRECT STOCK/MARKET IMPACT: Article directly mentions stock prices, market movements, earnings, or financial performance
2. MAJOR COMPANY NEWS: Significant corporate developments (mergers, acquisitions, leadership changes, product launches) for publicly traded companies
3. ECONOMIC INDICATORS: Federal Reserve decisions, inflation data, GDP reports, unemployment that directly affect markets
4. SECTOR ANALYSIS: Tech sector growth, financial sector regulations, industry trends affecting multiple stocks
5. STARTUP/IPO NEWS: Only if discussing funding rounds, IPOs, or acquisition potential
6. FINANCIAL TECHNOLOGY: Only fintech innovations that could disrupt existing financial institutions

ABSOLUTELY EXCLUDE:
- General tech news without financial/business angle
- Politics unless directly market-moving (like trade policy)
- Opinion pieces or analysis without concrete data
- Entertainment, sports, lifestyle, health, crime, weather
- General business news without stock market relevance
- Articles about individuals unless they're major company executives
- Regulatory news unless it directly affects specific industries/stocks
- General economic news without clear market implications
- Any article that doesn't help make specific investment decisions

BE EXTREMELY SELECTIVE. Only 20-40% of articles should pass this filter.

Return ONLY the numbers (indices) of relevant articles, separated by commas.
Example: 1,3,5,7,12

News relevance screening (user message)

Filter these articles for STRICT investment relevance:\n\n{articles_text}