DiligentIQ Agent Mode Testing Report

DiligentIQ AI Research Team

Introduction

In September 2024, the DiligentIQ Research team completed testing and analysis on the new Agent Mode feature. The Agent Mode feature in the DiligentIQ App introduces an intelligent decision-making layer to query processing. When activated, Agent Mode executes a two-step process. First, it performs a query rewrite, translating the user's input into a more precise, context-aware format. This refinement enables more accurate and relevant decision-making. Second, the agent engages in path selection. Based on the rewritten query and Chat settings, the agent determines the optimal action from four possible paths: standard RAG (Retrieval-Augmented Generation) on deal documents, web search using Cohere's Command R+, web scraping for specific URL analysis, or direct LLM interaction. This dynamic approach aims to deliver more flexible and comprehensive responses by intelligently choosing between local document analysis and web-based information retrieval.

The performance of ChatGPT4-Omni May, ChatGPT4-Omni August, and Claude V3.5 Sonnet were analyzed with the Agent Mode feature on and off. The objective of this testing was to examine the Agent's performance, user interactions, speed, and accuracy. The goal of this test was to evaluate whether the Agent Mode feature chose the correct path and if the Agent Mode's rewritten query had a major impact.

Methodology

For this study, 30 unique prompts were tested across three different LLM models, with the Agent Mode feature both on and off, resulting in a total of 6 tests per prompt. The prompts were categorized into six groups: Financial (Annual), Consumer Goods, Credit Agreement Summary, Sustainability, Legal, and Operations. The test employed a diverse range of prompts to verify that the Agent Mode feature could accurately determine the appropriate course of action across various prompt types. We also recorded the Time to First Token (TFT) and Response Generation Time (RGT) for each run, allowing us to see how Agent Mode impacts speed. Each response was analyzed and scored based on 4 key metrics: accuracy, relevance, problem-solving, and sources.

<u>Defining Key Metrics</u>

1. *Accuracy* - The degree to which the output provides information that is precise and factual, containing no errors or misrepresentations.
2. *Relevance* - The degree to which the output fully comprehends and addresses the intent of the prompt, addressing its core objectives.
3. *Problem Solving* - Whether or not the output generates information beyond what is directly stated in the provided documents, relying on calculations for additional insights.
4. *Sources* - The degree to which the output accurately references the correct document and specific page number for information that it is citing.

After each response was analyzed using these key metrics, a statistical analysis was conducted to identify trends and important findings within the data, allowing the team to further examine if the Agent's rewritten query had a major impact. The detailed methodologies for the speed, qualitative and quantitative will be explained in their respective sections.

Speed Analysis

Methodology

To evaluate the speed of the Agent Mode, the TFT and RGT were timed for each response in the test. TFT is the time it takes for the model to output the first word after pressing "Send". RGT is the time it takes for the model to complete a full response after pressing Send. The total time was then calculated for each run, by adding TFT and RGT. We then compared the responses generated with Agent Mode on and off for each model.

Findings

We observed interesting variations in response times with the Agent Mode enabled, providing valuable insights for optimizing performance. Agent Mode did have a tendency to slow down response times when engaged, this is due to the two step process outlined in the introduction. Claude V3.5 Sonnet outperformed both ChatGPT-4 Omni May and August versions with the Agent on. With the Agent off, ChatGPT-4 Omni August performed faster than Claude V3.5 Sonnet across most prompts. However, Claude V3.5 Sonnet consistently outperformed ChatGPT-4 Omni May.

Our analysis of speed performance revealed several key patterns:

- Operations-focused prompts consistently required longer processing times across all models when the Agent Mode was active. Without Agent Mode on, these times aligned more closely with other prompt categories.
- An outlier in the speed analysis was the bullet list prompt, which took Claude 3.5 Sonnet approximately twice as long as other models with Agent Mode enabled. This is atypical, as Claude 3.5 Sonnet usually processes tasks in half the time of other models in this mode.
- The Competitive Landscape Matrix task took both ChatGPT-4 Omni models significantly longer than Claude 3.5 Sonnet's processing time with the Agent Mode active.
- For all Credit Agreement Summary prompts, all models completed the responses much faster than other prompt categories when using the Agent Mode feature. This prompt category employs chain-of-thought prompting, which could explain the quick response times.

These findings demonstrate the variable performance of different LLMs across task types. The data suggests that model performance can vary significantly depending on the specific prompt type and whether the Agent Mode feature is enabled or disabled.

Qualitative Analysis

Methodology

To assess the performance of responses for each run, the accuracy, relevance, problem-solving, and use of sources for each response were analyzed. For each metric, responses were put into 3 buckets: “Good Performance”, “Moderate Performance”, and “Bad Performance”. The criteria of each performance level for each metric was based on subjective analysis shown in Figure 1, below. Notes were taken for each individual prompt after it was analyzed. Finally, after all of the prompts were analyzed and notes were taken for each one, we performed a comparative analysis to evaluate the performance of each unique prompt between each model with the Agent Mode feature on and off. This allowed us to see which models were consistently performing better than others, and which models were struggling. The qualitative analysis for each metric will be further described in their respective sections.

Performance Level	Accuracy	Relevance	Problem Solving	Sources
Good Performance	Provides accurate information that is cited.	Fully addresses the prompt.	Successfully computes solutions.	Accurately references all sources.
Moderate Performance	Provides accurate information without citing sources.	Partially addresses the prompt.	Partially computes solutions.	Correctly references sources but cites the wrong page.
Bad Performance	Provides inaccurate information and does not cite sources.	Fails to address the prompt.	Fails to compute solutions.	Incorrectly references sources or does not reference them.

Figure 1. Performance Grouping Descriptions for each Metric

Findings

Accuracy

With Agent Mode on and off, all models achieved “Good Performance” accuracy ratings, with only a few flaws and “Moderate Performance” ratings throughout. The August ChatGPT-4 Omni model with the Agent on proved to be the most reliable, offering the most accuracy. Additionally, its performance with the Agent on was a close second to its Agent-off mode, with only a slight difference in results. While Claude V3.5 Sonnet showed strong overall performance, its accuracy with the Agent off was slightly lower than the other models, indicating potential areas for future enhancement and growth. Although there were not many flaws, Claude V3.5 Sonnet

with the Agent off had a slightly lower accuracy performance compared to other models. Overall, the average performance for each model with the Agent off and on were consistent.

Relevance

With Agent Mode on and off, all models achieved “Good Performance” relevance ratings, with only a few flaws and “Moderate Performance” ratings throughout. The relevance scores were equal or higher than the accuracy scores for all of the models. The May ChatGPT-4 Omni model with Agent Mode on had the highest relevance performance score among all the models.

Problem Solving

For many of the prompts, problem-solving was not necessary to generate a response, so most of the column has N/A as a qualitative score. The model that used problem solving the most was Claude V3.5 Sonnet, using it a total of 6 times. The May version of ChatGPT4-Omni used problem solving 3 times. Finally, the August model of ChatGPT4-Omni did not use problem solving at all.

Sources

The sourcing performance varied more widely across models, offering valuable insights into areas for refinement and growth. For the August Chat GPT-4 Omni model, there were minor instances where the output included small misinterpretations and incorrect sourcing; this was consistent with Agent Mode both on and off. Sourcing issues such as embedded footnotes leading the user to the wrong page or providing the wrong source link for the data were found in more than one instance.

The May Chat GPT-4 Omni model with Agent Mode on was slightly less consistent and accurate than the August version. There were more instances where sourcing was an issue. Occasionally, the linked sources led to different pages or sources than expected. In terms of accuracy, the model was reliable, but it was weaker in sourcing and accuracy compared to the August model.

With Agent Mode off, the May Chat GPT-4 Omni model was more consistent in terms of sourcing capabilities than when Agent Mode was on, but less reliable than the August version. There were a few sourcing errors with incorrect links.

At first glance, Claude V3.5 Sonnet was able to provide additional insight in responses that other models didn’t. Claude V3.5 Sonnet displayed a creative approach to complex prompts, although there’s room to enhance response completeness and consistency. Sourcing presented some challenges for this model, suggesting an opportunity to strengthen its citation accuracy. In addition, there were sometimes irrelevant sources referenced in several of the responses and at

times, responses themselves were poorly formatted, hindering readability. This model seemingly relies more heavily on quantitative data in its responses, allowing the model more opportunities to showcase its problem solving capabilities, which were successful in some instances. When successful, this allowed for enhanced insights in some cases. Nonetheless, this reliance on quantitative data also resulted in more variable responses that were occasionally inaccurate.

With the Agent Mode off, the Claude V3.5 Sonnet model faced significant challenges surrounding accuracy and sourcing. Despite being limited to internal documents, its accuracy worsened compared to when the Agent was on, producing the most hallucinations of any model. The relevance of responses were comparable with other models, while also showing good problem solving capabilities. However, the accuracy of sourcing was the main area of concern, leading to incorrect pages, or sources altogether relatively often. Overall, this model had the most room for improvement in accuracy, sourcing, and relevance.

Prompt Categories

Across the various prompt categories, there were no visible trends in the accuracy, relevance, and problem-solving performance among the models when the Agent was both on and off. However, the sourcing performance exhibited the most variability, with the credit agreement summary prompts had higher sourcing scores compared to the other prompt types. The reason behind this variability in sourcing remains unclear and may be due to random factors.

Quantitative Analysis

Methodology

To better understand the impact of the Agent feature on each model's performance, we created a simple scoring system to enable a statistical analysis of the results. Based on our previous qualitative analysis, we assigned numbers to indicate performance rating, noting "Good Performance", "Moderate Performance", and "Bad Performance" as defined in Figure 1. across each metric. As seen in Figure 2 below, "Good Performance" was denoted by a 9, "Moderate Performance" by a 5, and "Bad Performance" by a 1.

Attaching numbers to model performance allowed us to calculate averages and standard deviations across each metric, as seen in our Agent Testing Results Table. We were then able to total the individual average scores across each metric to gain insight into the overall performance across each of the three models.

For more granularity, we also assessed the total scores of each model, evaluating them more in depth based on whether the Agent Mode feature was on or off, implying two separate scores per model. From these more granular totals, we were able to deduce the percent improvement in overall performance when the Agent Mode was on.

Metric	Score
Good Performance	9
Moderate Performance	5
Bad Performance	1

Figure 2. Performance Metric Numerical Scores

Findings

Accuracy

The ChatGPT-4 Omni August model produced impressive accuracy on average, yielding scores of 8.867 and 8.733 with the Agent-on and Agent-off modes, respectively. Its consistent performance was also notable, with the lowest standard deviations of 0.730 and 1.015 for the Agent-on and Agent-off modes, respectively. The Agent-on mode had a higher average accuracy and lower standard deviation score, proving that the feature enables more accurate and consistent responses.

While Claude V3.5 Sonnet's accuracy scores were somewhat lower, at 8.467 and 7.933 for the Agent-on and Agent-off modes, respectively, its performance was still competitive with the other models. While it did suffer with occasional data inaccuracies, producing hallucinations 4 times with the Agent mode off and 1 time with the Agent mode on, broadly speaking, the LLM was able to provide particularly insightful responses. Additionally, this demonstrates how even though there is room for improvement, the Agent feature adds significant value to Calude's responses.

In general, these results imply that accuracy was generally improved across the board with the Agent Mode on, suggesting that leveraging the Agent Mode enhances the models ability to generate precise outputs.

Relevance

All models performed roughly equally in terms of generating relevant responses, both with the Agent Mode feature on and off, with the ChatGPT-4 Omni May model holding a slight edge in consistently aligning with the expectations of the prompt. The May model in particular attained a perfect score of 9.000 with the Agent Mode on and a score of 8.867 with the Agent Mode off, highlighting its quality performance. Given that the ChatGPT-4 Omni August model and Claude V3.5 Sonnet scores were comparable, we can conclude that all three models comprehended and addressed the core objectives of each prompt quite well.

Problem Solving

Because not all models employed problem solving capabilities, the key focus here is the frequency in which computing capabilities were utilized. Claude V3.5 Sonnet was the most likely to rely on problem solving in its responses, using it a total of 6 times between the Agent-on and Agent-off modes collectively. This highlights its inclination to provide especially comprehensive responses whenever possible. In contrast, the ChatGPT-4 Omni August model didn't employ problem solving at all in its responses, while the ChatGPT-4 Omni May model used it a total of 3 times. Given that each model handled a total of 60 prompts, it is evident that Claude was more inclined to generate additional insights beyond what was already directly stated in the provided documents compared to the other models, underscoring an area where Claude differentiates itself in terms of response quality and depth.

Sources

The ChatGPT-4 Omni August model set itself apart from the ChatGPT-4 Omni May and Claude V3.5 Sonnet models in its sourcing capabilities. Due to there being significantly more differentiation across the sourcing capabilities of models, our analysis here allowed us to clearly distinguish between cases where the model provided both an accurate source and page number, an accurate source with an incorrect page number, or an entirely incorrect source.

More specifically, the ChatGPT-4 Omni August model yielded scores of 7.207 and 7.933 with the Agent-on and Agent-off modes, respectively, while also proving these sourcing capabilities consistently as evidenced by its corresponding low standard deviation values of 2.744 and 2.083. In contrast, the Claude V3.5 Sonnet model struggled a bit more with accurate sourcing, yielding lower scores of 5.258 and 3.323 with the Agent-on and Agent-off modes, respectively, with greater variability across sourcing performance as evidenced by higher standard deviation values. Although the Claude V3.5 Sonnet model displayed some challenges with accurate sourcing, these findings highlight areas where performance can be improved relative to other models.

Overall, the ChatGPT-4 Omni August model performed better than the other models in terms of sourcing accuracy, particularly when Agent Mode was off, while the ChatGPT-4 Omni May model and Claude V3.5 Sonnet model saw worse performances, also most notably when Agent Mode was off. As a general theme therefore, sourcing accuracy was generally enhanced when Agent Mode was on, with the August model serving as an exception to this.

<u>Results</u>

In terms of overall performance, the ChatGPT-4 Omni August model generally outperformed the other models, with notable strength in accuracy and sourcing capabilities. Alternatively, while the Claude V3.5 Sonnet model showed the most promise with respect to its problem solving capabilities, consistently generating insightful responses, it also revealed room for improvement across the other metrics we tested.

To effectively convey these findings, we totaled the scores for each individual metric, excluding problem solving scores. This exclusion was made because our problem solving data primarily provided insight into frequency of use, rather than quality, as with the other metrics. These total scores reaffirmed our findings, producing total scores of 50.340, 47.646, and 42.714, for the ChatGPT-4 Omni August, ChatGPT-4 Omni May, and Claude V3.5 Sonnet models, respectively.

Importantly, based on these totals, we can also derive that the ChatGPT-4 Omni May model improved in overall performance by 0.93% with the Agent Mode on while the Claude V3.5 Sonnet model saw an even more noteworthy improvement of 12.27% with the Agent Mode on. The August model served an exception to this, exhibiting an overall performance decline of -1.81% with the Agent Mode on. However, this decrease that we see can largely be attributed to its superior sourcing score with the Agent Mode turned off. Overall, the use of the Agent feature generally enhanced the LLMs performance, with Claude V3.5 Sonnet benefitting the most from its use.

Prompt Design and Effectiveness

The prompts used throughout this test varied in complexity, ranging from more straightforward to others requiring multi-step responses. This complexity directly impacted accuracy, relevance, and especially sourcing scores in each model. More complex prompts tended to yield lower average scores compared to simpler ones, as they require deeper reasoning, which often leads to challenges in source verification and fact-checking. In these scenarios, leveraging the Agent Mode can be beneficial, as it helps refine and structure complex prompts, resulting in more accurate responses.

With the Agent Mode feature enabled, the user input is rewritten to be more specific, which likely contributes to the improved accuracy scores observed in the statistical analysis. The Agent helps transform prompts into more direct queries, reducing the cognitive load on the model. By refining prompts with more clear instructions, ambiguity is minimized, leading to more precise responses. The model's ability to rewrite inputs to produce optimal responses is extremely valuable, as it provides an automated form of prompt engineering without requiring additional effort from the user. This feature can help expedite the PE due diligence process by allowing faster decision-making.

Conclusion

This study focuses on testing DiligentIQ's Agent Mode feature across three different LLM models: ChatGPT-4 Omni May, ChatGPT-4 Omni August, and Claude V3.5 Sonnet. Enabling Agent Mode led to noticeable improvements in accuracy and relevance for most responses, showcasing the positive impact of the feature's query refinement step. Additionally, the Agent Mode's intelligent path selection in the second step ensured that the Agent only performed web searches and scrapes when necessary. While we observed that the Agent Mode introduced a slight delay in processing times, it's worth noting that Claude V3.5 Sonnet demonstrated faster response times compared to other models with Agent Mode enabled. Meanwhile, the

ChatGPT-4 Omni May and August models excelled in speed when Agent Mode was off, highlighting each model's unique strengths in different scenarios.

Through our analysis, we gained invaluable insight into the more general performance of these respective models. Outside of the Agent Mode focus of this study, our results show that while Claude V3.5 Sonnet showed promise with its problem solving capabilities, uniquely positioning it to provide higher quality and more in depth responses, the ChatGPT-4 Omni August model ultimately came out on top based on our scoring system. As such, the model solidified itself as having the most impressive accuracy and sourcing capabilities compared to the other models tested regardless of the Agent Mode.

Index

Agent Testing Results Table

Model	Agent (On/Off)	Accuracy (Avg)	Accuracy (SD)	Relevance (Avg)	Relevance (SD)	Sourcing (Avg)	Sourcing (SD)	Score (PS Omitted)	% Improvement with Agent
Chat-GPT4 August	On	8.867	0.730	8.867	0.730	7.207	2.744	24.940	
Chat-GPT4 August	Off	8.733	1.015	8.733	1.015	7.933	2.083	25.400	
								50.340	-1.81%
Chat-GPT4 May	On	8.467	1.737	9.000	0.000	6.467	3.060	23.933	
Chat-GPT4 May	Off	8.467	2.030	8.867	0.730	6.379	3.256	23.713	
								47.646	0.93%
Claude V3.5 Sonnet	On	8.467	2.596	8.867	0.730	5.258	3.059	22.591	
Claude V3.5 Sonnet	Off	7.933	2.559	8.867	0.730	3.323	3.059	20.123	
								42.714	12.27%