vinley
AI for serious investors

VingeGPT quarterly benchmark report

DECEMBER 2025

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

El Mowafak SALIM
Quality Assurance Engineer

Candi CARRERA
Co-founder & CEO

NextGen Value Investing Sàrl
LHoFT - Luxembourg House of Fintech
9, rue du Laboratoire, L-1911 Luxembourg
contact@nextgenvi.com

Abstract

AI is moving fast, with new models launching constantly, so VingeGPT needs a structured and repeatable way to stay current and competitive. Since February 2025, we've run a quarterly benchmarking program to measure VingeGPT's performance against leading general-purpose and finance-focused models.

In this report, VingeGPT-5 refers to VingeGPT's configuration built on the OpenAI 5 engine. The results presented here come from the latest test cycle conducted between December 10th and 31st, 2025, and the next benchmark will cover Vinley, the upcoming version of VingeGPT expected in February 2026, with its final configuration to be confirmed in the next report. The benchmark also aims to verify whether VingeGPT shows a clear, measurable advantage as a custom GPT versus alternative market solutions.

This report benchmarks VingeGPT-5 against 13 other AI models across general-purpose and finance-specialized categories to assess how well they perform in investor-oriented workflows. To keep results fair and comparable, every model was tested in a fresh, isolated session using the same standardized prompts, preventing warm-up effects or session memory from influencing outputs. The evaluation uses multiple test sets designed to measure complementary dimensions such as accuracy, analytical depth, breadth of coverage, multilingual inclusiveness, global market coverage, and responsiveness.

Models are scored using a relative ranking method: for each criterion, the best model receives 14 points and the lowest receives 1 point, with others ranked in between. Exception to this are the data accuracy & realibility tests, as the multilinguality & market coverage that allow ex-aequos. The overall score is the unweighted sum of the criterion scores, giving a transparent comparison across analytical rigor, usability, and differentiation.

One of the main while surprising findings was that only 6 of the 14 models achieved full marks for data accuracy and reliability, while the others returned outdated financial information, ranging from being off by one quarter to more than a full year. The detailed model-by-model results and evidence are documented inside the report & the raw data appendix.

Overall, results show a clear performance hierarchy. VingeGPT ranks as the strongest all-around model, combining up-to-date financial accuracy, deep and broad financial analysis, and strong multilingual and global market coverage. While not the fastest, its response time remains competitive given the completeness of its outputs, making it the most robust option for professional and institutional investors.

A second tier is formed by the Grok models: Grok (Fast) stands out for speed and strong depth, while Grok (Expert) offers a more balanced trade-off between breadth, depth, and efficiency, well-suited for users who prioritize responsiveness without sacrificing rigor.

Several models perform well but show constraints that reduce suitability for globally diversified investment work. OpenAI ChatGPT-5 and ChatGPT-5 Thinking are strong in breadth and multilinguality, but are limited by data freshness and weaker non-US coverage. Perplexity and Gemini (Thinking) provide good synthesis and context, but are held back by older financial data or fewer distinctive analytical strengths.

Lower-tier models may offer niche value, such as conceptual depth, creative angles, or a specific investing narrative, but their weaker reliability on data accuracy and market coverage makes them less dependable for professional investment decision-making.

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 the capabilities of our specialised AI relative to the latest general-purpose and domain-specific AI models. For clarity, the term VingeGPT-5 is used throughout this report to denote VingeGPT's configuration built on the OpenAI 5 engine.

The overarching aim of this quarterly report is twofold. First, it seeks to provide a comparative performance assessment of VingeGPT in relation to other leading AI models currently available. VingeGPT itself integrates approximately 170 pages of specialised value investing expertise, a corpus exceeding 30 million data points, covers 60 stock markets with over 42.000 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.

The present report documents the results of the latest evaluation executed between December 10th and 31st, 2025, as part of our ongoing quality assurance framework. Readers should already consider that the next benchmark report will cover Vinley, the new version of VingeGPT, that runs on Microsoft Co-pilot together with Open AI as underyling subprocessor and that will be released in February 2026. We will confirm the exact configuration for Vinley in the next benchmark report.

Second, the report investigates whether VingeGPT, 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.

2 Methodology

2.1 Models

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

The benchmarking compared VingeGPT-5 against the following models:

- **OpenAI General-Purpose Models :** OpenAI ChatGPT-5, OpenAI ChatGPT-5 Thinking
- **Other General-Purpose AI Models :** Google Gemini fast, Google Gemini thinking, DeepSeek Standard DeepSeek DeepThink, Anthropic Claude Sonnet 4.5, Perplexity AI, Grok fast, Grok expert
- **Finance-Specific Custom AI Models :** FiscalAI (formerly known as Finchat), WarrenAI, InvestingAI

2.2 Model warm-up & session memory

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.

2.3 Test sets

The evaluation framework for this report consists of various sets of tests, each designed to assess the complementary aspects of each model and elements like performance, accuracy, breadth & depth of analysis:

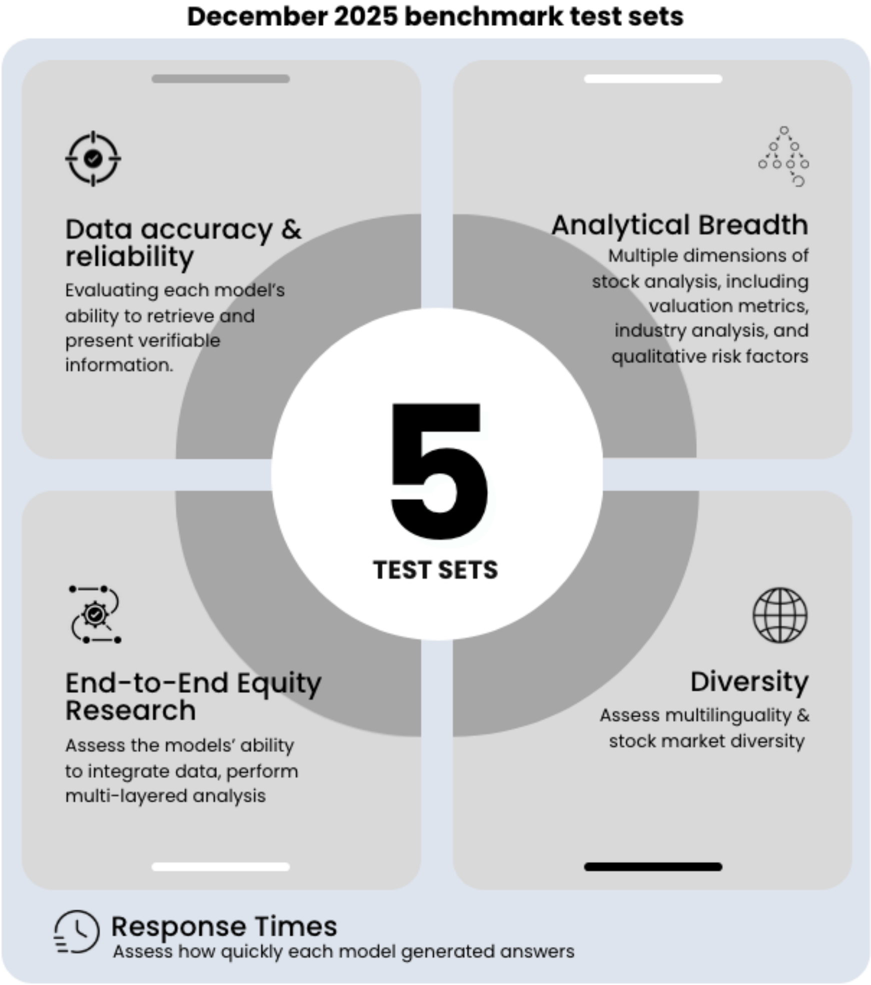

December 2025 benchmark test sets
Data accuracy & reliability
Evaluating each model's ability to retrieve and present verifiable information.
Analytical Breadth
Multiple dimensions of stock analysis, including valuation metrics, industry analysis, and qualitative risk factors
5
TEST SETS
End-to-End Equity Research
Assess the models' ability to integrate data, perform multi-layered analysis
Diversity
Assess multilinguality & stock market diversity
Response Times
Assess how quickly each model generated answers

Figure 1 : Methodological overview

2.3.1 Test set 1 – Data accuracy & reliability

The first test set comprising 2 prompts, focused on data reliability and accuracy, evaluating each model’s ability to retrieve and present verifiable information.

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 **October 1st, 2025, covering the quarter Q1 of the 2026 fiscal year ended August 31st, 2025.**

Two standardised prompts were submitted to each model around mid of December 2025:

- *“Show me the latest balance sheet of Nike.”*
- *“What is the date of the balance sheet?”*

By the date of mid-December 2025, 10 weeks had passed since the official release of the results on October 1st, 2025, which shall allow sufficient time to all AI models to incorporate accurate and updated data.

For each model, we recorded:

- The balance sheet date, which should correspond to **August 31st, 2025.**
- The reported figures for four key line items: **Cash and cash equivalents**, **Total current assets**, **Long-term debt** and **Shareholders equity with expected values being respectively 7024, 23898, 7996 & 13468.**

For data reliability, models that provided accurate and up-to-date figures received 100% of the points. Models that returned outdated data of maximum 2 quarters received 50% of points, while models showing more than 1 year of outdated data received 0 points.

2.3.2 Test set 2 – analytical breadth

The second test set comprising 14 prompts, consists 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 14 points and the lowest receiving 1 point, and all others scored proportionally between these two extremes.

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.

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?*

Table 1 : List of T2 prompts

2.3.3 Test set 3 – End-to-end equity research

The third set comprising 15 prompts, 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.

This test set shall assess 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. with the highest-performing model in each category receiving 14 points and the lowest receiving 1 point, and all others scored proportionally between these two extremes.

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?*

Table 2 : List of T3 prompts

2.3.4 Test set 4 – Diversity

The fourth test set comprising 10+3 prompts focuses on multilinguality of each model and checks the diversity of the investment universe. The languages tested were English, Arabic, French, Chinese (Mandarin) and Spanish. For the stock market diversity, we assessed the availability of company general information and the balance sheet for the companies Reliance Industries Ltd (listed on the Bombay Stock Exchange) and Mitsubishi Corporation (listed on the Tokyo Stock Exchange)

The selection of the companies has been done based on the market cap. Both companies were at the moment of testing amongst the Top3 market caps on each respective stock exchange.

2.3.5 Test set 5 – Response times

The fifth test set, measures response times for all previous test sets, 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 14 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.

2.4 Potential bias in test sets & conclusion

While every effort is 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 and specialised AI models.

In order to minimize the risk of biasing the conclusion, as for every benchmark report, we do not write the conclusion ourselves but provide the raw results together with a detailed chain of thought to OpenAI to summarize the conclusions. Appendix 1 provides the detailed Chain of Thought (CoT) used for the quarterly benchmark report.

2.5 Scoring methodology and interpretation

The scoring framework employed in this study is based on a comparative, multi-criteria evaluation design intended to assess the suitability of large language models for investor-oriented applications. Scores reflect relative performance within the evaluated cohort rather than absolute or intrinsic model quality. All models were assessed under identical experimental conditions, including fresh sessions, standardized prompts, and controlled test sets, to minimize carryover effects and contextual bias.

For each criterion, models were ranked using a relative ordinal scale, with the highest-performing model assigned 14 points and the lowest-performing model assigned 1 point. Intermediate models received proportionally ranked scores. The overall score for each model represents the unweighted sum of the three criterion scores, providing a composite indicator of relative performance across analytical rigor, usability, and differentiation. This scoring approach is intended to support transparent comparison while acknowledging that individual use cases may warrant alternative weightings.

3 Results

3.1 Test set 1 – Data accuracy & reliability

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 **October 1st, 2025, covering the quarter Q1 of the 2026 fiscal year ended August 31st, 2025.**

Two standardised prompts were submitted to each model around mid of December 2025:

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

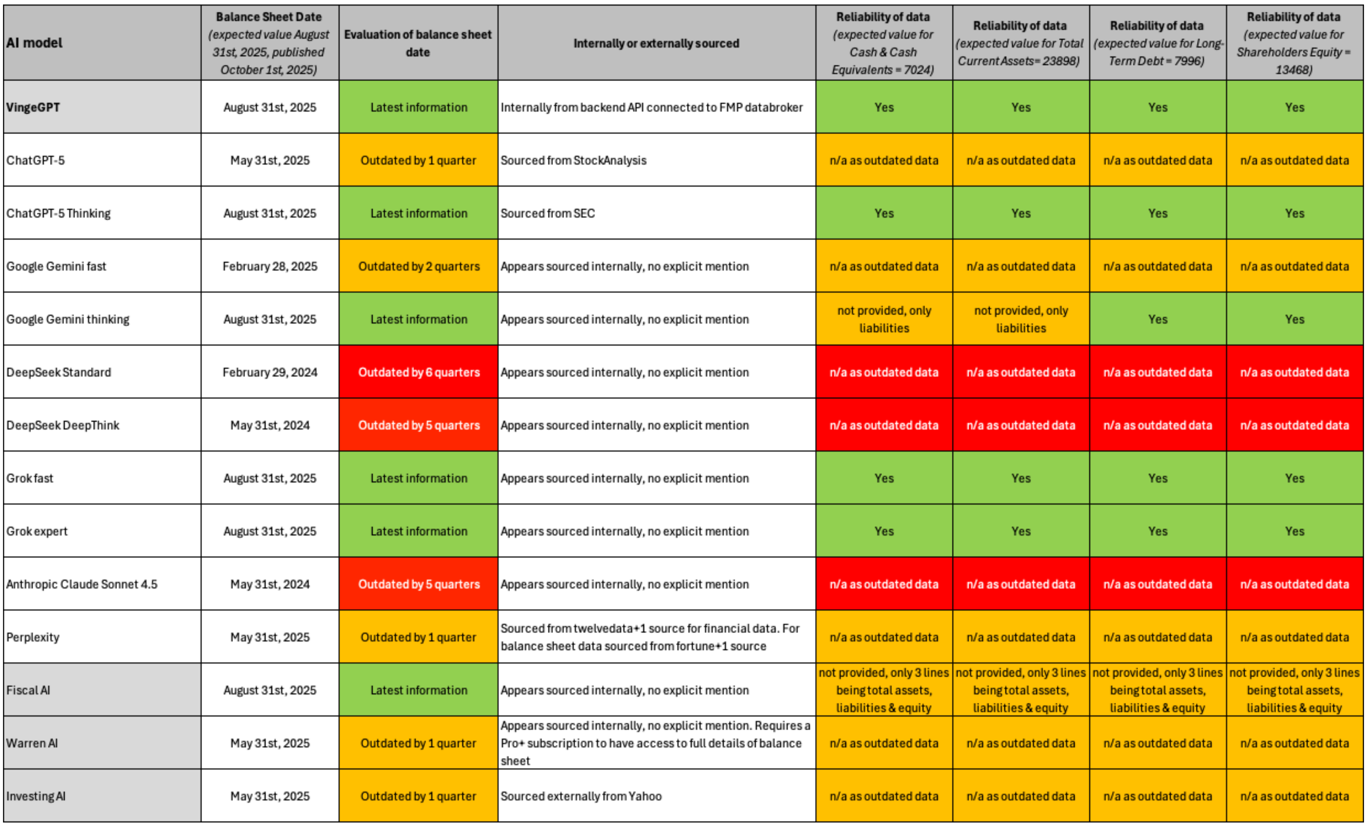
By the date of mid-December 2025, 10 weeks had passed since the official release of the results on October 1st, 2025, which shall allow sufficient time to all AI models to incorporate accurate data.

For each model, we recorded:

- The balance sheet date, which should correspond to **August 31st, 2025.**
- The reported figures for four key line items: **Cash and cash equivalents**, **Total current assets**, **Long-term debt** and **Shareholders equity with expected values being respectively 7024, 23898, 7996 & 13468.**

The results are summarised in the table 1 below. The only accurate and reliable performers were VingeGPT, ChatGPT 5 Thinking, Google Gemini Thinking, and both Grok models Grok Fast and Grok Expert returning both the correct balance sheet date and the correct values for the four financial metrics. These results are extremely surprising but confirm a tendency that we have observed over the last months that more & more models focus on searching data on the Internet but do not really consider if the data makes sense in terms of accuracy & reliability to support investing processes & potentially related decisions.

By contrast and in stark quality decrease versus the previous quarterly benchmark, 8 models out of 13 failed to provide the most up-to-date information. The worst performer again has been DeepSeek with their models showing financial information outdated by 6 respectively 5 quarters (represents 1 ½ years ot outdated data) up to a couple of models that include this quarter ChatGPT-5 with outdated data between 1 to 2 quarters.



AI model	**Balance Sheet Date** *(expected value August 31st, 2025, published October 1st, 2025)*	**Evaluation of balance sheet date**	**Internally or externally sourced**	**Reliability of data** *(expected value for Cash & Cash Equivalents = 7024)*	**Reliability of data** *(expected value for Total Current Assets= 23898)*	**Reliability of data** *(expected value for Long-Term Debt = 7996)*	**Reliability of data** *(expected value for Shareholders Equity = 13468)*
VingeGPT	August 31st, 2025	Latest information	Internally from backend API connected to FMP databroker	Yes	Yes	Yes	Yes
ChatGPT-5	May 31st, 2025	Outdated by 1 quarter	Sourced from StockAnalysis	n/a as outdated data	n/a as outdated data	n/a as outdated data	n/a as outdated data
ChatGPT-5 Thinking	August 31st, 2025	Latest information	Sourced from SEC	Yes	Yes	Yes	Yes
Google Gemini fast	February 28, 2025	Outdated by 2 quarters	Appears sourced internally, no explicit mention	n/a as outdated data	n/a as outdated data	n/a as outdated data	n/a as outdated data
Google Gemini thinking	August 31st, 2025	Latest information	Appears sourced internally, no explicit mention	not provided, only liabilities	not provided, only liabilities	Yes	Yes
DeepSeek Standard	February 29, 2024	Outdated by 6 quarters	Appears sourced internally, no explicit mention	n/a as outdated data	n/a as outdated data	n/a as outdated data	n/a as outdated data
DeepSeek DeepThink	May 31st, 2024	Outdated by 5 quarters	Appears sourced internally, no explicit mention	n/a as outdated data	n/a as outdated data	n/a as outdated data	n/a as outdated data
Grok fast	August 31st, 2025	Latest information	Appears sourced internally, no explicit mention	Yes	Yes	Yes	Yes
Grok expert	August 31st, 2025	Latest information	Appears sourced internally, no explicit mention	Yes	Yes	Yes	Yes
Anthropic Claude Sonnet 4.5	May 31st, 2024	Outdated by 5 quarters	Appears sourced internally, no explicit mention	n/a as outdated data	n/a as outdated data	n/a as outdated data	n/a as outdated data
Perplexity	May 31st, 2025	Outdated by 1 quarter	Sourced from twelvedata+1 source for financial data. For balance sheet data sourced from fortune+1 source	n/a as outdated data	n/a as outdated data	n/a as outdated data	n/a as outdated data
Fiscal AI	August 31st, 2025	Latest information	Appears sourced internally, no explicit mention	not provided, only 3 lines being total assets, liabilities & equity	not provided, only 3 lines being total assets, liabilities & equity	not provided, only 3 lines being total assets, liabilities & equity	not provided, only 3 lines being total assets, liabilities & equity
Warren AI	May 31st, 2025	Outdated by 1 quarter	Appears sourced internally, no explicit mention. Requires a Pro+ subscription to have access to full details of balance sheet	n/a as outdated data	n/a as outdated data	n/a as outdated data	n/a as outdated data
Investing AI	May 31st, 2025	Outdated by 1 quarter	Sourced externally from Yahoo	n/a as outdated data	n/a as outdated data	n/a as outdated data	n/a as outdated data

Table 3 : Analysis of data reliability & accuracy

3.2 Test set 2 – Analytical breadth

The T2 prompt set assessed each model's capacity to deliver a broad, structured, and methodologically sound financial analysis, covering core financial statements, performance drivers, valuation considerations, and risk factors. Performance varied substantially across models, with clear differentiation between those capable of integrating multiple analytical frameworks coherently and those providing only partial or generic coverage. The strongest performers demonstrated not only completeness across income statement, balance sheet, and cash flow analysis, but also the ability to contextualize financial metrics within valuation and risk perspectives without excessive verbosity.

A second tier of models delivered generally competent but less comprehensive analyses, often addressing the main financial components while lacking integration, prioritization, or specificity. Lower-performing models exhibited narrower analytical scope, fragmented reasoning, or reliance on high-level financial commentary that limited their usefulness for professional investment decision-making. Overall, the T2 results highlight that breadth of analysis is not merely a function of length, but of structured coverage, relevance, and the ability to synthesize financial dimensions into a coherent analytical narrative.

The model-by-mode results in alphabetical order :

- **Anthropic Claude Sonnet 4.5 :** Claude Sonnet 4.5 provided moderate breadth, touching on most financial analysis components at a conceptual level. However, it often emphasized narrative explanation over structured financial coverage. This resulted in breadth that is broad in topic range but it tried to do a bit of everything in its responses.
- **DeepSeek (DeepThink) :** DeepSeek (DeepThink) exhibited strong analytical breadth, addressing multiple financial dimensions with notable conceptual depth. It incorporated valuation logic and risk considerations effectively. Despite its breadth, reliability issues related to outdated data that could again be observe in T2 prompt set and outside T2 limit its professional applicability.
- **DeepSeek (Standard) :** DeepSeek (Standard) showed narrow analytical breadth, covering only selected financial metrics and concepts. Important components of comprehensive financial analysis were missing or weakly developed. As a result, its T2 performance ranks among the lowest in the cohort.
- **Fiscal AI :** Fiscal AI delivered limited breadth, focusing on basic financial elements without deeper integration or expansion into valuation and risk. Several analytical dimensions were either underdeveloped or omitted. The breadth is insufficient for professional-grade financial analysis.
- **Investing AI :** Investing AI demonstrated limited breadth by covering key financial statements and basic performance indicators. However, its analysis often lacked integration across financial dimensions and provided limited context often claiming that it was lacking the necessary information. The breadth is not sufficient for advanced analysis.
- **Google Gemini (Thinking) :** Google Gemini (Thinking) provided balanced and consistent coverage of major financial analysis components. While it addressed income, balance sheet, and cash flow considerations, its treatment of valuation and risk was more descriptive than analytical. The overall breadth is adequate but not distinctive.
- **Google Gemini (Fast) :** Gemini (Fast) delivered partial financial coverage, focusing primarily on headline metrics and high-level trends. Several analytical dimensions were addressed only superficially, limiting overall breadth. The model prioritizes responsiveness over comprehensive financial framing.
- **Grok (Expert) :** Grok (Expert) delivered strong breadth across income statement, balance sheet, and cash flow analysis, with clear identification of performance drivers. It incorporated valuation and risk considerations more consistently than its fast counterpart. The analysis was broad without becoming diffuse, supporting structured financial interpretation.
- **Grok (Fast) :** Grok (Fast) provided selective but focused financial coverage, emphasizing core drivers and key metrics rather than exhaustive statement-by-statement analysis. While it addressed the main analytical dimensions, some secondary aspects such as valuation framing or balance sheet nuance were less developed. Its breadth is adequate but optimized for speed rather than completeness.
- **OpenAI ChatGPT-5 :** OpenAI ChatGPT-5 demonstrated solid breadth, covering the major components of financial analysis with reasonable consistency. Its responses generally addressed financial statements and drivers but occasionally relied on high-level explanations. The breadth is suitable for general analysis, though less rigorous than top-tier performers.
- **OpenAI ChatGPT-5 Thinking :** The Thinking variant showed good analytical coverage, particularly in integrating multiple financial dimensions within a single response. It addressed valuation drivers and risk considerations adequately. However, some areas lacked prioritization, slightly diluting analytical focus.
- **Perplexity :** Perplexity offered good breadth with a strong emphasis on synthesis and cross-referencing across financial dimensions. It covered core financial and contextual factors, though some responses leaned toward summarization rather than structured financial decomposition or even for some questions, the model could not answer. Its analytical breadth is average.
- **VingeGPT :** VingeGPT demonstrated the most comprehensive breadth of financial analysis, consistently covering all major financial statements, key performance drivers, valuation considerations, and risk factors. Its

responses were well-structured and avoided generic filler, integrating multiple analytical perspectives into a coherent framework. This level of breadth is well-aligned with professional and institutional investment analysis standards.
- **Warren AI :** Warren AI referenced key financial concepts, coverage of formal statements and valuation mechanics was limited due to the fact that many datapoints require a Pro or Pro+ subscription. Its breadth is selective and subscription dependent.

3.3 Test set 3 – End-to-end equity research

The T3 prompt set evaluated each model's capacity to deliver deep, company-specific financial analysis focused on Nike, extending beyond standard financial metrics to include advanced and granular information such as auditor-related disclosures, insider activity, governance considerations, and filing-level details. This test set explicitly distinguished between models capable of anchoring their analysis in concrete, company-specific facts and those relying primarily on generic corporate finance narratives.

High-performing models demonstrated not only conceptual understanding but also operational depth, accurately contextualizing Nike's financial structure, governance, and disclosures within a coherent analytical framework.

The results reveal a clear stratification across models. A leading group consistently addressed both standard and advanced Nike-specific queries with precision and completeness, while a second tier provided partial depth but exhibited gaps in harder prompts or a tendency toward generalization.

Lower-performing models, despite sometimes demonstrating strong conceptual finance knowledge or creative insight, failed to deliver sufficiently granular, verifiable, and Nike-specific information, limiting their usefulness for professional equity analysis. Overall, the T3 results emphasize that analytical depth in company analysis depends on specificity, factual anchoring, and the ability to engage with detailed corporate disclosures rather than on response length or narrative sophistication.

The model-by-mode results in alphabetical order:

- **Anthropic Claude Sonnet 4.5 :** Claude Sonnet 4.5 demonstrated systematically textual ballooning of its responses, overshooting the expected answers. Advanced prompts were addressed at a high level without sufficient reference to concrete disclosures. This resulted in depth that was more interpretative than analytical.
- **DeepSeek (DeepThink) :** DeepSeek (DeepThink) showed exceptional analytical depth on Nike, particularly in addressing advanced and less commonly requested information. It demonstrated strong conceptual and operational understanding of company disclosures and governance. Despite broader reliability issues linked to outdated data that is NOT being evaluated in T3 prompts set, its T3 depth ranks among the highest.
- **DeepSeek (Standard) :** DeepSeek (Standard) showed uneven Nike-specific depth, with some competent responses alongside notable omissions (cf. insider trades) potentially suffering from extremely outdated information. While certain prompts were addressed adequately, advanced disclosure-related questions were weak. Overall depth was inconsistent and limited.
- **Fiscal AI :** Fiscal AI demonstrated adequate and consistent depth across the full set of Nike-specific T3 prompts, successfully addressing both standard financial questions and more detailed company-related inquiries. Its responses were coherent, company-anchored, and sufficiently detailed to meet the objectives of the depth analysis. While not differentiated by exceptional granularity or distinctive analytical angles, its overall depth was appropriate and reliable for professional company-level analysis.
- **Google Gemini (Fast) :** Gemini (Fast) exhibited limited depth in Nike analysis, focusing primarily on high-level company information. Advanced prompts related to auditor opinion, earnings manipulation and audit fees analysis were only partially addressed or treated superficially. This constrained its usefulness for detailed company-specific investigation.
- **Google Gemini (Thinking) :** Google Gemini (Thinking) provided adequate but unexceptional Nike-specific depth. It addressed most T3 prompts competently but often provided generic responses or even outdated forensic or filing-level analysis. The model's depth is adequate for general equity research but less suited to detailed due diligence.
- **Grok (Expert) :** Grok (Expert) delivered consistently deep Nike-specific analysis, covering both financial and non-financial dimensions with precision. It demonstrated strong command of advanced disclosure topics and integrated them into a coherent analytical narrative. Its depth was comprehensive without being excessively verbose. It only missed an answer on T3.15 related to the analysis of audit fees.
- **Grok (Fast) :** Grok (Fast) achieved a similar depth of analysis in the T3 set as Grok (Expert), while providing shorter answer and effectively responding to advanced Nike-specific prompts including governance and disclosure-related topics. It provided concrete, focused answers with minimal dilution. The model's depth is particularly well-suited for rapid, detailed company-level analysis.
- **Investing AI :** Investing AI exhibited limited depth in its Nike-specific analysis across the T3 prompt set. While some standard financial questions were addressed, the model frequently provided generic responses or failed to

directly answer more detailed and advanced prompts. As a result, its analysis lacked sufficient specificity and completeness to meet the requirements of a rigorous company-level depth assessment.

- **OpenAI ChatGPT-5 :** OpenAI ChatGPT-5 provided moderate depth in Nike analysis, adequately addressing core company information and standard financial metrics. However, its responses to advanced prompts were occasionally incomplete or generalized (cf. insider trades, dividend history, Nike's industry). This limited its overall depth relative to top-tier models.
- **OpenAI ChatGPT-5 Thinking :** The Thinking variant showed improved reasoning structure but did not consistently translate this into deeper Nike-specific factual detail. Some advanced prompts were addressed conceptually rather than through concrete company-specific information. As a result, its depth was uneven across the T3 test set.
- **Perplexity :** Perplexity demonstrated moderate depth in Nike analysis, often synthesizing information across multiple corporate dimensions. It handled advanced prompts with moderate specificity, (cf. auditor opinion, earnings manipulation, audit fees analysis) with some responses relied on secondary summarization rather than direct disclosure-level detail. Overall, its depth is robust for simple questions but not uniformly precise.
- **VingeGPT :** VingeGPT demonstrated strong Nike-specific depth, consistently addressing both standard financial questions and more advanced prompts related to filings and disclosures. Its responses were anchored in company-specific detail and avoided generic commentary. Its depth was balanced, reliable, and professionally actionable with one miss on the audit fees analysis providing outdated information
- **Warren AI :** While it offered insightful observations, it provided limited responses with detailed Nike-specific disclosures (cf. auditor opinion, audit fees analysis, auditor opinion). Its depth is selective and narrative-driven rather than quantitative.

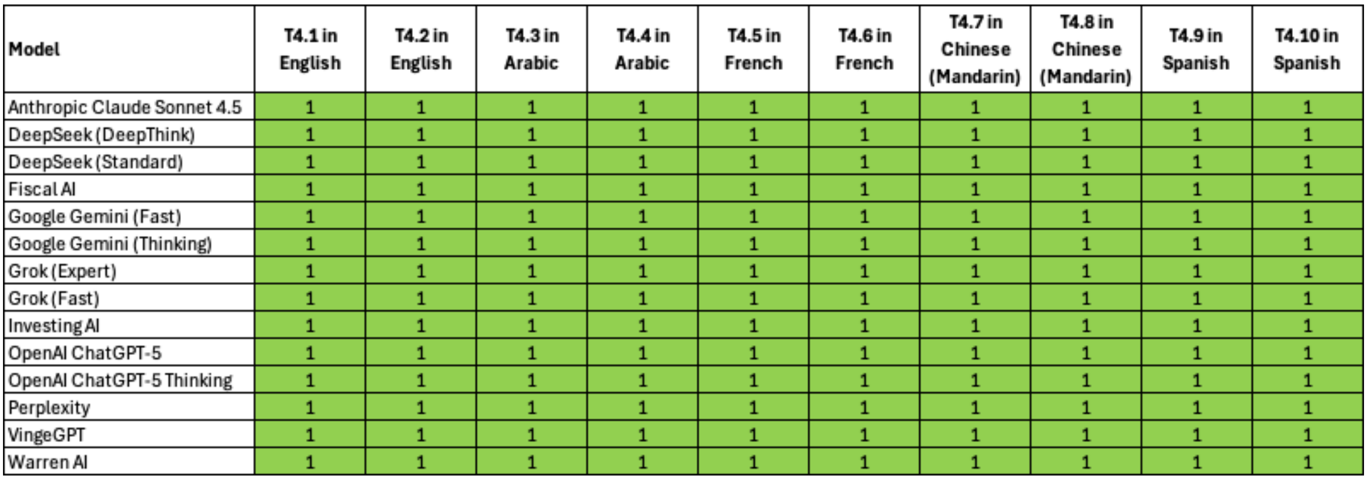
3.4 Test set 4 – Diversity

The T4 prompt set evaluated the multilingual and cross-market accessibility of each model by testing responses in English, Spanish, French, Mandarin Chinese, and Arabic, alongside the ability to retrieve company information from non-US equity markets.

All evaluated models demonstrated full multilingual capability, successfully generating coherent and contextually appropriate responses across the five tested languages, indicating a broad level of linguistic inclusiveness. In parallel, models were assessed on their capacity to provide company-level information for Reliance Industries Ltd (Bombay Stock Exchange) and Mitsubishi Corporation (Tokyo Stock Exchange).

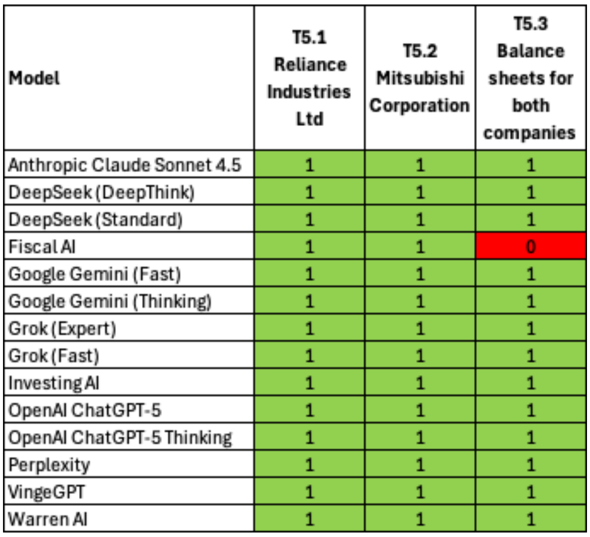
With one exception, all models were able to supply both descriptive company information and corresponding balance sheet data for these non-US issuers, reflecting adequate coverage of global equity markets. Fiscal AI constituted the sole exception, as it was unable to provide balance sheet information for either company, indicating a limitation in non-US financial data retrieval rather than in multilingual language generation.

The table below summarizes the results of the T4 prompts set :



Model	T4.1 in English	T4.2 in English	T4.3 in Arabic	T4.4 in Arabic	T4.5 in French	T4.6 in French	T4.7 in Chinese (Mandarin)	T4.8 in Chinese (Mandarin)	T4.9 in Spanish	T4.10 in Spanish
Anthropic Claude Sonnet 4.5	1	1	1	1	1	1	1	1	1	1
DeepSeek (DeepThink)	1	1	1	1	1	1	1	1	1	1
DeepSeek (Standard)	1	1	1	1	1	1	1	1	1	1
Fiscal AI	1	1	1	1	1	1	1	1	1	1
Google Gemini (Fast)	1	1	1	1	1	1	1	1	1	1
Google Gemini (Thinking)	1	1	1	1	1	1	1	1	1	1
Grok (Expert)	1	1	1	1	1	1	1	1	1	1
Grok (Fast)	1	1	1	1	1	1	1	1	1	1
Investing AI	1	1	1	1	1	1	1	1	1	1
OpenAI ChatGPT-5	1	1	1	1	1	1	1	1	1	1
OpenAI ChatGPT-5 Thinking	1	1	1	1	1	1	1	1	1	1
Perplexity	1	1	1	1	1	1	1	1	1	1
VingeGPT	1	1	1	1	1	1	1	1	1	1
Warren AI	1	1	1	1	1	1	1	1	1	1

Table 4 : Results of T4 prompts assessing multilinguality

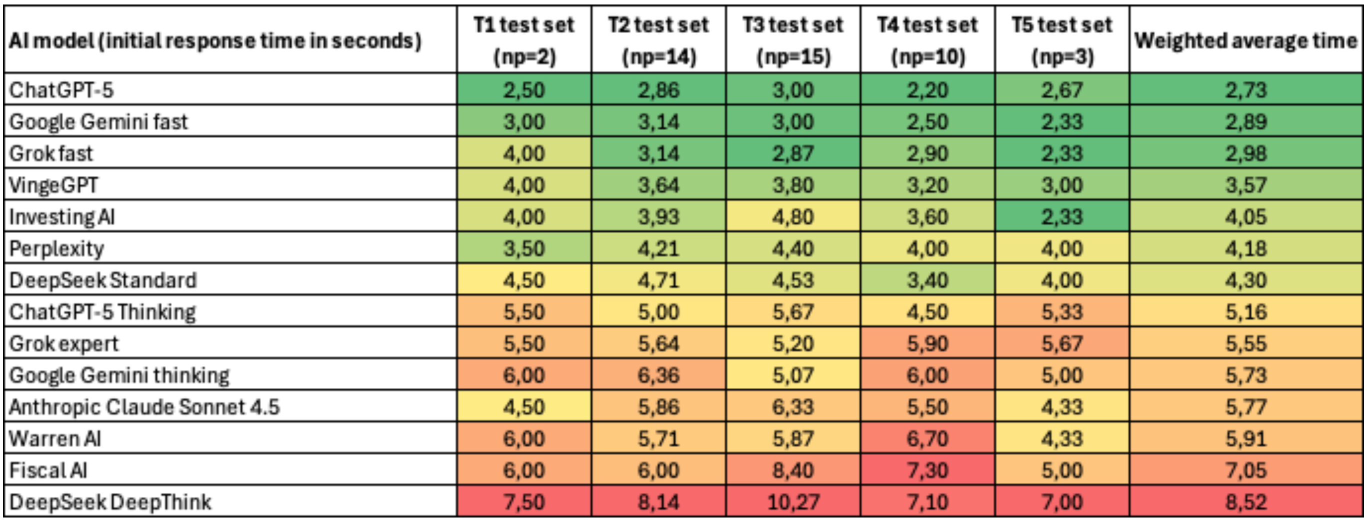


Model	T5.1 Reliance Industries Ltd	T5.2 Mitsubishi Corporation	T5.3 Balance sheets for both companies
Anthropic Claude Sonnet 4.5	1	1	1
DeepSeek (DeepThink)	1	1	1
DeepSeek (Standard)	1	1	1
Fiscal AI	1	1	0
Google Gemini (Fast)	1	1	1
Google Gemini (Thinking)	1	1	1
Grok (Expert)	1	1	1
Grok (Fast)	1	1	1
Investing AI	1	1	1
OpenAI ChatGPT-5	1	1	1
OpenAI ChatGPT-5 Thinking	1	1	1
Perplexity	1	1	1
VingeGPT	1	1	1
Warren AI	1	1	1

Table 5 : Results of T4 prompts assessing diverse market coverage

3.5 Test set 5 - 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 models analyzed.



AI model (initial response time in seconds)	T1 test set (np=2)	T2 test set (np=14)	T3 test set (np=15)	T4 test set (np=10)	T5 test set (np=3)	Weighted average time
ChatGPT-5	2,50	2,86	3,00	2,20	2,67	2,73
Google Gemini fast	3,00	3,14	3,00	2,50	2,33	2,89
Grok fast	4,00	3,14	2,87	2,90	2,33	2,98
VingeGPT	4,00	3,64	3,80	3,20	3,00	3,57
Investing AI	4,00	3,93	4,80	3,60	2,33	4,05
Perplexity	3,50	4,21	4,40	4,00	4,00	4,18
DeepSeek Standard	4,50	4,71	4,53	3,40	4,00	4,30
ChatGPT-5 Thinking	5,50	5,00	5,67	4,50	5,33	5,16
Grok expert	5,50	5,64	5,20	5,90	5,67	5,55
Google Gemini thinking	6,00	6,36	5,07	6,00	5,00	5,73
Anthropic Claude Sonnet 4.5	4,50	5,86	6,33	5,50	4,33	5,77
Warren AI	6,00	5,71	5,87	6,70	4,33	5,91
Fiscal AI	6,00	6,00	8,40	7,30	5,00	7,05
DeepSeek DeepThink	7,50	8,14	10,27	7,10	7,00	8,52

Table 6 : Evaluation of responsiveness results for the models across Test Sets 1 to 5

The fastest models in this benchmark, such as ChatGPT 5, Google Gemini Fast and Grok Fast, take around 3 seconds per initial response time. 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, DeepSeek Standard, Perplexity and WarrenAI, 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.

For VingeGPT, the results are normal and the 0,8 seconds incremental time compared to its underlying engine ChatGPT-5 are linked to the API calls done by VingeGPT to its data backend.

Slower models, averaging 6 seconds and beyond such as FiscalAI (formerly known as Finchat), DeepSeek DeepThink, 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.

4 Conclusion

The comparative evaluation of fourteen AI models across accuracy, analytical capability, inclusiveness, and responsiveness reveals a clear performance hierarchy. VingeGPT emerges as the strongest overall performer, combining fully up-to-date financial accuracy, maximal breadth of financial analysis, strong company-specific depth, and comprehensive multilingual and global market coverage.

Although not the fastest model, VingeGPT's response time remains competitive relative to its analytical completeness, positioning it as the most robust all-in-one solution for professional and institutional investors. The Grok variants constitute a strong second tier: Grok (Fast) excels in analytical depth and responsiveness, while Grok (Expert) provides a more balanced trade-off between breadth, depth, and efficiency. Together, these models are particularly well-suited for users prioritizing speed without materially sacrificing analytical rigor.

A second group of models demonstrates solid but constrained performance due to trade-offs in accuracy, global coverage, or depth. OpenAI ChatGPT-5 and its Thinking variant perform well in breadth and multilinguality, though limitations in data freshness and non-US financial coverage reduce their suitability for globally diversified investment analysis. Perplexity and Google Gemini (Thinking) offer strong contextual and synthetic capabilities but are penalized by outdated financial data or lack of distinctive analytical advantages.

Lower-tier models exhibit niche strengths—such as conceptual depth, creative perspectives, or investment-style narratives—but suffer from structural weaknesses, particularly in data accuracy and market coverage, which undermine their reliability for professional investment workflows.

Model	Response times	T1 Data accuracy & reliability	T2 Breadth of financial analysis	T3 Depth of analysis	Unique Contributions & Angles	T4 Multilinguality	T5 Stock Market Exchanges coverage	Total Score
VingeGPT	11	14	14	11	12	14	14	**90**
Grok (Fast)	12	14	8	14	9	14	14	**85**
Grok (Expert)	6	14	12	13	7	14	14	**80**
OpenAI ChatGPT-5 Thinking	7	14	13	8	8	14	14	**78**
OpenAI ChatGPT-5	14	7	12	6	8	14	14	**75**
Google Gemini (Thinking)	5	14	10	7	8	14	14	**72**
Perplexity	9	7	7	9	11	14	14	**71**
Google Gemini (Fast)	13	7	9	2	7	14	14	**66**
DeepSeek (DeepThink)	1	0	12	12	13	14	14	**66**
Warren AI	3	7	6	4	14	14	14	**62**
Investing AI	10	7	6	1	6	14	14	**58**
Anthropic Claude Sonnet 4.5	4	0	8	3	14	14	14	**57**
Fiscal AI	2	10	7	10	4	14	7	**54**
DeepSeek (Standard)	8	0	4	5	6	14	14	**51**

Table 7 : Global results for the 14 models across all test sets

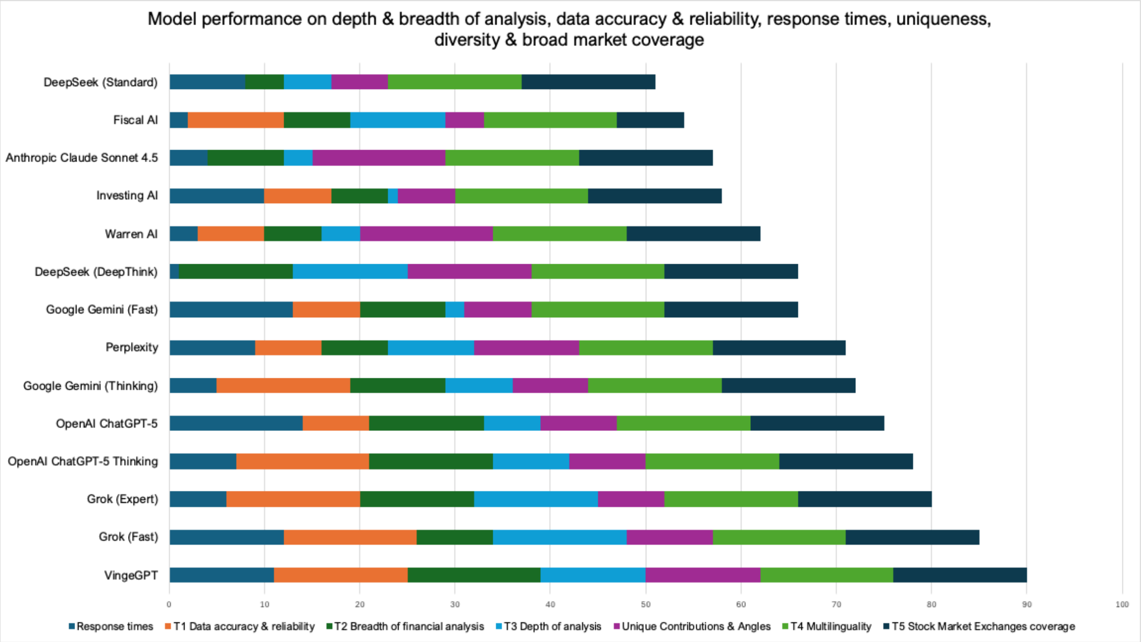


Figure 2 : Model performance on depth & breadth of analysis, data accuracy & reliability, response times, uniqueness, diversity & broad market coverage

Across the benchmark, data accuracy emerges as the primary differentiator, followed by analytical depth and global inclusiveness, underscoring that speed and creativity are only valuable when grounded in timely and verifiable financial information.

5 Appendices

5.1 Appendix 1 – Chain of Thought/prompt used for analysis in ChatGPT 5.2

The attached document compares 14 different AI models amongst them. the document contains only raw result data for each of the 14 models without any preliminary scoring. I want you to act as a neutral researcher and compare the models amongst them. in terms of testing methodology we have submitted the same group of prompts prefixed as T1, T2 , T3, T4 and T5 prompts to each of the 14 models.

- *T1 analyses the reliability & accuracy of data. for T1, each model should be able to provide the balance sheet of Nike dated August 31st 2025.*
- *The group of prompts T2 analyses the breadth of financial analysis of the model.*
- *The group of prompts prefixed T3 analyses the depth of analysis for Nike which is a very well known company. The T3 prompts analyse standard financial information. These T3 test set prompts covers from basic financial information questions up to analysing the auditor fees or providing information about insider trades.*
- *The T4 group of prompts analyses the multilinguality of each model to determine if all models are inclusive of multiple languages and not just English.*
- *the group T5 of prompts analyses the span of its investment universe to see if companies from other stock exchanges than only US are included in the model. We have requested for each model in T5 test set to get information and the balance sheet for 1 company on the Bombay Stock Exchange and another company on the Japan Stock Exchange.*

Your task is to analyse the raw results for each prompt for each model with the following steps :

1) *Firstly, determine which is the model with the best accuracy & reliability of information analysing the results and contents of the prompts T1. Outdated financial information should be penalised the farther it is from August 31st, 2025 the lower the score shall be. For the models that have provided accurate information they shall be ex-aequo on this category with the maximum grade of 14 points, models that are between 1 and 2 quarters outdated give them 7 points, models that have outdated information of more than 2 quarters give them 0 points.*
2) *Secondly, analyse the breadth of analysis using the results of prompts T2.*
3) *Thirdly, analyse the depth of analysing by looking how far models are able to provide information on a specific company like Nike. Models that are unable to provide accurate answers to all T3 prompts should be penalised.*
4) *In step 4, I want you to assess for each model if they provide any unique analysis angles or data points that the other models do not provide. Unique analysis angles or datapoints can be elements that complement the financial metrics*
5) *Then in step 5, I want you to determine the model that supports best investors in their investment process by being most synthetic instead of ballooning into long, diffuse responses.*

After having done the analysis for the 5 steps, give to the best model 14 points in each category and 1 point for the lowest. In order to grade the models in each analysis category, provide the results in a table with the following columns :

Here the list of columns to be used :

- *1st column being the full name of each model analyse, so that users can identify the model when reading the table*
- *2nd column being the score for data accuracy & reliability related to T1 prompts test set. As said earlier, for the models that have provided accurate information dated August 31st, 2025 they shall be ex-aequo on this category with the maximum grade of 14 points, models that are between 1 and 2 quarters outdated give them 7 points, models that have outdated information of more than 2 quarters give them 0 points.*
- *3rd column grade the breadth of financial analysis by looking at the results of T2. To grade the breadth of analysis consider first if the model was able to answer to all questions. Then consider the accuracy of the responses. Models that tend to balloon the answers with generic inaccurate answers shall be penalised*
- *4th column grade the depth of analysis capabilities of each model related to the T3 prompts*
- *5th column grade the unique contributions and datapoints of each model*
- *6th column grade the diversity & inclusiveness of the models by analysing the results of the T4 prompts related to multilinguality. Models that respond in all tested languages being English, Arabic, Spanish, French and Chinese shall receive 14 points and be considered ex-aequo.*
- *7th column grade the inclusiveness & stock market diversity by looking at T5 prompts results. Models that can provide firstly information about the company and secondly the latest balance sheet dated Sept 30th, 2025 for Reliance Industries & Mitsubishi Corporation shall receive 14 points and be ex-aequo. Models that can only provide company information but not the balance sheet for Reliance Industries and Mitsubishi Corporation should be penalised and get 7 here. Models that cannot provide neither information nor the balance sheet for both companies shall receive 0 points here. Models that can provide information about the company and the balance sheet but the balance sheet being outdated meaning not dated Sept 30th, 2025 shall only receive 7 points.*
- *the last column of the summary table should be the sum of points of each category for each model.*

Make sure to read the whole document, take your time for the various analysis & grading steps and do not rush.