DiligentIQ vs Standalone Large Language Models Performance Study

DiligentIQ ResearchTeam

Introduction

In October 2024, the DiligentIQ Research Team conducted a test and analysis comparing DiligentIQ to Large Language Models (LLMs). This report compares the performance of DiligentIQ with ChatGPT-4 Omni and Claude V3.5 Sonnet and the models on their own. The objective was to see the differences in content and structure in the responses. Additionally, our team analyzed which model performed best in generating insights for three prompt categories: Sustainability, Consumer Goods, and Financial. To perform this testing, a Sonos VDR was utilized.

DiligentIQ is designed to leverage the capabilities of multiple LLMs to optimize performance based on user needs. The platform allows users to select from different LLMs, enabling flexibility and customization in generating outputs. In our testing, we used both ChatGPT and Claude, two prominent LLMs integrated within DiligentIQ, to compare their performance within DiligentIQ and independently across various prompts. The models were given the same set of prompts, and their responses were compared based on the speed, accuracy, relevance, and overall quality. Based on the evaluation metrics and overall user performance, the best performing model was chosen for each prompt type.

Methodology

Our research team tested ChatGPT-4-Omni and Claude V3.5 Sonnet both in their own user interface and through the DiligentIQ platform. For the DiligentIQ testing, the prompts were each queried twice, once using ChatGPT-4-Omni and the other using Claude V3.5 Sonnet. We compared responses from ChatGPT within DiligentIQ to those generated by ChatGPT independently, applying the same approach for Claude. This process provided insights into how each model performs both inside and outside of DiligentIQ. The testing helps us understand the impact of DiligentIQ's added software on the performance of Claude and ChatGPT.

The test utilized a set of 21 prompts, each of which was run four times: twice within DiligentIQ, once in ChatGPT, and once in Claude. The prompts were divided into three categories: Sustainability, Consumer Goods, and Financial. For Sustainability, the models were prompted on topics like sustainability practices, risk management, and governance. Consumer Goods prompts focused on assessing company performance, product sustainability, and consumer impact. The Financial prompts required the models to summarize and analyze financial and operational data from corporate 10-K filings. Testing with a diverse range of prompt categories is crucial, as it helps our research team identify potential trends in model performance across different types of tasks.

This testing was conducted with a Sonos VDR containing 57 files. DiligentIQ is able to ingest thousands of documents at once, however, ChatGPT has a limit of ten files per chat, while Claude limits users to five files. Because of this limitation, our researchers sorted through the Sonos VDR to attach the most relevant files for each independent Claude and ChatGPT chat.

Speed Methodology

Each time a prompt was run through a model, speed metrics were recorded to evaluate the model's overall performance. Time to First Token (TFT) and Response Generation Time (RGT) were measured for each response. TFT refers to the time it takes for the model to generate the first word after pressing "Send," while RGT represents the time required for the model to complete the entire response. The total time for each run was calculated by adding TFT and RGT. We then compared the TFT and RGT metrics for each prompt, assessing the performance of Claude and ChatGPT within DiligentIQ against the models operating independently.

Qualitative Methodology

Once all responses were generated, each one was evaluated based on relevance, accuracy, and problem-solving. The descriptions of the qualitative metrics used to analyze each response are listed below.

- *Accuracy* - The degree to which the output provides information that is precise and factual, containing no errors or misrepresentations.
- *Relevance* - The degree to which the output fully comprehends and addresses the intent of the prompt, addressing its core objectives.
- *Problem Solving* - Whether or not the output generates information beyond what is directly stated in the provided documents, relying on calculations for additional insights.

The responses were bucketed in one of three categories (Good Performance, Moderate Performance, and Bad Performance) for each of the metrics. The description for each performance grouping under each metric is described in the table below.

Performance Level	Accuracy	Relevance	Problem Solving
Good Performance	Provides accurate information that is cited.	Fully addresses the prompt.	Successfully computes solutions.
Moderate Performance	Provides accurate information without citing sources.	Partially addresses the prompt.	Partially computes solutions.
Bad Performance	Provides inaccurate information and does not cite sources.	Fails to address the prompt.	Fails to compute solutions.

After evaluating the responses for accuracy, relevance, and problem-solving, we reviewed and compared them for comprehensiveness and overall quality. ChatGPT responses from within DiligentIQ were compared directly with those generated independently, and the same comparison was made for Claude. For each pair of responses, we noted key differences in

content and structure for each model. This analysis was crucial as it enabled our research team to assess how using the LLM within DiligentIQ affected the responses compared to using the models on their own. Based on performance scores and comprehensiveness, the best-performing model for each prompt was identified. If the responses from DiligentIQ and the independent models were similar without significant differences, the models were considered equally effective. The analysis of the best-performing responses was conducted separately for Claude and GPT.

Quantitative Methodology

Once the qualitative analysis was complete, our team had a best performer for each prompt. From this, we were able to get a total count of DiligentIQ, the LLM alone (Claude or ChatGPT), or equal performance. The data was analyzed and bar charts were created to show the overall performance difference between DiligentIQ and the independent LLM responses.

Speed Analysis

The Time to First Token (TFT) and Response Generation Time (RGT) were both recorded for each response in this test to show which model responded fastest and if there were any trends across prompt categories. The results indicate that Claude V3.5 Sonnet demonstrates superior efficiency, consistently achieving the lowest total processing times, predominantly in the 20-30 second range, with TFT often under 10 seconds. ChatGPT4-Omni, despite quick TFT, displays the longest overall processing times due to extended RGT, frequently exceeding 40 seconds and often surpassing 50 seconds.

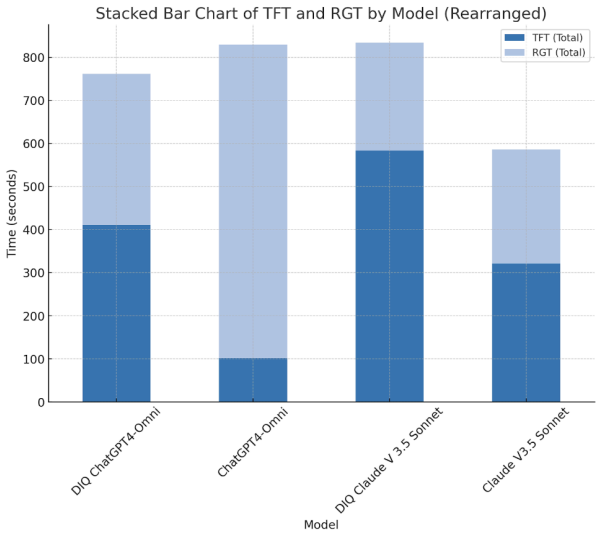
DiligentIQ Claude V 3.5 Sonnet demonstrates consistently moderate processing times across all prompt categories. Its total time averages around 35-45 seconds for most tasks, with a relatively balanced distribution between TFT and RGT. The model shows particular efficiency in ESG-related prompts, with most total times falling between 33-41 seconds. Its performance appears stable across different task types, suggesting reliable and predictable processing speeds.

DiligentIQ ChatGPT4-Omni exhibits variable performance across different prompt categories. It generally has faster TFT compared to DIQ Claude V 3.5 Sonnet, but often longer RGT, resulting in total times that are sometimes faster and sometimes slower than Claude. The model seems to perform particularly well on some ESG prompts, with several instances of total times under 30 seconds. However, it also shows some notably longer processing times, especially for certain Consumer Goods and 10-K prompts, where total times exceed 50 seconds in a few cases.

The standard Claude model demonstrates the fastest overall performance among the four models tested. It consistently shows very low TFT, often under 10 seconds, and moderate RGT. The total processing times are predominantly in the 20-30 second range, with some tasks

completed in under 20 seconds. This model appears to be highly efficient across all prompt categories, maintaining fast and consistent performance regardless of the task type.

The standard ChatGPT4-Omni model shows the longest overall processing times among the four tested. While its Time to First Token is generally quick, often comparable to Claude V3.5 Sonnet, its Response Generation Time is significantly longer. This results in total processing times that frequently exceed 40 seconds, with many instances surpassing 50 seconds. The model's performance seems consistent across different prompt categories, but it's notably slower compared to the other models, particularly in terms of response generation. The stacked bar chart displayed below shows a visual representation of the speed data collected in this test.



These findings suggest significant variations in processing efficiency among the tested models, with performance differences more pronounced in certain prompt categories. Claude V3.5 Sonnet emerges as the most time-efficient across diverse prompt types, while the other models show varying degrees of efficiency depending on the task.

The overall takeaway from the speed analysis is that Claude had the fastest completion times, but there is a trade off in occasionally compromised response depth and performance, which will be discussed in the qualitative section.

Qualitative Analysis

The first part of the qualitative analysis involved rating each response in accuracy, relevance, and problem-solving. The description of each performance category is described in the table above, in the Methodology section..

Overall, the DiligentIQ Claude and DiligentIQ ChatGPT models performed very well. The responses were all rated in “Good Performance” for relevance and accuracy except for one. The

DiligentIQ Claude model had a moderate accuracy for the prompt “Summarize the company's approach to human capital resources based on the information provided in the 10K, including employee diversity, training, and development programs”, which was in the Financial prompt category. The issue found in this response was that it included a fact we couldn't verify through the source documents or online.

The Claude and ChatGPT models on their own tended to have slightly lower performance scores in accuracy and relevance. Claude had one prompt with bad performance and one prompt with moderate performance in relevance. These issues related to not answering the prompt fully or providing data that was irrelevant to the prompt. ChatGPT had four prompts with bad performance in accuracy and one prompt with moderate performance. These issues related to providing data that was not consistent or different from the information provided in the source documents.

Once each response was scored in terms of accuracy, relevance, and problem-solving performance, the next step of the qualitative analysis involved reading the model’s responses within and outside of DiligentIQ, comparing the content and structure. For each response pairing, the key differences between content and structure were noted. It was found that the models used within DiligentIQ lead to more thorough and to-the-point responses. The DiligentIQ responses tended to be more structured and detailed, whereas Claude and ChatGPT tended to give broader overviews.

There were also some differences across prompt categories between the response differences of the LLMs within and outside DiligentIQ. For sustainability prompts, DiligentIQ provided comprehensive and factually accurate responses on sustainability issues, outperforming ChatGPT in terms of governance details. However, ChatGPT excelled in generating socially oriented responses, but lacked specific governance insights compared to DiligentIQ. Additionally, Claude's responses were less detailed, especially in terms of environmental sustainability and governance structures.

For Consumer Goods prompts, DiligentIQ produced strong evaluations on company sustainability practices and aligned its responses closely with the product’s environmental impact. ChatGPT was creative but occasionally veered off-topic, offering more general insights rather than company-specific evaluations. Claude's responses were concise but lacked the depth and relevance that DiligentIQ offered, particularly for company performance assessments.

In the Financial prompt responses, DiligentIQ demonstrated a better understanding of financial terminology and offered detailed insights into company operations, especially around risk management and financial performance. ChatGPT’s responses were more surface-level, and had good summaries but missed key operational details found in the 10-K filings. Claude performed weakest in this category, where it often misinterpreted financial data or missed key points altogether.

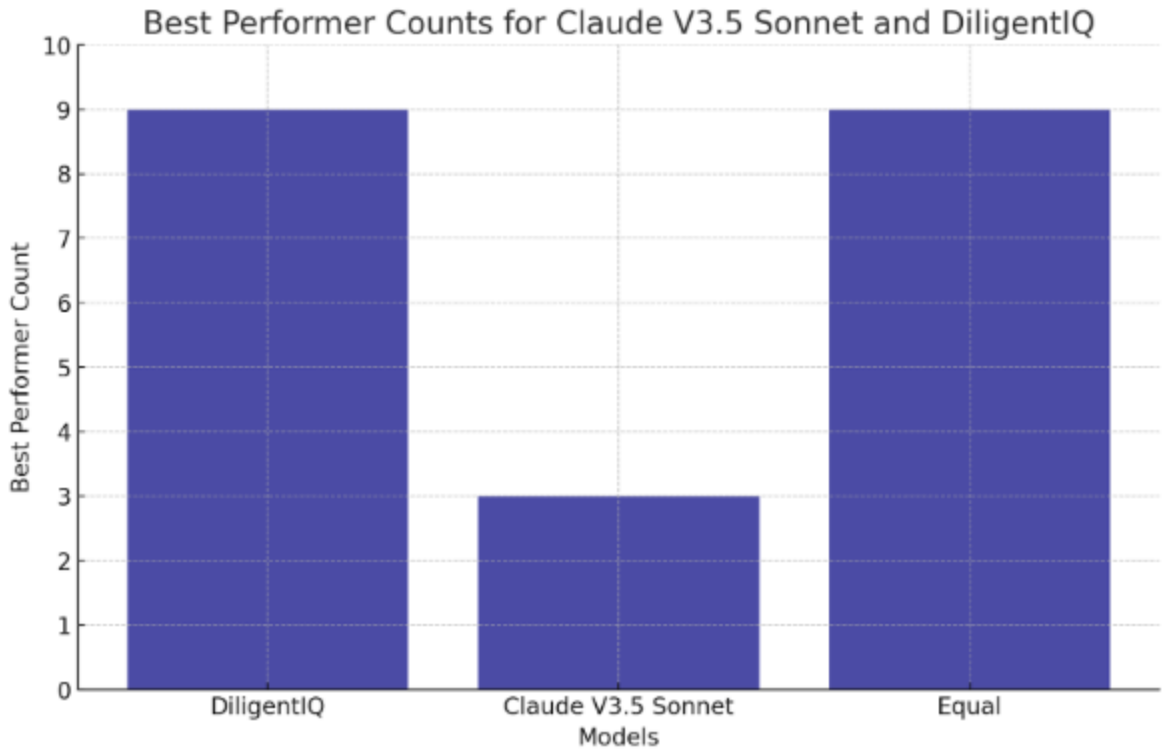
Quantitative Analysis

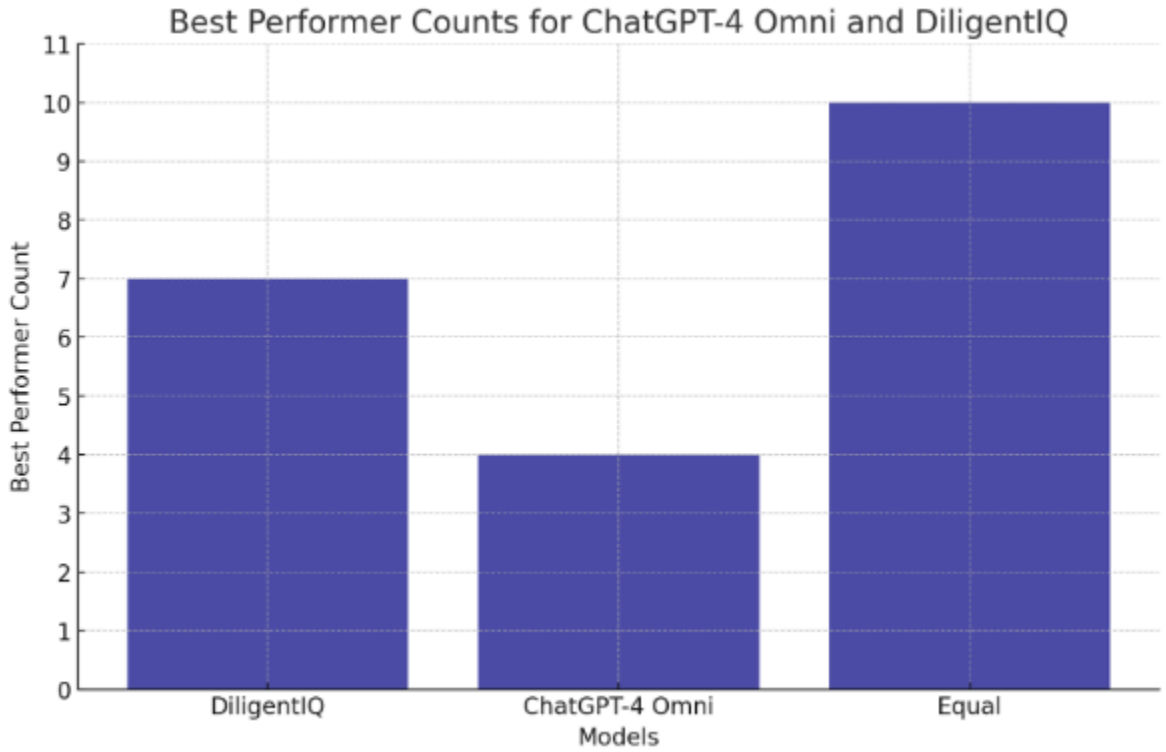
After the qualitative analysis was completed, the best performer for each prompt was chosen. This process was done separately for ChatGPT and Claude. The best performer was chosen between ChatGPT used inside and outside of DiligentIQ for each prompt, and the same was done for Claude. In order to choose the best performer, the performance scores and differences in content and structure were taken into account. If the model produced two responses with no clear “better” answer, the best performer was noted as Equal. In the table below, the best performer count is displayed for Claude within and outside DiligentIQ.

Model	Best Performer Count
DiligentIQ	9
Claude	3
Equal	9
Total	21

Model	Best Performer Count
DiligentIQ	7
ChatGPT	4
Equal	10
Total	21

As seen in the table, ChatGPT by itself outperformed ChatGPT within DiligentIQ only 3 times. In order to better visualize the findings, our research team generated bar graphs for the best performers, which are shown below.





These bar graphs display how using the LLMs within DiligentIQ leads to overall better responses. The models also produced several equal responses; however, when taking out those prompts, the likelihood of DiligentIQ performing better than the model on its own is much higher.

Key Differences between DiligentIQ and LLMs on their Own

In this testing, a key difference found between DiligentIQ and LLMs on their own is the thoroughness of responses. DiligentIQ was more thorough compared to ChatGPT and Claude for many prompts. The LLMs within DiligentIQ provided more structured, detailed, and nuanced responses compared to the LLMs on their own. Claude and ChatGPT on their own tended to generalize responses and give a broader overview. DiligentIQ was more likely to emphasize key metrics and kept responses targeted towards the prompt.

Another large difference between using the LLMs on their own versus in DiligentIQ is the amount of files you can query at a time. In DiligentIQ, the user can upload an entire VDR to the deal. In this case, the VDR contained 57 files. The maximum number of files that could be uploaded in ChatGPT was 10, and only 5 files were able to be uploaded at a time in Claude. Therefore, for each prompt, the best files were manually chosen in order to answer the prompt. The limited amount of files that can be uploaded into Claude and ChatGPT independently can lead to responses being less comprehensive and accurate, since important information from other sources could have been left out.

As for user experience, the main difference between using the LLMs on their own versus within DiligentIQ is that users can rely more heavily on DiligentIQ for providing accurate and relevant responses. When using the LLMs on their own for these prompts, there was a higher risk of responses having poor performances in accuracy and relevance. Responses with bad accuracy or relevance could lead to professionals making incorrect investment decisions, which is why it is extremely important for responses to be accurate. With using the LLMs in DiligentIQ, professionals can rely more heavily on Generative AI to assist them with analyzing documents and drawing conclusions.

Conclusion

The overall takeaway from this testing is that using LLMs within DiligentIQ can significantly improve response accuracy, relevance, and overall quality. DiligentIQ also can analyze a much larger batch of files within a chat compared to the LLMs on their own, which is extremely useful for answering prompts related to large datasets. However, Claude and ChatGPT are sufficient for providing general responses to prompts if that is what is desired. DiligentIQ is useful for more in-depth and specialized tasks.

Using LLMs within DiligentIQ can greatly improve workflow for professionals. The ability to handle a large number of files can allow for comprehensive and accurate responses. Additionally, DiligentIQ provides relevant responses, ensuring that only desired information will be extracted from documents. The accuracy and relevance of responses can help improve overall efficiency and reduce the chances of making investment mistakes.