Evaluation of Modern Large Language Models as Tools for Financial Due Diligence: Gemini 2.5 Pro Preview, GPT 4.1, & Claude Sonnet 4

Alfast Bermudez - 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.)

Abstract
This study is a comparative evaluation of the currently released versions of three leading large language models (LLMs), GPT 4.1, Claude Sonnet 4, Gemini 2.5 Pro Preview, and their performance on several financial due diligence tasks relevant to private equity. Utilizing TolIQ's database of hundreds of real-world use cases, 16 use cases holistically reflecting different aspects of the due diligence process were selected and given to the models for testing. Each model was also given a VDR (virtual deal room) with a large set of source documents consisting of financial filings, investor materials, press releases, and news articles relevant to publicly traded company, Amazon.com, Inc., to base its answers on. Model performance was measured across quantitative traits such as speed, source usage, and output length, and tied into qualitative traits such as relevance, accuracy, reasoning, and industry understanding. Our findings indicate that while GPT 4.1 and Claude Sonnet 4 models demonstrate strong capabilities, they excel in different aspects of the due diligence process. Claude Sonnet 4 achieved the highest scores in analytical depth and reasoning, while GPT 4.1 offered superior speed and consistent structure. Gemini 2.5 Pro Preview shows distinctive promise in its ability to consider 2.3x more unique sources than GPT 4.1 and 1.89x more unique sources than Claude Sonnet 4 on average per output. However, optimization will be required for Gemini before its outputs receive similarly high qualitative scores as GPT 4.1 and Claude Sonnet 4.

Model Name	Strengths	Best For	Limitations
GPT 4.1	Fastest generation times, concise outputs, well structured, least variance in output quality and speed, and reliably accurate.	Time sensitive analysis requiring concise and informed outputs, first pass reviews, or creating structured reports.	Less interpretive depth and least number of sources cited.
Claude 4 Sonnet	Highest quality outputs, excels at deep analysis, strong industry context, second fastest generation times, consults more sources than GPT 4.1 on average.	Complex valuations, trend analysis, detailed research, various other due diligence functions.	Slower generation than GPT 4.1 and slightly higher variance in performance.
Gemini 2.5 Pro Preview	Highest number of unique sources considered per output, cites sources at 3.75x the rate of GPT 4.1 and 2.29x the rate of Claude Sonnet 4.	Basic questions with answers that may require more document searching than the other models.	Overly verbose at times, longest generation times, lowest qualitative scores.

1 - Introduction

ToltIQ enables financial professionals to leverage their choice of LLMs from industry-leading providers for the streamlining of financial due diligence. To ensure that we offer the most accurate and powerful selections, our research team is dedicated to continuous evaluation of LLM performance in due diligence as models rapidly transform and are released. These tests are possible due to our platform's unique data processing architecture, enabling the LLMs to interact with large VDRs dense with real-world financial documents for the testing of hundreds of due diligence use cases.

Google's recent release of Gemini 2.5 Pro offers another exciting potential addition to our current selection. As part of our rigorous evaluation process before adding a model to our platform, we sought a new preliminary assessment of the model's holistic due diligence capabilities with a focus on private equity. To do so, we designed a simplified VDR and created prompt lists for testing on the model. Our research team then conducted performance benchmarking against Anthropic's Claude Sonnet 4 and OpenAI's GPT 4.1 with the goal of understanding each model's strengths and nature.

2 - Methodology

2.1. VDR Design

The VDR we presented to the models was filled with several documents relating to Amazon (AMZN). As one of the largest publicly traded companies in the world, Amazon has highly detailed filings and high press coverage which are optimal for this research. Over other publicly traded companies, the depth and size of available information challenges the models to seek the most relevant information from a larger pool of data while also supplying enough material to derive unique conclusions better reflecting the model's reasoning nature. The depth of this publicly available information is also better suited to reflect what investment professionals may confidentially obtain as opposed to assembling public information on a private company.

Included in the VDR were the company's 10-K filings for years 2008, 2014, 2020, 2022, 2024; several other financial documents such as 8-K, 10-Q, and special disclosure filings from varying time periods, as well as proxy statements, earnings call transcripts, Seeking Alpha analysis articles both bullish and bearish in nature, press releases regarding company activity, and other investor materials.

By including a broad temporal range of data and third-party analysis, we aim to assess not only the quality of model outputs, but also how well models prioritize newer over older information, how actively they form independent conclusions versus repeating external opinions, and whether they overly treat such opinions as fact.

2.2. Prompt List Designs

Our platform allows users to individually choose whether to enable the models to consider sources, utilize our web agent, and consider previous chat history to construct an answer.

2.2.1. Prompt List

The prompt list was intended to condense our use case database into 16 prompts curated to represent a wide variety of due diligence scenarios. For this test, only source consideration was on while chat history consideration was off and web agent capability was off. This was intended to ensure the models based their answers on new searches through the provided source material rather than previous conclusions or information sourced from the internet. Some prompt categories and example prompts from the tested list were as follows:

2.2.1.1. Financial
Prompts regarding analysis of financial accounting metrics and their sources and implications.

> Example from Prompt List A: *"Model the company's ability to take on debt based on EBITDA, interest coverage, and leverage ratios. Discuss feasibility of leveraged recap scenarios and impact on returns (IRR/MOIC)."*

2.2.1.2. Product Financial
Prompts regarding the financial analysis of products and services.

> Example from Prompt List A: *"Provide a detailed breakdown of unit economics by product or service line, including customer acquisition cost, contribution margin, and payback period. Identify breakeven points and how scale affects margins."*

2.2.1.3. Market & Customer
Prompts regarding the analysis of the business's market positioning and customer base.

> Example from Prompt List A: *"Provide an overview of the target company's total addressable market (TAM), segmented by product, geography, and customer type. Assess growth drivers, headwinds, and recent market trends. Evaluate how the company's current share compares with competitors and suggest areas for strategic growth."*

2.2.1.4. Environmental, Social, and Governance (ESG)
Prompts regarding the business's ethical practices.

> Example from Prompt List A: *"Evaluate the company's environmental, social, and governance (ESG) practices. Assess sustainability disclosures, carbon footprint, diversity metrics, board independence, and any known ESG-related risks or controversies."*

2.3. Evaluation Methodology
Outputs for Prompt List A and Prompt List B will both receive identical but separate quantitative and qualitative evaluation for output quality. It is important to note that these evaluation metrics, while valuable, require both the quantitative and qualitative results compared together to make final conclusions about model output quality and nature.

2.3.1. Quantitative Evaluation
Leveraging a Microsoft Excel Visual Basic for Applications macro created for this research, each model will have its responses evaluated then have several model performance metrics calculated based on the evaluations. These metrics are:

2.3.1.1. Average Time to First Token (Avg. TFT)
A measurement of time in seconds reflecting the average output's time to first generated token.

2.3.1.2. Response Generation Time (Avg. RGT)
A measurement of time in seconds reflecting the average output's time completed generation.

2.3.1.3. Average Number of Citations Made (Avg. # of Citations Made)
The average number of citations made per output.

2.3.1.4. Average Number of Unique Citations Made (Avg. # of Unique Citations)

The average number how many of the citations made correspond to a unique source rather than utilizing previously cited sources.

2.3.1.5. Percentage of Total Sources Used (% of Total Sources Used)
Figure calculated by dividing the number of unique sources cited across every output created by a model by the number of total available sources in the VDR.

2.3.1.6. Information Diversity Score
Figure calculated by multiplying Avg. # of Unique Citations with % of Total Sources Used to scale metric down toward better reflecting how much unique information from the whole is being presented in any given citation. This is then multiplied again by the Avg. # of Citations Made to reflect how much of this information is present in an output.

2.3.1.7. Average Output Character Count
A count of how many characters an output consists of.

2.3.1.8. Information Density Score
This is calculated by dividing the information diversity score by the average output character count to reflect how much information is reflected across the average character. As this tends to be a very small decimal number, it is then multiplied by 1000 for ease of comparison.

2.3.2. Qualitative Evaluation
Qualitative evaluation for the outputs involves human grading of several different output characteristics. All characteristics are to receive a score of either 1, 3, 5, 8, or 10, based on the level of success regarding that characteristic. The characteristics are:

2.3.2.1. Relevance
Relevance is the measure of how relevant an output is to the question being asked by the prompt. A true 10 score for relevance implies that the output precisely addresses every single requirement laid out in the prompt comprehensively; demonstrates perfect understanding of explicit and implicit needs; delivers tightly focused content with highly relevant additional insights; and follows the organizational protocol of the output requested by the prompt perfectly.

2.3.2.2. Accuracy
Accuracy measures the factual and technical correctness of the output, including proper use of data, terminology, and domain knowledge. A true 10 score for accuracy means the output is entirely free of factual or conceptual errors, represents concepts with technical precision, uses appropriate terminology with contextual correctness, and delivers claims that are fully substantiated.

2.3.2.3. Reasoning
Reasoning assesses the logical progression and justification of ideas in the response. A true 10 score of reasoning indicates the output demonstrates exceptional critical thinking, with sophisticated arguments, clearly articulated logic, and a tight causal flow between ideas. The response handles complexity and nuance effortlessly, drawing well-supported and insightful conclusions.

2.3.2.4. Problem Solving
Problem Solving measures the response's ability to identify, understand, and address challenges or questions effectively. A 10 score in problem solving demonstrates thorough identification of all core issues, with deeply reasoned, innovative solutions, and proactive insight that transforms obstacles into strategic opportunities.

2.3.2.5. Industry Relevance
Industry Relevance reflects how well the response aligns with the language, practices, and expectations of professionals in the relevant field. A true 10 score means the response exhibits expert-level domain knowledge, uses precise and accurate industry terminology, and demonstrates real-world applicability that could directly inform or influence high-level work.

2.3.2.6. Human Opinion
Human Opinion assesses how natural, usable, and credible the response feels to a human reader. A 10 score in this category means the response is indistinguishable from that of a skilled human expert — it reads authentically, is highly engaging, and could be immediately delivered to a client or stakeholder without revision.

These scores are then averaged out to receive an average qualitative score for each model.

2.3.3. Final Assessment
The final assessment is a comprehensive analysis of model behavior through output characteristics and observations of patterns in model output. Each model is given a final score in which the scoring formula prioritizes the average qualitative score, giving it a dominant weight. Generation time is lightly penalized while character count has no effect unless the average qualitative score is low. This version ensures that strong qualitative values can push the final score into a higher band while others adjust proportionally. The result is a model that rewards high-quality traits over raw length or speed. A comprehensive analysis of the results will be conducted and subsequently discussed in detail

2.3.3.1. Final Score Methodology
Input variables:

T = Average qualitative score (from 1 to 10)
G = Generation time (in seconds)
C = Character count

Calculation:

$$\text{BaseScore} = \left(\frac{T}{10}\right) \times 95$$

$$\text{PenaltyTime} = \min\left(\frac{G}{12},\ 15\right)$$

$$\text{PenaltyLength} = \begin{cases} \left(\frac{C-3000}{2000}\right) \times 20 & \text{if } T \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{FinalScore} = \text{BaseScore} - \text{PenaltyTime} - \text{PenaltyLength}$$

3 - Results

3.1 Quantitative Results

Model Name	Information Diversity Score	Avg. Output Character Count	Information Density
GPT 4.1	37.20	5106.30	7.29
Claude Sonnet 4	79.25	4628.90	17.12
Gemini 2.5 Pro Preview	366.08	6318.10	57.94

These metrics are the raw statistics for average speeds, citations made, unique citations, and what percentage of the total available sources were cited per model. GPT 4.1 remains the fastest model at the cost of having the least average citations per output, unique citations, and coverage of source materials. Gemini 2.5 Pro Preview had the longest generation times out of the tests, likely due to a vast increase in citations, unique citations, and having the highest coverage of the available sources. It is important to note that this does not necessarily imply that Gemini is the superior model.

Model Name	Avg. TFT	Avg. RGT	Avg. # of Citations Made	Avg. # of Unique Citations	% of Total Sources Used
GPT 4.1	7.2	36.2	15.5	3.2	75.00%
Claude Sonnet 4	16.6	61.7	25.4	3.9	80.00%
Gemini 2.5 Pro Preview	39.1	95.3	58.2	7.4	85.00%

After making the appropriate calculations as laid out in the methodology, these are the remaining quantitative results. From these it can be noted that Claude Sonnet 4 on average was able to output more information in more concise formats than GPT 4.1. Gemini 2.5 Pro Preview is also the most informationally dense as suggested by these metrics.

3.2 Qualitative Results

Test Name	Relevance	Accuracy	Reasoning	Problem Solving	Industry Relevance	Human Opinion
GPT 4.1	6.53	6.66	6.59	6.44	6.75	6.75
Claude Sonnet 4	7.44	8.25	8.06	8.13	8.19	8.06
Gemini 2.5 Pro Preview	6.00	6.31	6.25	4.88	6.00	5.44

From these, the following averages were calculated:

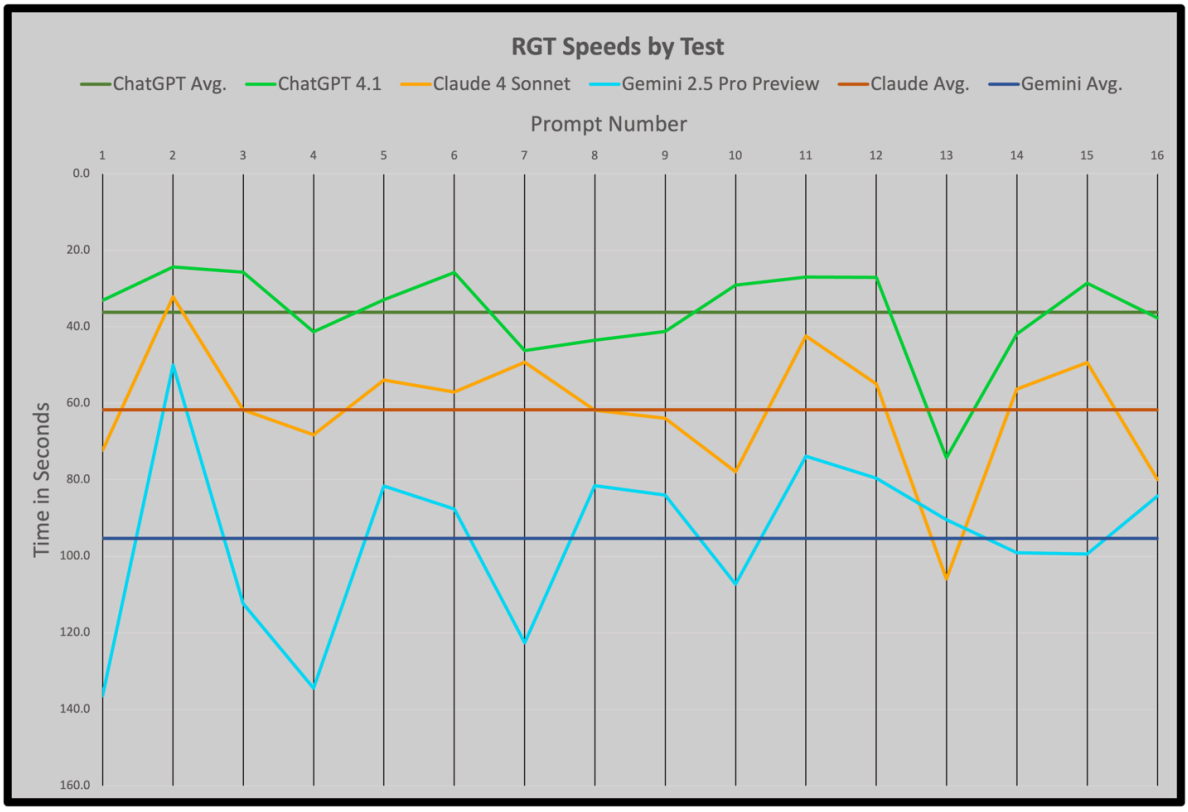
Model Name	Avg. Qualitative Score
GPT 4.1	6.62
Claude Sonnet 4	8.02
Gemini 2.5 Pro Preview	5.81

3.3 Final Assessment
After finalizing calculations with all relevant variables, the following are the final scores for each model:

Model Name	Final Score
GPT 4.1	60
Claude Sonnet 4	71
Gemini 2.5 Pro Preview	47

Claude scored highest in our weighted evaluation, particularly excelling in detailed analysis tasks, while GPT 4.1 demonstrated advantages in speed and consistent structure. Though Gemini showed strong potential by considering the most unique sources and making the most citations, it trailed with generally more superficial and occasionally off-target responses. This signifies that its larger outputs and source consideration could be functioning more as a liability than a benefit to end users, but if optimized could lead to much stronger performance in the future.

In pursuit of analyzing situational behavior, we graphed response generation times across all tests within Prompt List A as well as overlayed the model average times.



This offered insight into how models deviated from their average performance in specific situations and suggested prompts of interest to investigate. Gemini's greater and more frequent

deviations from its average suggest struggle with certain concepts while GPT 4.1 and Claude Sonnet 4 remained closer to their averages, implying stronger and more well-rounded performance.

An example of a prompt of interest is Prompt 13, which resulted in the largest deviation from the average for Claude Sonnet 4 and GPT 4.1. The deviation for the two moves almost in tandem downward implying similar ingestion processes for this specific situation.

3.3.1. GPT 4.1 Assessment

Based on performance and quality of outputs, GPT is very useful as a due diligence tool for analysts and is ranked second of the models. The model's generation speed was the fastest tested and its consistent quality ensures that professionals can rapidly obtain reliable information with nothing obvious overlooked, and information clearly presented. Its dependable accuracy and clarity also mean much less time correcting errors or reorganizing content. This made its responses comprehensive and directly useful as a first pass diligence item. Ultimately, GPT delivers critical information and structured presentation while leaving the human user to derive the most critical conclusions for investment decisions.

GPT 4.1 specifically excelled at providing clear, structured descriptions of financial data, prioritizing speed and reliability. For tasks requiring rapid turnaround, its 41.3% faster generation time over Claude and 62% faster generation time over Gemini offers significant advantages when time matters most.

However, to turn GPT's work into a truly insightful analysis, an investor or analyst would still need to layer on additional analysis such as calculating metrics, comparing against industry standards, and drawing conclusions or recommendations. For instance, the model would list revenue segments and their growth qualitatively but stop short of calculating exact percentages or making forward-looking inferences. In an income statement analysis, GPT recounted trends (revenue up, net income down in 2022, etc.) and flagged a big one-time loss, but it did not mention seasonality effects or provide much interpretation beyond stating the facts.

Similarly, for a question on unit economics, GPT explained the concept of unit economics and noted that AWS has better unit margins than physical retailers but provided no figures or detailed cost breakdown to quantify those differences. Another example showed logically sound reasoning in laying out cause and effect such as linking the 2022 loss to the investment's write-down, yet it didn't volunteer much implication commentary that a human analyst might add.

Regarding structure, GPT's answers were consistently well-formatted and on-point. It excelled at organization and clarity such as breaking down complex questions (like TAM or financial trends) into categorized sub-sections of relevant markets with bullet points or headings, making its output simple to follow. GPT almost always addressed the specific tranches within questions asked, such as if a prompt had multiple parts GPT tended to touch on each part systematically.

In terms of accuracy, GPT performed well, demonstrating an in-depth analysis of the sources, and then leveraging the provided source data to avoid hallucinations. It rarely asserted facts that weren't supported by the documents, offering more insightful analysis instead if possible. For example, regarding a prompt about Amazon's top 10 customers (where no such data was available), GPT searched through all source material and openly stated that the information wasn't disclosed then chose to explain Amazon's diversified customer base instead of guessing. This is a prudent and factually correct approach that shows the model's understanding of how to address user needs.
Across financial questions, it accurately reported key figures that were in the sources (noting specifics such as a 2022 net loss and its cause) and avoided making up numbers if data wasn't readily provided.

In terms of industry relevance, GPT showed a decent understanding of Amazon's business itself, referencing known elements like Prime membership, AWS's role, and antitrust concerns. However, this often would only reflect when prompted by the question. It generally did not bring in a lot of outside industry comparison or insight on its own.

For example, when comparing KPIs to industry benchmarks, GPT correctly surmised that Amazon's CAC might be lower than a typical retailer's (given its strong brand and organic traffic), but this point was made briefly and without citing specific benchmark figures. It further demonstrated that it was aware of broad industry facts in later prompts (such as Amazon's lack of customer concentration or its high payables relative toother retailers) and suggests that professionals prompting GPT should bolster prompts to stimulate injection of the model's greater relevant knowledge.

On human-like opinion or insight, GPT's didn't insert evaluative judgments or forward-looking speculation. This conservatism is often a positive behavior as it further reinforces the model's tendency to avoid unsupported opinions but also resulting in some answers "feeling" neutral or lacking a true takeaway.

For instance, the following prompt,

"*Assuming Amazon were a private company, analyze likely exit paths (strategic buyer, IPO, secondary PE) and their valuation multiples. Model base, bull, and bear case exit valuations using precedent transactions and public comps*"

was intended to measure the model's ability to ingest illogical questions such as this one. Amazon's business and size would make it relatively impossible to perform any traditional private "exit paths."

In discussing its answer, GPT listed IPO or sale as potential exit paths and noted the company's size but did not elaborate on the challenges or likelihood of each scenario in-depth. A human professional likely would have noted that a sale would be nearly impossible due to antitrust and financing constraints.

3.3.2. Claude Sonnet 4 Assessment

Upon assessment, Claude appears to be exceptionally well-suited for deep financial due diligence work and is ranked first of the models. Impressively, it could function almost like a highly competent research analyst and its use can elevate the rigor of analysis to ensure no important detail or reasoning step is overlooked. It often demonstrated a strong ability to find and incorporate precise data, and to reason through complex multi-part questions.

It's ability to do this resulted in a 8.02 average qualitative score, the highest of all tested models. While this increase in quality resulted in an average 25.5 second slower generation time per output when compared to GPT, this difference is likely negligible for professionals valuing stronger responses. The increase in both quality and generation time can likely be attributed to Claude utilizing around 10 more sources per output on average with those citations tending to reference more unique materials than GPT's citations.

For a private equity professional, Claude can save significant time by handling number-heavy analysis (trend identification, calculations, etc.) and compiling exhaustive lists of considerations. Claude reliably could analyze a company's financial statements and output advanced analysis with drivers and anomalies identified. A professional would then mainly need to review and decide which points matter most, rather than having to dig for the facts from scratch. Claude was also able to do this with a much lower average character count per output than the other tested models, implying stronger efficiency in generation and more concise text.

Importantly, Claude showed almost no instances of factual hallucination in this test, remaining grounded in real data, and clearly flagging unknowns. This built a strong trust in its output and thus for scenarios where thoroughness and accuracy are paramount, such as analyzing a target's financial health, assessing detailed operational metrics, or reviewing compliance risks, Claude is the best of the three. It delivers a level of depth that could allow parts of its answers to be lifted directly into an investment committee memo with minimal editing.

For example, when analyzing financial statements, Claude not only noted that Amazon had a net loss in 2022 but explicitly cited the $12.7 billion loss on the Rivian investment that drove it, a level of detail that demonstrates it combed through the data diligently. In the TAM overview, Claude referenced that Amazon accounts for only about 1% of global retail, implicitly highlighting how large the total market is, an insight that added valuable context beyond just listing categories.

Claude performed exceptionally well at following output structure requests. On more open-ended or data-sparse questions (like customer cohorts or top customers by revenue), Claude often provided a structured approach or a form of proxy data. It acknowledged when information wasn't available, but unlike the other models, it very frequently followed this up with "available information" or logical extrapolations.

For instance, in the question asking for a breakdown of Amazon's top 10 customers, after stating that specific top-customer info was not disclosed, Claude's answer went on to summarize Amazon's revenue breakdown by segment, a much more effective answer toward the intent (where revenue is concentrated) without naming explicitly customers.

This demonstrated a highly analytical mindset where if one angle is a dead end it would find another relevant way to address the underlying question while more relevant to the user's expectations than GPT's offering in the same situation. In a cohort analysis question (where no actual retention data was given), Claude discussed how an analyst might approach it, suggesting that newer customer cohorts (say those gained during the pandemic surge) could be compared to earlier cohorts in terms of spend and retention, and it logically surmised that Amazon likely enjoys strong retention especially via Prime.

In doing this, it was also transparent about the lack of concrete data yet still delivered a reasoned qualitative analysis of customer behavior. This approach is foundational for due diligence reporting, showing that Claude can fill in analytical gaps with sound reasoning without veering into unsupported speculation.

Claude's industry relevance was the strongest of all three models, frequently benchmarking or contextualizing Amazon's metrics against external points. For example, Claude mentioned that growth had "normalized" to single-digit percentages after the COVID boom, implicitly comparing Amazon's recent performance to prior extraordinary periods, an insight that links the company's results to broader market trends.

When discussing KPIs, Claude noted how Amazon's CAC is low compared to others and how Prime's churn is much lower than typical retail loyalty programs, showing familiarity with industry norms. In the ESG answer, it brought up concrete initiatives (like Amazon's Climate Pledge, renewable energy projects) and specific criticisms (unionization battles, etc.), demonstrating a broad awareness of what ESG is as a concept as well as what stakeholders are particularly interested in.

In terms of Human-like Opinion, Claude's tone remained analytical, but it did sometimes venture into slight interpretations. It didn't shy from making logical judgments such as implying that Amazon has significant untapped TAM (given its small share of global retail), or that Amazon could leverage more debt safely given its strong EBITDA and coverage ratios. These are

implied opinions based on data. Claude generally stopped short of subjective or speculative commentary (it didn't "recommend" actions or make value judgments in a personal voice), but it provided all the analytical pieces needed for a human reader to form an opinion.

3.3.1. Gemini 2.5 Pro Preview Assessment

In a rigorous financial due diligence process, Gemini is the least suited of the three models and ranked third, or last. Its lack of detail and tendency to generalize mean it could miss important nuances or fail to flag critical issues. Relying on it alone might be acceptable for preliminary research or for summarizing very basic information about a company (say, if one needs a quick company profile or a sanity check that a certain issue isn't overlooked entirely). For anything beyond basics, Gemini outputs would require heavy augmentation by a human analyst. Essentially, using Gemini in diligence would result in the burden of analysis remaining largely on the human.

In scenarios where time is short and a user needs just an executive summary, Gemini's style might be passable. But given that Gemini was the slowest and poorest performing of the tested models, the others would likely be superior. Gemini might be best reserved for non-critical tasks or as a last resort; it does not yet demonstrate the reliability or analytical capability expected for the complexities and precision required in PE due diligence.

This conclusion is surprising given that Gemini had the highest rate of average citations per output at 58.2 and the highest average of unique sources cited at 7.4. Originally, this could have justified its average output character count of 6318.1 as the inclusion of relevant and new information. However, when subjected to calculating the final score as laid out in the methodology, it's very low qualitative scores imply that these extra citations are largely irrelevant to achieving the goal of the prompt and the larger character count is likely due to 'fluff' in the output.

Upon deeper analysis, Gemini's responses were notably less comprehensive and sometimes off-target, especially in comparison to the other two models. It often provided broad, generic answers where specific or nuanced analysis was needed. For example, in the TAM overview prompt, instead of methodically segmenting the market, Gemini gave a high-level narrative about Amazon's diverse markets and even mentioned the company's "guiding principles" like customer obsession. This distracted from the core question and failed to utilize the user's preferred structure, sounding somewhat like a press release or corporate profile rather than a focused TAM analysis. This pattern suggests that Gemini struggled with Relevance, sometimes including extraneous information, or failing to drill down into the exact details asked.

In multiple cases, it answered in generalities: for the product portfolio breakdown, it named major business lines but did not quantify their revenue share or growth; for the KPI question, it asserted things like "Amazon likely has high customer loyalty" without providing any comparative numbers or evidence. It also did not break down calculations or provide rationales to the extent the others did. In the working capital question, for example, Gemini correctly noted Amazon's basic cash conversion cycle advantage (low DSO, high DPO) but provided no trend analysis or comparison, simply stating the fact and moving on. It didn't highlight how exceptional Amazon's negative cash cycle is, something a more analytical answer would stress.

For the exit scenario prompt, Gemini gave a boilerplate answer (IPO or acquisition, with an unspecified timeline) without discussing the real-world feasibility issues; this indicates a lack of strategic reasoning or at least an unwillingness to engage with the complexity of the scenario.

Industry relevance in Gemini's answers was minimal, seldom going beyond generic statements. It didn't incorporate competitive benchmarks or external data points. For instance, where Claude or GPT might reference industry trends (like pandemic-driven growth or competitor practices), Gemini typically did not. In the ESG question, it listed a couple of broad initiatives and issues

but lacked detail such as mention of specific programs by name or any statistics such as how many electric vehicles Amazon is rolling out. The answer was more consistent with repackaging common knowledge, rather than deeper insight or lesser-known details. This might point to the model falling back on stock phrases and general knowledge when pressed for detail, possibly due to having a weaker grasp of the specifics. This signifies that it failed to understand and adapt its style to the needs of an investor audience, who would prefer specifics over platitudes.

A positive trait of the model was its ability to grasp the basic intent of most questions and usually it covered the obvious points while demonstrating a noticeable tendency to refrain from stating unsupported facts. However, this fear likely also led to issues deriving conclusions from existing facts. As such, Gemini performed poorly in human opinion rankings.

It is worth exploring in the future how Gemini's ability to search through many sources can be optimized. By being naturally inclined to explore more chunks, the possibility exists for outputs to include higher levels of detail than the other two models tested.

3.4 Limitations

The results of this study have some limitations due to the nature of the experiments. A singular evaluator was responsible for assigning qualitative scores and another evaluator's opinion could differ. The study could also be expanded to a greater number of prompts and VDRs utilized to track result consistency over differing industry and due diligence use cases. The final scoring methodology places strong weight on qualitative score and may vary if weights are adjusted to emphasize speed. It is important to note that at any given point, all these LLMs are subject to unannounced changes from their providers that may impact how the models behave. These results were derived from the most current models and can be shifted pending modifications.

4 - Conclusion

Based on our multi-dimensional testing across diverse financial due diligence scenarios, Claude Sonnet 4 emerges as the most capable and reliable model for private equity workflows. Compared to both GPT 4.1 and Gemini 2.5 Pro Preview, Claude demonstrated consistent superiority across core evaluations.

Claude Sonnet 4 delivered the strongest qualitative performance, with exceptional scores in relevance, accuracy, reasoning, and industry applicability. Its ability to synthesize detailed financial analysis, flag data limitations, and build structured reasoning chains positions it as the closest analog to a skilled human analyst. Its ability to draw logical inferences from incomplete data and to reframe prompts analytically also sets a new standard for our offerings in AI-driven diligence.

Further, its ability to maintain a high information density score validated by qualitative scores while with notably shorter outputs than its peers strike a positive balance between conciseness and depth. Unlike Gemini, which suffered from generic and verbose outputs, or GPT, which leaned more descriptive than evaluative, Claude consistently offered analytical commentary, relevant context, and precise data usage rapidly. This results in higher quality first drafts, reduced revision time, improved confidence in AI-supported investment analysis, and ultimately drives faster decision making.

As such, Claude's performance meaningfully enhances analyst productivity, raises the bar for insight generation, and minimizes the risk of oversight in complex financial investigations. Users can expect ongoing performance improvements as Anthropic continues advancing Claude's capabilities, with Claude Sonnet 4 already establishing itself as a cornerstone tool in modern private equity diligence.

While Gemini 2.5 Pro Preview demonstrated significant potential through its superior source utilization capabilities, its current outputs do not yet meet the analytical rigor and precision required for private equity due diligence. We will continue to evaluate future iterations as Google optimizes this promising foundation. Both Claude Sonnet 4 and GPT 4.1 are currently available in ToltIQ's due diligence platform. Based on our testing, we recommend Claude Sonnet 4 for tasks requiring deep analytical reasoning and comprehensive source analysis, while GPT 4.1 remains valuable for rapid, well-structured initial assessments. Gemini, while showing potential, is not yet recommended for client use.