vingegpt
stock market investing with AI

VingeGPT Quarterly Benchmark Report

AUGUST 2025

A comparative analysis of VingeGPT performance across general and custom intelligence models

El Mowafak SALIM
Quality Assurance Engineer
Maw.salim@nextgenvi.com

NextGen Value Investing Sàrl
LHoFT - Luxembourg House of Fintech
9, rue du Laboratoire, L-1911 Luxembourg

Executive Summary

This quarterly benchmarking report evaluates VingeGPT against 14 other leading AI models, both general-purpose and finance-specific, using a rigorous methodology with fresh sessions, identical prompts, and four test sets covering data reliability, analytical breadth, and an end-to-end equity research case study. The aim of this report is to assess VingeGPT's performance and determine whether it demonstrates a clear and sustainable advantage over competing solutions.

Results show that VingeGPT achieved the highest overall score across depth of analysis, clarity & readability, and unique contributions. It consistently delivered structured, concise, and investor-friendly outputs, combining factual accuracy with clear synthesis across a broad spectrum of analytical tasks—from company-specific valuation to macroeconomic information and portfolio diagnostics. Its ability to integrate quantitative and qualitative insights, and present coherent, actionable conclusions set it apart from other models. In responsiveness, VingeGPT ranked in the moderate range (4–6 seconds), providing a balance between speed and thoughtful output. This pace remains acceptable for professional investor interactions, where accuracy, completeness, and clarity of insights are prioritized over instantaneous replies.

In conclusion, while leading general-purpose systems such as ChatGPT 4o and Gemini Pro demonstrated strong reasoning, and finance-specific models like FinChat offered niche strengths, none matched VingeGPT's combination of breadth, analytical structure, and tailored relevance for investors.

1 Introduction

In light of the exceptionally rapid pace of innovation in artificial intelligence (AI) and the continual release of new models globally, a systematic and recurring evaluation process is essential to ensure both relevance and competitive performance for VingeGPT.

Since February 2025, we have implemented a comprehensive quarterly benchmarking program for VingeGPT, designed to measure its capabilities relative to the latest general-purpose and domain-specific AI systems. For clarity, the term VingeGPT-4o is used throughout this report to denote VingeGPT's configuration built on the OpenAI 4o engine.

The present report documents the results of the latest evaluation as part of our ongoing quality assurance framework.

2 Objectives

The overarching aim of this quarterly report is twofold. First, it seeks to provide a comparative performance assessment of VingeGPT-4o in relation to other leading AI models currently available.

VingeGPT-4o itself integrates approximately 150 pages of specialised value investing expertise, a corpus exceeding 30 million data points, covers 58 stock markets with over 40.000 publicly listed securities, other aggregated and curated data sources, and a carefully calibrated blend of custom instructions designed to serve investors on a global scale.

Second, the report investigates whether VingeGPT-4o, as a custom GPT, demonstrates a distinctive and measurable competitive advantage over alternative solutions in the market, including both general-purpose systems and those tailored specifically to the finance sector.

3 Benchmarking scope

For this evaluation, VingeGPT-4o was benchmarked against a representative sample of contemporary AI models spanning different categories.

The August 2025 benchmarking compared VingeGPT-4o against the following models:

- **OpenAI General-Purpose Models :** GPT-4o, GPT-o3 Advanced Reasoning, GPT-o4-mini, GPT-4.5
- **Other General-Purpose AI Models :** Google Gemini 2.5 Flash, Google Gemini 2.5 Pro, DeepSeek Standard, DeepSeek DeepThink, Anthropic Claude Sonnet 4, Perplexity AI, Grok AI
- **Finance-Specific Custom AI Models :** FinChat, WarrenAI, InvestingAI

4 Methodology

To ensure methodological rigor and eliminate potential confounding factors such as prompt warm-up effects or residual session memory, each model was tested within fresh, isolated sessions using identical baseline text inputs. This approach ensured that no prior context could influence the results and that all systems operated under strictly comparable conditions.

The evaluation framework comprised 4 distinct sets of tests, each designed to assess complementary aspects of model performance :

- The first set focused on data reliability and accuracy, evaluating each model's ability to retrieve and present verifiable information. For data reliability, models that provided accurate and up-to-date figures in both tests received 15 points. Models that returned outdated data in both the initial and retest phases were assigned 5 points, while the single model that was outdated in the initial test but corrected its data in the retest received an intermediate score.
- The second set consisted of a diverse range of prompts related to multiple dimensions of stock analysis, including valuation metrics, industry analysis, and qualitative risk factors, thereby simulating the varied nature of investor queries. Models were ranked on depth of analysis and clarity & readability, with the highest-performing model in each category receiving 15 points and the lowest receiving 1 point, and all others scored proportionally between these two extremes.
- The third set concentrated on a comprehensive investment analysis of a single company, Nike Inc., to assess the models' ability to integrate data, perform multi-layered analysis, and deliver a coherent, end-to-

end investment process. For this set of tests, similar to the second set of tests, models were ranked on depth of analysis and clarity & readability, with the highest-performing model in each category receiving 15 points and the lowest receiving 1 point, and all others scored proportionally between these two extremes.

- The fourth set measured response times, capturing how quickly each model generated answers across all prompt categories. This analysis provided insights into performance efficiency and latency, both of which can influence the practical usability of an AI tool in real-world investment contexts. Responsiveness was scored using a simple formula starting at 15 points, from which the model's average response time in seconds was subtracted, ensuring that faster models received higher scores while slower models were proportionally penalized.

Figure 1 : Overview of 4 test sets

While every effort was made to maintain neutrality in prompt selection, it is acknowledged that the design process may inevitably reflect unconscious biases shaped by our expertise in value investing. A detailed table listing the exact prompts used in this evaluation is presented in the following sections.

5 Future Benchmarking Enhancements

In future editions of this report, the benchmarking framework will be expanded to assess each model's geographic market coverage and multilingual interaction capabilities. This will include evaluating the extent to which models incorporate equities and market data from global stock exchanges beyond the United States, as well as their ability to accurately process and respond in the most widely spoken languages worldwide.

6 Benchmarking test results

6.1 Evaluation of Data Reliability and Accuracy

The first series of tests for both general-purpose and finance-focused custom intelligence models was designed to assess the **accuracy and consistency of financial data retrieval**. As a benchmark, we used Nike's latest financial statements and more specifically the balance sheet published on June 26th, 2025, covering the fiscal year ended May 31st, 2025.

Two standardised prompts were submitted to each model on July 17th, 2025:

1. *"Show me the latest balance sheet of Nike."*
2. *"What is the date of the balance sheet?"*

By this date, three weeks had passed since the official release of the results, allowing sufficient time for accurate data to be incorporated into reliable sources. For each model, we recorded:

- The balance sheet date, which should correspond to May 31st, 2025.
- The reported figures for three key line items: **Total assets**, **Cash and cash equivalents**, and **Total current liabilities**.

The results are summarised in the tables 1 & 2 below. The most accurate and reliable performers were VingeGPT-4o, ChatGPT-4o, ChatGPT-o3, ChatGPT-o4-mini, ChatGPT-4.5, Google Gemini 2.5 Flash, Google Gemini 2.5 Pro, Grok AI, FinChat, Warren AI, and Investing AI, returning both the correct balance sheet date and precise values for the three metrics.

By contrast, Anthropic Claude Sonnet 4, Perplexity AI, DeepSeek Standard, and DeepSeek Deep returned outdated financials, in some cases from the previous fiscal year or earlier quarters. A retest conducted on August 10th, 2025, showed that only Perplexity AI had updated its data. Anthropic Claude Sonnet 4, DeepSeek Standard, and DeepSeek Deep continued to return outdated figures.

Model	Reliability of Data	Balance Sheet Date	Outdated Data
VingeGPT-4o	Accurate and reliable. Figures match the official 10-K report: Total Assets: $36,579M, Cash & Cash Equivalents: $7,464M, Total Current Liabilities: $10,566M.	May 31, 2025	No
ChatGPT 4o	Accurate and reliable. Figures match the official 10-K report: Total Assets: $36,579M, Cash & Cash Equivalents: $7,464M, Total Current Liabilities: $10,566M.	May 31, 2025	No
ChatGPT o3 (advanced reasoning)	Accurate and reliable. Figures match the official 10-K report: Total Assets: $36,579M, Cash & Cash Equivalents: $7,464M, Total Current Liabilities: $10,566M.	May 31, 2025	No
ChatGPT o4-mini (fast at advanced reasoning)	Accurate and reliable. Figures match the official 10-K report: Total Assets: $36,579M, Cash & Cash Equivalents: $7,464M, Total Current Liabilities: $10,566M.	May 31, 2025	No
ChatGPT 4.5 (good for writing and exploring ideas)	Accurate and reliable. Figures match the official 10-K report: Total Assets: $36,579M, Cash & Cash Equivalents: $7,464M, Total Current Liabilities: $10,566M.	May 31, 2025	No
Google Gemini 2.5 Flash	Accurate and reliable. Figures match the official 10-K report: Total Assets: $36,579M, Cash & Cash Equivalents: $7,464M, Total Current Liabilities: $10,566M.	May 31, 2025	No
Google Gemini 2.5 Pro (advanced reasoning)	Accurate and reliable. Figures match the official 10-K report: Total Assets: $36,579M, Cash & Cash Equivalents: $7,464M, Total Current Liabilities: $10,566M.	May 31, 2025	No
Anthropic Claude Sonnet 4	Outdated data. Provided balance sheet data for the previous fiscal year (May 31, 2024).	May 31, 2024	**Yes**
Perplexity AI	Outdated data. Provided balance sheet data for the previous fiscal year (May 31, 2024).	May 31, 2024	**Yes**
DeepSeek Standard	Outdated data. Provided balance sheet data for the third quarter of the previous fiscal year (February 29, 2024).	February 29, 2024	**Yes**
DeepSeek DeepThink	Outdated data. Provided balance sheet data for the third quarter of the previous fiscal year (February 29, 2024).	February 29, 2024	**Yes**
Grok AI	Accurate and reliable. Figures match the official 10-K report: Total Assets: $36,579M, Cash & Cash Equivalents: $7,464M, Total Current Liabilities: $10,566M.	May 31, 2025	No
FinChat	Accurate and reliable. Figures match the official 10-K report: Total Assets: $36,579M, Cash & Cash Equivalents: $7,464M, Total Current Liabilities: $10,566M.	May 31, 2025	No

Warren AI	Accurate and reliable. Figures match the official 10-K report: Total Assets: $36,579M, Cash & Cash Equivalents: $7,464M, Total Current Liabilities: $10,566M.	May 31, 2025	No
Investing AI	Accurate and reliable. Figures match the official 10-K report: Total Assets: $36,579M, Cash & Cash Equivalents: $7,464M, Total Current Liabilities: $10,566M.	May 31, 2025	No

Table 1 : Evaluation results of data reliability and accuracy

Model with outdated data during initial test	Balance sheet date (initial test July 17th, 2025)	Balance sheet date (retest August 10th, 2028)
Anthropic Claude Sonnet 4	May 31, 2024	August 31, 2024
Perplexity AI	**May 31, 2024**	**May 31, 2025**
DeepSeek Standard	February 29, 2024	May 31, 2024
DeepSeek DeepThink	February 29, 2024	February 29, 2024

Table 2 : Retest results conducted on August 10th, 2025

6.2 Evaluation of Analytical Breadth in Equity Research & End-to-End Equity Research Case Study – Nike Inc.

The second series of tests examined each model's ability to address a diverse and representative set of equity research tasks**.** The prompt set was deliberately varied in scope, combining macroeconomic assessment, company-specific valuation, strategic and competitive analysis, portfolio evaluation, and market composition queries. This breadth was designed to simulate the complex, multi-topic conversations that retail and professional investors frequently conduct when researching equities.

By encompassing multiple dimensions of analysis this test set assessed not only factual accuracy, but also analytical breadth, adaptability, and synthesis capability. The variety of prompts offered a comprehensive view of how each model performs when switching rapidly between distinct analytical contexts.

Between mid-July and the end of July 2025, the following prompts were submitted to each model:

1. *Summarize the current macroeconomic environment. show the most recent datapoints in a comprehensive table and provide a one sentence assessment for each metric. add to each metric description its series ID between brackets*
2. *Is Coca Cola currently undervalued or overvalued?*
3. *What is the intrinsic value of Procter?*
4. *Can you perform a strategic analysis of Unilever?*
5. *Who are the competitors of Unilever?*
6. *What is the current share price of Microsoft?*
7. *Calculate the historical intrinsic value of Microsoft in 2016 and 2018?*
8. *Study the following portfolio: 25% of Microsoft, 20% of Apple, and the rest in Louis Vuitton MC.PA*
9. *What is the appropriate cost of capital for Unilever?*
10. *How has the growth of Unilever been over the last 3 years?*
11. *What is the geographical exposure of QQQ?*
12. *What are the top holdings of QQQ?*
13. *What is the intrinsic value of Tata Consulting India?*
14. *How many markets and companies do you cover?*

The third series of tests assessed each model's ability to conduct a full end-to-end investment analysis within the context of a natural, conversational exchange. The objective was to replicate how an investor might interact with an AI assistant from the initial query through to a final investment decision, while progressively deepening the analysis.

By structuring the test as a continuous dialogue, this assessment measured not only factual accuracy and analytical depth, but also the ability to maintain contextual continuity across multiple topics, integrate quantitative and qualitative information, and deliver a coherent, investor-ready conclusion.

Between mid-July and the end of July 2025, the following 15 prompts were submitted to each model:

1. *Good morning, how are you?*
2. *Who has created you?*

3. *Ok. So now I would like to analyze a company I am interested in. What do you suggest?*
4. *So, the company is Nike.*
5. *Perform a fundamental analysis first.*
6. *So, what do you think about their ROIC and dividend payout ratio?*
7. *Ok. Can you calculate the IV of the company? Analyze if I have a safety margin on the current share price.*
8. *Can you analyze Skechers side-by-side to Nike?*
9. *And what is the employee and customer sentiment for both companies?*
10. *Show me the dividend history for Nike*
11. *Ok interesting. and can you share the latest insider trades with me?*
12. *What can you tell me about the industry?*
13. *So given those elements I prefer to invest in Nike. Before doing that can you let me know what the auditor's opinion is and if there are any disagreements with management?*
14. *Are there any signs of earnings manipulation?*
15. *Can you analyze the audit fees as well?*

Results from the second set (Evaluation of Analytical Breadth in Equity Research) and the third set (End-to-End Equity Research Case Study – Nike Inc.) have been consolidated for joint analysis.

To ensure objectivity in the interpretation and analysis of the raw results, we initially replicated the approach used in previous quarterly benchmarks by submitting the findings to Microsoft Copilot in Office 365, selected as a neutral model outside the scope of this benchmark. However, the model demonstrated limitations in processing and interpreting the more than 300 pages of raw data.

We therefore transitioned to ChatGPT 4o, providing it with the complete dataset along with a detailed analytical framework. The instructions specified that the document compared 15 different AI models, with only raw result data for each model and no preliminary scoring. ChatGPT 4o was tasked with evaluating the models based on their financial literacy, their effectiveness in supporting investors throughout the investment process, and the ease of use they offer to investors. The same set of prompts, categorised as T2 and T3, had been submitted to each of the 15 models.

The analysis framework required the assessment model to identify:

1. The model delivering the greatest depth of analysis.
2. The model that best supports investors by providing information that is readable, synthetic, and clearly structured.
3. Any unique contributions offered by each model. Unique contributions capture any capabilities, insights, or analytical approaches that materially enhance an investor's decision-making process beyond standard data retrieval. This may include proprietary valuation models, forward-looking scenario analysis, sector-specific intelligence, integration of alternative data (e.g., ESG scores, supply chain data), interactive portfolio simulations, or the ability to synthesise multi-source inputs into actionable recommendations.

The results were to be presented in a table with five columns: the full name of each model, its score for depth of analysis, its score for clarity and readability, its score for unique contributions, and the total score across all categories. Scoring followed a comparative ranking method, assigning 15 points to the highest-performing model in each category and 1 point to the lowest.

The detailed multi-step instruction prompt is provided in the Appendices, along with a screenshot of the results generated by ChatGPT 4o.

Model (full name)	Depth of analysis	Clarity & readability	Unique contributions	Total
VingeGPT-4o	15	14	15	44
ChatGPT 4o	14	15	9	38
Google Gemini 2.5 Pro	13	10	11	34
ChatGPT 4.5	12	13	8	33
Anthropic Claude Sonnet 4	11	12	10	33
Perplexity AI	7	9	13	29
FinChat	6	7	14	27
ChatGPT o4-mini	8	11	7	26
DeepSeek DeepThink	10	8	5	23
ChatGPT o3 (advanced reasoning)	9	6	6	21
Grok AI	3	3	12	18
DeepSeek Standard	5	4	3	12
Warren AI	4	2	4	10
Google Gemini 2.5 Flash	1	5	2	8
Investing AI	2	1	1	4

Table 3 : Evaluation results for the 15 models in Test Sets 2 and 3

Based on the total points from all three criteria (depth of analysis, clarity & readability and unique contributions), the best overall model in the T2 & T3 benchmark is VingeGPT-4o. It edges out others by having consistently high scores in all three categories, with particular strength in clarity/readability and unique contributions (like FRED series IDs, portfolio concentration metrics, and solvency/profitability panels).

Model (full name)	Depth of analysis - Summary	Clarity & readability - Summary	Unique contributions - Summary	Negative points - Summary	Total evaluation points
VingeGPT-4o	Strong macro & portfolio analysis with relevant metrics; slightly less narrative depth than 4.5.	Highly structured tables, concise interpretations, clear visual hierarchy.	FRED series IDs, portfolio concentration, solvency/profitability rollups.	Slightly less narrative depth than top-ranked for analysis; may rely heavily on tabular presentation.	44
ChatGPT 4o	Solid, consistent analysis across prompts, though less expansive than 4.5.	Top-tier clarity with clean tables and concise bullet summaries.	Balanced macro & portfolio dashboards with quick takeaways.	Less detail than 4.5 in deep dives; some macro explanations can be overly concise.	38
Google Gemini 2.5 Pro	Very strong depth, especially on complex reasoning tasks.	Good structure though sometimes verbose.	Detailed breakdowns and multi-perspective analysis.	Verbose at times; structure can feel dense.	34
Anthropic Claude Sonnet 4	High-detail outputs with strong diagnostic coverage.	Clear and well-organized narrative style.	Actionable rebalancing advice with explicit allocation targets.	Slightly less concise; may include more context than needed for quick scanning.	33
ChatGPT 4.5	Deepest analysis with extensive detail and explanations.	Readable but longer-form; may require scanning to extract key points.	Comprehensive narrative insights and thematic exploration.	Long-form answers may require effort to distill key points.	33
Perplexity AI	Good breadth and factual accuracy, slightly less depth than top-tier models.	Very readable dashboards with bullet-point assessments.	Macro dashboards with one-sentence insights.	Lacks the analytical depth of leaders; focuses more on presentation than deep reasoning.	29
FinChat	Limited scope; covers essentials without extended reasoning.	Functional clarity but minimal structure.	Investor-focused tone with conservative recommendations.	Lacks extended reasoning; basic presentation style.	27
ChatGPT o4-mini	Surface-level responses; focuses on brevity over depth.	Readable but minimal structuring; concise to a fault.	Speed-focused, light-touch summaries.	Overly brief; lacks depth and detailed structuring.	26
DeepSeek DeepThink	Decent depth; more thoughtful than Standard version.	Moderately clear, but more text-heavy.	Deeper reasoning sequences and context awareness.	Can be wordy; structure less polished than top clarity models.	23
ChatGPT o3	Well-reasoned and comprehensive, slightly behind top three in breadth.	Organized structure but more text-heavy than 4o or VingeGPT.	Nuanced reasoning paths and scenario considerations.	Text-heavy format can reduce quick readability; slightly behind top tier in breadth.	21
Grok AI	Mid-tier depth; balanced but not standout.	Readable, some structural consistency.	Occasional creative framing of investment themes.	Inconsistent depth and clarity; occasional meandering narrative.	18
DeepSeek Standard	Limited depth; covers basics but lacks sophistication.	Readable but basic formatting.	Straightforward, no-frills summaries.	Very basic analysis; minimal added value beyond essentials.	12
Warren AI	Minimal analytical depth; generic outputs.	Plain text, low readability for quick scanning.	Basic investor guidance but lacks specifics.	Generic and vague; poor structure and low insight density.	10
Google Gemini 2.5 Flash	Adequate factual coverage but less detailed than Pro variant.	Moderately structured with mixed formatting consistency.	Quick, high-level outputs with minimal elaboration.	Shallow compared to Pro; uneven formatting quality.	8
Investing AI	Shallow analysis with little beyond surface metrics.	Poor clarity; unstructured responses.	Generic statements with no distinctive features.	Minimal substance; unstructured, generic outputs.	4

Table 4 : Textual evaluation results for the 15 models in Test Sets 2 and 3

6.3 Response times

We have also recorded **initial response times** for each group of prompts and each model as well, which measures how long it takes from prompt submission to initial output—to better evaluate both performance and responsiveness. The table below shows the response times for each of the 15 models analyzed.

AI model (initial time in seconds)	T1	T1	T2.1	T2.2	T2.3	T2.4	T2.5	T2.6	T2.7	T2.8	T2.9	T2.10	T2.11	T2.12	T2.13	T2.14	T3	T3	T3	T3	T3	T3	T3	T3	T3	T3	T3	T3	T3	T3	minimum	maximum	average
VingeGPT-4o	4	3	4	5	2	2	9	5	7	3	3	6	5	3	3	5	2	4	3	2	5	3	3	4	6	3	7	3	6	4	5		2	9	4
ChatGPT 4o	3	3	5	4	4	3	1	3	5	6	2	2	3	3	3	1	2	1	2	2	3	3	2	5	3	3	8	2	6	3	4		1	8	3
ChatGPT o3 (advanced reasoning)	4	5	3	2	2	2	9	2	2	3	2	#	5	7	4	4	2	5	4	6	5	6	6	#	#	4	8	5	5	#	#	6	2	13	5
ChatGPT o4-mini (fast at advanced reasoning)	3	4	2	2	2	2	4	2	3	2	2	2	3	3	3	3	2	1	2	2	2	3	3	3	5	5	5	3	4	4	4	5	8	1	8	3
ChatGPT 4.5 (good for writing and exploring ideas)	3	3	7	4	4	4	4	3	4	5	4	2	#	3	4	3	2	2	2	3	2	2	2	2	2	4	4	3	7	3	2	4	3	4	2	22	4
Google Gemini 2.5 Flash	6	3	4	3	4	3	3	3	3	6	2	2	2	4	2	2	2	4	3	2	3	2	3	3	3	3	5	3	3	3	3		2	6	3
Google Gemini 2.5 Pro (advanced reasoning)	8	4	3	3	2	3	7	2	5	4	3	2	#	8	#	#	8	9	#	#	#	7	#	#	#	8	5	9	#	#	#		2	27	10
Anthropic Claude Sonnet 4	2	1	3	4	3	2	3	2	3	3	4	3	4	4	2	2	3	3	2	6	6	3	2	4	3	3	3	3	3	4	4		1	6	3
Perplexity AI	3	3	5	3	3	4	5	5	4	3	3	3	2	3	4	5	1	2	1	3	3	2	2	4	3	3	1	4	3	3	2		1	5	3
DeepSeek Standard	4	2	#	2	3	5	2	5	3	3	2	4	4	4	5	4	4	7	5	5	5	5	6	5	6	6	5	5	4	8	5		2	12	5
DeepSeek DeepThink	4	3	1	1	1	2	1	1	2	3	1	2	4	4	6	5	5	5	#	6	6	#	#	#	6	7	6	#	#	#	9		1	14	5
Grok AI	3	2	#	4	3	6	5	3	#	#	3	2	4	5	4	5	1	4	4	5	1	3	4	5	4	3	3	2	4	5	4		1	22	5
FinChat	#	4	7	5	4	4	7	8	6	4	6	6	5	7	4	3	3	5	3	#	9	8	7	#	#	7	7	#	#	#	9		3	21	8
Warren AI	9	5	3	8	4	3	#	6	5	4	6	8	3	5	4	7	7	3	2	7	5	9	2	#	8	3	5	8	5	5	8		2	11	6
Investing AI	5	3	3	2	2	3	4	1	2	3	2	2	2	3	4	5	3	6	4	4	7	6	8	8	#	7	5	#	8	5	5	#	1	15	5

Table 5 : Evaluation of responsiveness results for the 15 models across Test Sets 1, 2 and 3

The fastest models in this benchmark, such as ChatGPT 4o, Google Gemini 2.5 Flash, Claude Sonnet 4, and Perplexity AI, average around 3 seconds per response. This timing is close to a thoughtful pause in human conversation, allowing for a smooth, natural flow without feeling rushed or abrupt. For investor-facing interactions, this speed maintains engagement while conveying that the system is “thinking” before answering.

Models in the 4–6 second range, like VingeGPT-4o, DeepSeek Standard, Grok AI, and Warren AI, mimic the rhythm of a human who is considering their answer carefully. While still within an acceptable range for professional dialogue, these delays may be noticeable in rapid Q&A sessions. In advisory contexts, this timing can reinforce the perception of depth if the answer quality justifies the extra wait.

Slower models, averaging 8–10 seconds such as FinChat and Gemini Pro, feel more like a human pausing to take notes or look up data. This is acceptable when delivering highly detailed or analytical responses, especially in long-form financial guidance. However, for ongoing conversational exchanges, this speed risks breaking the flow unless paired with clear signals that the additional time is delivering greater value.

7 Conclusion

Results show that VingeGPT-4o achieved the highest overall score across depth of analysis, clarity & readability, and unique contributions. It consistently delivered structured, concise, and investor-friendly outputs, combining factual accuracy with clear synthesis across a broad spectrum of analytical tasks—from company-specific valuation to macroeconomic dashboards and portfolio diagnostics. Its ability to integrate quantitative and qualitative insights, and present coherent, actionable conclusions set it apart from other models.

Model (full name)	Depth of analysis	Clarity & readability	Unique contributions	Data reliability	Response times	Total
VingeGPT-4o	15	14	15	15	11	70
ChatGPT 4o	14	15	9	15	12	65
ChatGPT 4.5	12	13	8	15	11	59
Google Gemini 2.5 Pro	13	10	11	15	5	54
ChatGPT o4-mini	8	11	7	15	12	53
Perplexity AI	7	9	13	10	12	51
FinChat	6	7	14	15	7	49
ChatGPT o3 (advanced reasoning)	9	6	6	15	10	46
Anthropic Claude Sonnet 4	11	12	10	5	5	43
Grok AI	3	3	12	15	10	43
DeepSeek DeepThink	10	8	5	5	10	38
Google Gemini 2.5 Flash	1	5	2	15	12	35
Warren AI	4	2	4	15	9	34
Investing AI	2	1	1	15	10	29
DeepSeek Standard	5	4	3	5	10	27

Table 6 : Global results for the 15 models across all test sets

While top general-purpose systems like ChatGPT 4o and Gemini Pro demonstrated strong reasoning, and some finance-specific models like FinChat offered niche strengths, none matched VingeGPT-4o's balance of breadth, structure, and tailored investor relevance.

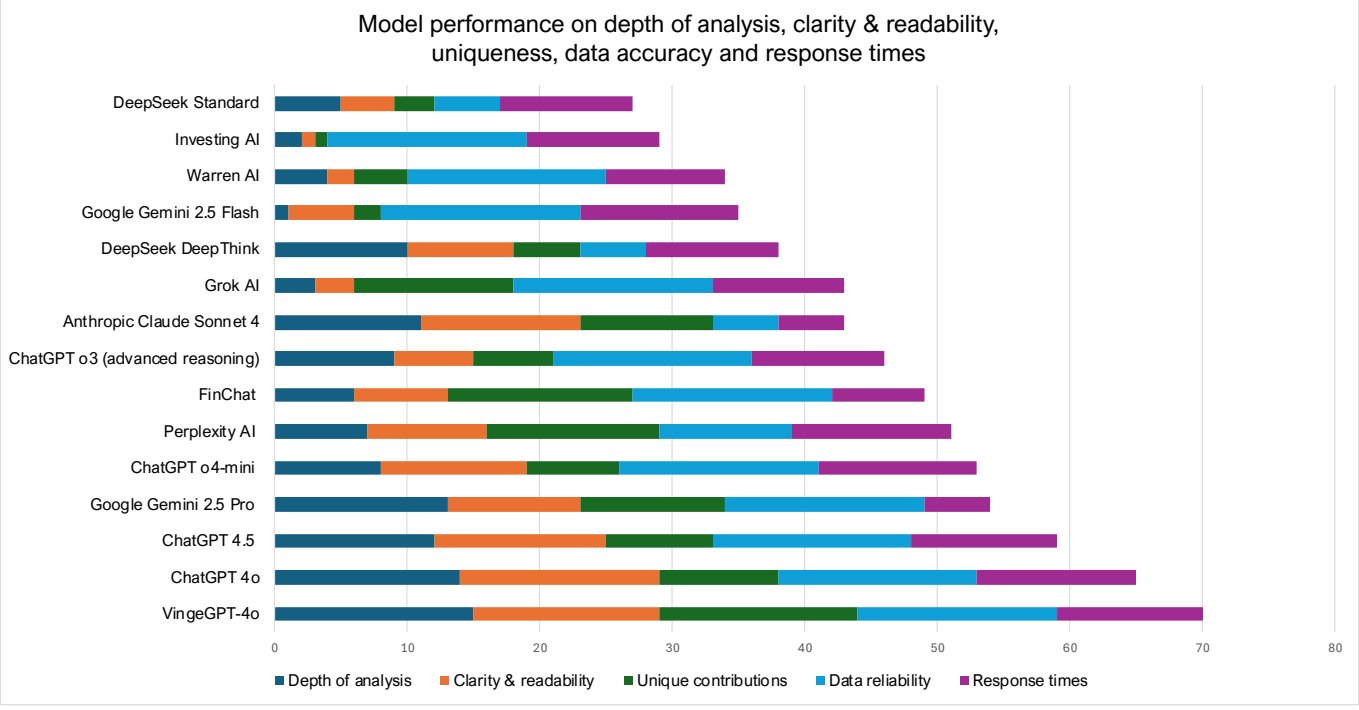

Figure 2 : Model performance on depth of analysis, clarity & readability, uniqueness, data accuracy and response times

8 Appendix

8.1 Appendix 1 - Prompt used for the analysis

“this document compares 15 different AI models amongst them. the document contains only raw result data for each of the 15 models without any preliminary scoring. I want you to compare the financial literacy of the models, how good they are helping and supporting investors in their investment process and also being able to evaluate how easy to use they are for investors. we have submitted the same prompts categorized as T2 and T3 prompts to each of the 15 models. please analyse the results for each model and determine 1) which is the model with the best depth of analysis, 2) the model that supports best investors in their investment process by being most readable, synthetic and providing readable structured information, 3) assess for each model if they provide any unique contributions. provide the results in a table with 5 columns: 1st column being the full name of each model analysed. so that users can identify the model, 2nd column being the depth of analysis, 3rd column clarity & readibility, 4th column unique contributions and 5th the sum of points of each category. in order to grade the models on each of these 3 attributes being depth of analysis, clarity & readibility and unique contributions, give to the best model 15 points in each category and 1 point for the lowest.”

Appendix 2 – Screenshot of ChatGPT 4o analysis of raw results

Full comparative table (15 models, 1 = lowest, 15 = best)

#	Model (full name)	Depth of analysis	Clarity & readability	Unique contributions*	Total
1	VingeGPT	**15**	14	**15**	**44**
2	ChatGPT 4o	14	**15**	9	38
3	Google Gemini 2.5 Pro	13	10	11	34
4	ChatGPT 4.5	12	13	8	33
5	Anthropic Claude Sonnet 4	11	12	10	33
6	Perplexity AI	7	9	13	29
7	FinChat	6	7	14	27
8	ChatGPT o4-mini	8	11	7	26
9	DeepSeek DeepThink	10	8	5	23
10	ChatGPT o3 (advanced reasoning)	9	6	6	21
11	Grok AI	3	3	12	18
12	DeepSeek Standard	5	4	3	12
13	Warren AI	4	2	4	10
14	Google Gemini 2.5 Flash	1	5	2	8
15	Investing AI	2	1	1	**4**

*“Unique contributions” captures anything materially different the model adds for an investor (live quotes, embedded sources, stylistic innovations, risk-scenario simulators, etc.).