Comparative Performance Analysis of Leading AI Models in ToltIQ

This report highlights the distinctive traits of leading AI models utilized at ToltIQ. The models included in this benchmark were **ChatGPT 5**, **ChatGPT 4.1**, **Gemini 2.5 Pro**, and **Claude 4 Sonnet**. We analyzed their performance when leveraged in the ToltIQ platform and scored their responses based on six different evaluation criteria: accuracy, prompt relevance, reasoning, problem solving, industry relevance, and human opinion. We previously conducted a similar benchmark test on earlier models. However, given that AI systems continue to evolve rapidly, we repeated the analysis to capture performance improvements and included ChatGPT 5 in this updated comparison. This ensures our evaluation reflects the most current model capabilities and provides a reliable guide for practical use in Private Equity workflows.

Authors

Steiner Williams - Private Equity AI Researcher (Led the benchmarking exercise, data analysis, and reporting.)

Maya Boeye - Head of AI Research (Provided conceptual guidance, supervised the research process and report review.)

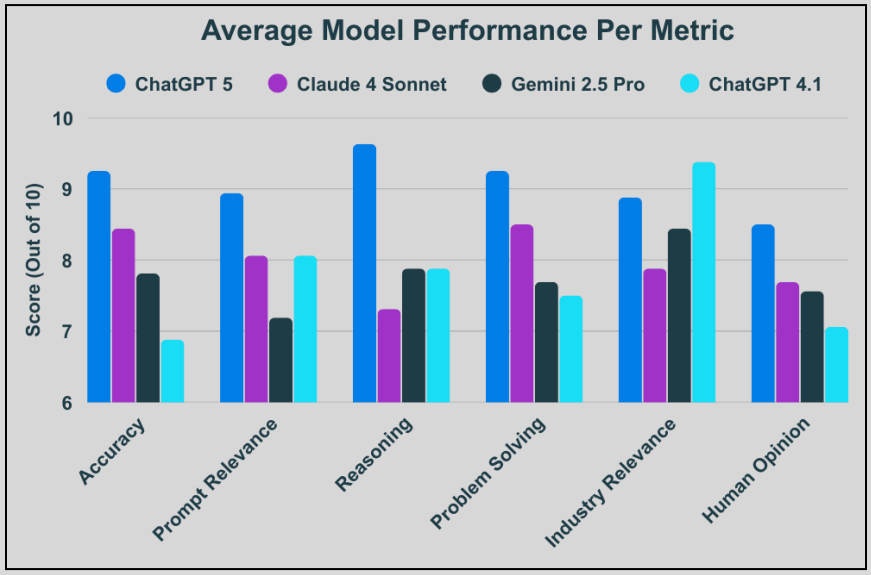
Methodology and Criteria

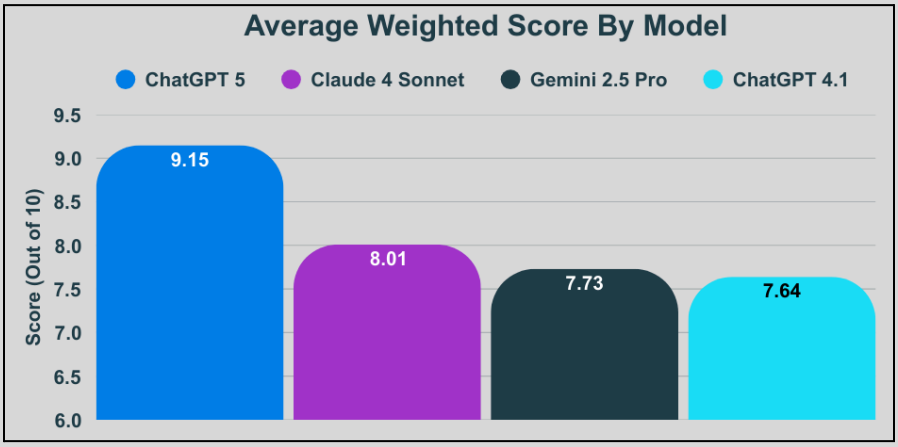
We evaluated the models on the ToltIQ platform using prompts based on Amazon public filings, designed to mirror analytical tasks relevant to Private Equity. Each model received the same prompts, and responses were scored across six weighted criteria: Accuracy (30%), Relevance (20%), Reasoning (20%), Problem Solving (10%), Industry Relevance (10%), and Human Opinion (10%).

Human evaluators manually scored each response using rubrics, without considering which model produced it. Scores were aggregated and weighted to produce the final benchmark results, ensuring both fairness and relevance to Private Equity decision-making.

Model Analysis

In our benchmark, **ChatGPT 5 performed the best with an average weighted score of 9.15 out of 10, followed by Claude 4 Sonnet at 8.01 out of 10**. ChatGPT 5 consistently excelled across most tasks. However, these results don't suggest a one-size-fits-all approach for Private Equity due diligence. Throughout this benchmark, we discovered a few idiosyncrasies for each of these models.





Discussion

ChatGPT 5 was the strongest performer in this latest benchmark testing with a score of 9.15 out of 10, distinguished by its meticulous approach to numerical data, always presenting values with exact decimal places. Other models will use shortened notation (2.3 million) instead of full numerical format (2,300,000.00). It is also notable for its frequent use of disclaimers, often stating that there is not enough contextual information when data is incomplete, even when only minor contextual details are missing. While other models and ChatGPT 5 both attempt to provide complete answers, ChatGPT 5 also indicates when additional information would improve its accuracy. Structurally, it produces highly detailed tables with comprehensive headers and extensive categorization.

Private Equity Context: These behaviors make ChatGPT 5 a very effective model for LBO validation, CIM comparisons, and drafting IC memos. Its precision, transparency, and structure align directly with the workflows where accuracy and completeness are non-negotiable.

Claude 4 Sonnet performed strongly with a score of 8.01 out of 10. It places a greater emphasis on risks and limitations than its peers. Its responses are structured with moderate depth, ensuring that all elements of a prompt are addressed without unnecessary elaboration. It organizes its analysis systematically, aligning responses closely with the requested structure. Claude 4 Sonnet did not have any single notable aspect that stood out in our benchmark as it demonstrated fairly consistent reliability across all metrics.

Private Equity Context: Claude 4 Sonnet's risk-conscious idiosyncrasy makes it particularly effective for red team analysis, regulatory diligence, and operational reviews. It helps investors stress-test CIM assumptions, highlight regulatory risks, and uncover operational vulnerabilities before capital deployment.

Gemini 2.5 Pro earned a score of 7.73 out of 10. It was the most comprehensive in providing historical context, often explaining the background and evolution of business segments in greater depth. Its analysis tends to be narrative-driven, relying on explanatory prose rather than structured tables or lists. One of its idiosyncrasies is the way it reverses rows and columns in tables; while the information remains accurate, its formatting differs from the conventions of other models. Human evaluators prefer the tables Gemini 2.5 Pro creates over the others because it displays information consistently with the way we read content in everyday life.

Private Equity Context: Gemini 2.5 Pro supports sector due diligence by highlighting industry evolution and regulatory shifts. Its scenario modeling is useful for exit planning and sensitivity analysis, while its narrative style aids portfolio monitoring by framing market and operational

developments. Though less precise than ChatGPT 5, it adds depth to diligence and review workflows.

ChatGPT 4.1 scored the lowest overall, with a weighted score of 7.64 out of 10. While it is the oldest model on this list released to the public and has now been replaced by ChatPGT 5, it still has its uses. It is concise and action-oriented, focusing on clarity and brevity with strong emphasis on strategic recommendations. The downside is it often employs approximation symbols, such as “~38–40%” for market share, prioritizing directional accuracy over absolute precision. Its communication style leans heavily on bulleted lists, making findings easy to scan.

Private Equity Context: ChatGPT 4.1 excels in deal screening, rapid market sizing, and preparing executive summaries for management meetings, where brevity and actionability matter more than full precision.

Conclusion

Ultimately, ChatGPT 5 sets the benchmark for precision and structured analysis in Private Equity, while the other models provide complementary strengths across diligence, scenario planning, and risk assessment.

Based on the results of this most recent benchmark, ChatGPT 5 remains the default model on the ToltIQ platform for many high-precision functions, while Claude 4 Sonnet is applied in specific due diligence workflows where it outperformed ChatGPT 5. Gemini 2.5 Pro continues in Beta testing, supporting use cases where narrative depth and scenario modeling add value.

At ToltIQ, we remain model agnostic, giving users access to multiple leading LLMs within one secure and approved environment. The ongoing, rigorous testing we conduct directly informs what models are available on the platform and the use cases we suggest for each model, ensuring that our clients are equipped to generate the highest quality results for any given task.

Access to three of the world’s most advanced models, within a single trusted platform, remains one of ToltIQ’s most powerful differentiators.