SHIFT

The State of AI in Insurance

A Comparison of LLMs (Vol. III)

www.shift-technology.com

Executive summary

- **LLM technology is advancing at a rapid pace** with both new versions of existing models and entirely new models being introduced.
- **We are reaching a confluence point** with several models achieving highly comparable performance.
- **Price/performance comparisons** may be an important determining factor when selecting LLMs until new performance gains can be established.

From the editor

The introduction of the F1 score in **Vol. II** of this ongoing series of publications has allowed us to think a little differently about how we report our findings related to the evaluation of Large Language Model (LLM) performance when applied to specific insurance use cases. You will see that evolved thinking reflected in this report.

We believe the aggregated F1 score, both per scenario and overall, provides the relevant insights required to understand how the tested LLMs perform against common insurance industry use cases and if the costs associated with their deployment are in line with performance. This approach also makes it easier to understand how each model performs against what could be considered "simple scenarios" (information extraction from text fields, amounts, dates, etc.) as opposed to "complex scenarios" (tasks including several steps and/or information extraction from lists or fields that are themselves complex objects).

Based on advancements in LLM technology which have been introduced to the market since Vol. II our data science and research teams have included six new LLMs - GPT4o-Mini, Claude3.5 Sonnet, Mistral Large 2407, Llama3.1-405b, Llama3.1-70b, and Llama3.1-8b - into the testing. Command r and Command r+ were removed from evaluation.

As always, this report would not be possible without the efforts of Shift's data science and research teams.

LLM Model Comparison for Information Extraction, Select Insurance Documents

Methodology

The Shift Technology data science and research teams devised four test scenarios to evaluate the performance of 16 different publicly available LLMs: GPT3.5, GPT4, GPT4o, GPT4o-Mini, Mistral Large, Mistral Large 2407, Llama3.1-8b, Llama3.1-70b, Llama3.1-405b and their corresponding Llama3 variants, Claude3 Opus, Claude3 and 3.5 Sonnet, and Claude3 Haiku.

The scenarios include:

- Information extraction from English-language airline invoices (complex)
- Information extraction from Japanese-language property repair quotes (simple)
- Information extraction from French-language dental invoices (simple)
- Document classification of English-language documents associated with travel insurance claims (complex)

The LLMs were tested for:

Coverage - did the LLM in fact, extract data when the ground truth (the value we expect when we ask a model to predict something) showed that there was something to extract.

Accuracy - did the LLM present the correct information when something was extracted.

Prompt engineering for all scenarios was undertaken by the Shift data science and research teams. For each individual scenario, a single prompt was engineered and used by all of the tested LLMs. It is important to note that all the prompts were tuned for the GPT LLMs, which in some cases may impact measured performance.

Reading the Results

Evaluating LLM performance is based on the specific use case and the relative performance achieved. The table included in this report reflect that reality and are color-coded based on relative performance of the LLM applied to the use case, with shades of blue representing the highest relative performance levels, shades of red representing subpar relative performance for the use case, and shades of white representing average relative performance.

As such, a performance rating of 90% may be coded red when 90% is the lowest performance rating for the range associated with the specific use case. And while 90% performance may be acceptable given the use case, it is still rated subpar relative to how the other LLMs performed the defined task.

Context Size

Although still relevant, context size is fast becoming what one may consider table stakes in the evaluation of LLMs. Most of the leading models feature a 128k tokens context which is robust enough to support a majority of insurance use cases.

Results & analysis

LLM Metrics Comparison

F1 Score	GPT4o	GPT4o-Mini	GPT3.5	Claude3.5 Sonnet	Claude3 Opus	Claude3 Sonnet	Claude3 Haiku	Mistral Large 2407	Mistral Large	Llama 3.1-405b	Llama 3.1-70b	Llama 3.1-8b	Llama 3-70b	Llama 3-8b
Price 100k docs	€1,840	€112	€321	€2,503	€12,519	€2,503	€208	€2,604	€5,186	€3,471	€1,326	€168	€2,443	€238
French Dental Invoice	93.7%	90.0%	91.8%	94.3%	93.7%	93.3%	93.2%	93.5%	91.5%	93.1%	91.8%	90.4%	90.9%	89.0%
Japanese Home Quote	83.0%	78.4%	79.2%	83.0%	83.0%	82.2%	78.4%	82.5%	81.6%	83.3%	79.8%	71.0%	79.1%	72.1%
English Flight Invoice	82.6%	75.3%	69.9%	81.5%	82.2%	77.8%	72.9%	82.1%	78.7%	83.5%	79.2%	65.0%	75.0%	65.5%
Classif	89.5%	84.8%	81.5%	88.2%	87.1%	85.5%	78.2%	87.2%	70.7%	83.7%	68.3%	62.2%	78.5%	20.4%
Classif With Id	91.6%	85.6%		88.7%	86.7%	88.7%	79.3%	87.8%		87.2%	87.9%	67.5%		
Classif Without Id	91.3%	85.8%		88.1%	88.0%	86.2%	79.3%	88.0%		88.2%	87.7%	70.4%		
All use cases aggreg	87.2%	82.1%	80.6%	86.7%	86.5%	84.7%	80.7%	86.3%	80.6%	85.9%	79.7%	72.2%	80.9%	61.8%

Referring to the above table we can draw the following conclusions regarding the LLMs tested:

- **GPT4o** outperformed other models with an aggregate performance score of 87.2%. In addition, since Vol. II was released the price for GPT4o has effectively been cut in half. The model performs well on both simple and complex tasks, making GPT4o a solid choice for many organizations.
- **GPT4o-Mini** achieved a respectable aggregate performance score of 82.1%. Intended to replace GPT3.5, we witnessed slight performance degradation (~2%) on simple tasks when compared with its predecessor, but a marked performance increase (~3-5%) on complex tasks. Based on price/performance metrics, with GPT3.5 being approximately three times the cost of GPT4o-Mini, we believe the new model is an appropriate replacement for GPT3.5.
- At 86.7, **Claude3.5 Sonnet** achieved the second highest aggregate performance score. We found it interesting that this new model slightly outperformed Claude3 Opus but at a price comparable to Claude3 Sonnet. While its overall performance is comparable to GPT4o, it remains more expensive.
- **Mistral Large 2407** showed a performance improvement of approximately six percentage points over Mistral Large at nearly half the price. Its aggregate score of 86.3% puts it in close competition with GPT4o and Claude3.5 Sonnet. Here again, the GPT4o price reduction greatly impacts the price/performance ratio.

(Continued on next page)

(Continued)

- Although performance compared to the earlier Llama3 versions has significantly improved, **Llama3.1 70b** and **Llama3.1 8b** underperformed in aggregate and when asked to complete complex tasks. When compared against other LLMs applied to specific insurance scenarios they simply underachieve.
- With an aggregate performance score of 85.9% **Llama3.1 405b** is close to the leading models evaluated. However, its underperformance when compared with the leaders in "Classification of English-language Travel Documents" and its relatively higher cost make this model less attractive than several of the othersw tested.

Conclusion

The state of the art for LLMs is advancing at a rapid pace with both new versions and new models being released in short periods of time. It can be incredibly difficult to keep up. At the same time, we appear to be coming to an interesting confluence point where several of the models we tested achieved quite similar results. In the near future, price may be the key factor in determining which LLM is best suited for your organization until we begin to see more significant performance gaps emerge. At this time, the GPT models, specifically GPT4o and GPT4o-Mini are demonstrating the best price/performance comparisons when tested against these four insurance industry scenarios.