SHIFT

The State of AI in Insurance

A Comparison of LLM Performance (Vol. V)

www.shift-technology.com

Executive summary

- **Developers continue to introduce LLM models —** both brand new and extensions of existing LLM model families — bringing with them new questions about cost, performance and appropriateness for different use cases
- **We continue to find that "best" is a relative term** when comparing LLM performance, which is tightly bound to individual use case
- As the LLM landscape becomes more diverse, **understanding the intended purpose of an LLM** becomes an important evaluation criteria
- **The price/performance ratio continues** to be a critical metric for evaluating which LLM is right for each use case
- **Deepseek R1 demonstrates the viability** of large models coming from the open source community

From the editor

The LLM landscape continues to evolve at a rapid pace, which can make it feel impossible to keep up. Established models introduce new versions, and new players enter the mix. It becomes critically important to understand how these changes may impact the way LLMs are used in support of critical insurance processes and use cases.

This publication series began with the inaugural **The State of AI in Insurance report**, where we explored the performance of six different Large Language Models (LLMs) when applied to various insurance-specific use cases. Since that first publication, some of the models tested have been retired from the testing suite and new models added. Shift researchers do so to ensure that the report best reflects the current state-of-the-art of available LLMs, highlights models receiving significant interest from the technology community (e.g. Deepseek R1), and includes those most likely to be considered for deployment against insurance-specific use cases. The report is intended to not only compare relative performance against a set of predetermined tasks, but also illustrate the cost/performance comparisons associated with each of the LLMs tested.

In Vol. V, we are testing a total of 19 LLMs, 11 of which are new to the report, including Deepseek R1. We continue to use an F1 score generated for each model to report performance. The F1 score aggregates coverage and accuracy against two axes - the specific use case (e.g. French-language Dental Invoices) as well as the individual fields associated with the use case. This approach allows us to generate a single performance metric per use case as well as an aggregated overall score including the cost associated with analyzing 100,000 documents. The following formula was used to generate the F1 score: 2 x Cov x Acc / (Cov + Acc).

LLM Model Comparison for Information Extraction & Classification, Select Insurance Documents

Methodology

The data science and research teams devised four test scenarios to evaluate the performance of 19 different publicly available LLMs: GPT4o, OpenAI o1-preview*, OpenAI o1-mini*, OpenAI o3-mini*, GPT4o-mini, Deepseek R1*, Claude3.5 Sonnet v2*, Claude3.5 Sonnet, Claude3.5 Haiku*, Claude3 Haiku, Mistral Large 2411*, Mistral Large 2407, Llama3.3* (70b), Llama3.2* (90b & 11b), Llama3.1(405b, 70b & 8b), and Microsoft Phi4*

The scenarios include:

- Information extraction from English-language airline invoices (complex)[1]
- Information extraction from Japanese-language property repair quotes (simple)[2]
- 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 tables 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.

A Note on Costs

Beginning with Vol. 1 of this benchmark report we based our cost estimate related to processing 100k documents on the assumption of 0.5k tokens for the output. However, this assumption does not hold true for the reasoning models now included in the testing. By definition these models will output dedicated additional reasoning tokens. As such, we updated the cost computation with the assumption of 1.5k tokens for the output for reasoning models.

---

*= new to the report

[1] tasks including several steps and/or information extraction from lists or fields that are themselves complex objects

[2] information extraction from text fields, amounts, dates, etc.

Results & analysis

LLM Metrics Comparison

F1 Score	GPT4o	GPT4o-Mini	o1-preview	o3-mini	o1-mini	Deepseek R1	Claude3.5 Sonnet v2	Claude3.5 Haiku	Claude3.5 Sonnet	Claude3 Opus	Claude3 Sonnet
Price 100k docs	€1,840	€112	€22,800	$1,672	€1,672	€0	€2,503	€698	€2,503	€12,519	€2,503
French Dental Invoice	93.7%	90.0%	93.2%	94.7%	92.7%	93.9%	94.0%	91.6%	94.3%	93.7%	93.3%
Japanese Home Quote	83.0%	78.4%	84.9%	82.2%	82.0%	82.2%	82.7%	83.7%	83.0%	83.0%	82.2%
English Flight Invoice	82.6%	75.3%	82.0%	78.3%	78.9%	78.9%	79.7%	78.7%	81.5%	82.2%	77.8%
Classif WithoutId	91.3%	85.8%	91.0%	89.1%	88.0%	89.5%	89.5%	87.0%	88.1%	88.0%	86.2%
All use cases aggreg	**87.6%**	**82.4%**	**87.8%**	**86.1%**	**85.4%**	**86.1%**	**86.5%**	**85.3%**	**86.7%**	**86.7%**	**84.9%**

F1 Score	Claude3 Haiku	Mistral Large 2411	Mistral Large 2407	Llama3.3-70b	Llama3.2-90b	Llama3.2-11b	Llama3.1-405b	Llama3.1-70b	Llama3.1-8b	Phi4
Price 100k docs	€208	€2,604	€2,604	€344	€989	€179	€3,471	€1,326	€168	€95
French Dental Invoice	93.2%	94.1%	93.5%	90.9%	92.1%	90.1%	93.1%	91.8%	90.4%	92.1%
Japanese Home Quote	78.4%	82.5%	82.5%	79.2%	79.8%	71.1%	83.3%	79.8%	71.0%	81.0%
English Flight Invoice	72.9%	82.1%	82.1%	77.2%	78.5%	65.6%	83.5%	79.2%	65.0%	78.8%
Classif WithoutId	79.3%	88.0%	88.0%	86.9%	84.3%	69.5%	88.2%	87.7%	70.4%	81.5%
All use cases aggreg	**80.9%**	**86.7%**	**86.5%**	**83.5%**	**83.7%**	**74.1%**	**87.0%**	**84.6%**	**74.2%**	**83.3%**

The latest OpenAI models, o1-preview, o1-mini and o3-mini, delivered excellent performance, with o1-preview achieving the highest aggregate F1 score. However, their respective costs (o1-preview is 10x more expensive than GPT4o) and computation time (~ 45-60 seconds for o1-preview and 15 seconds on average for o1-mini and o3-mini) makes them unlikely to be used in production for the use cases present in this benchmark. This is not surprising as these latest OpenAI models were designed for complex reasoning tasks as opposed to information extraction or classification.

Much like the OpenAI models tested, Deepseek R1 performed well. Its results were comparable to that of the OpenAI o3-mini and slightly below those of OpenAI o1-preview. What is perhaps most interesting in this evaluation is that Deepseek R1 costs significantly less to train than the OpenAI models tested. In addition, being an open source LLM, Deepseek R1 is basically free to use. Deepseek R1 demonstrates that an effective large model can come from the world of open source and deliver performance comparable with commercial options.

(Continued on next page)

(Continued)

At the same time, our testing revealed two areas of concern. We experienced high latency — reaching 10 minutes for some responses. Despite repeated testing we could not definitively determine if the latency is due to the model itself or the infrastructure on which it was deployed. For the described use cases, the latency experienced would make the models unsuitable in production. Additionally, we found that the output from Deepseek R1 often required post processing prior to automated parsing. As a reasoning model this LLM frequently output intermediary results that were read as final, impacting downstream parsing.

New versions of the Mistral and Claude models — respectively Mistral Large 2411 and Claude3.5 Sonnet v2 — performed on par with their earlier versions. Our testing did not reveal any significant performance increases. Interestingly, Claude3.5 Haiku showed a significant performance gain — nearly +5% correlating to 3x price increase when compared to its last version. We are now seeing that Claude3.5 Haiku demonstrates near "big model" performance while remaining cost competitive.

Digging into the Llama models we find that the performance of the small version, Llama3.2-11b, could be considered disappointing. Overall it tested comparable to its previous 3.1 version in terms of performance and cost. When looking at the two medium models — Llama3.3-70b and Llama3.2-90b — we see that these models are similar in terms of performance to both each other and the previous Llama3.1-70b version. However, we do witness an improvement on the cost side, with Llama3.3-70b being 3x less expensive than Llama3.2-90b and 4x less expensive than Llama3.1-70b. These performance and price improvements related to Llama3.3-70b makes the model comparable to GPT4o-mini in terms of both performance and price.

While we found that Phi4 delivered performance comparable to GPT4o-mini and featured similar pricing, its smaller context window (16k) may make it a less attractive option for those use cases where the context window could be exceeded. However, for those use cases where the size of the context window is not a consideration, based on our evaluation Phi4 could be a reasonable substitution for GPT4o-mini.

Conclusion

When we began producing this report with Vol. 1, we witnessed a clear distinction between performance of what would be considered "big models" and those that would be considered "small models." And although we did begin to see performance equalize to a certain extent in later reports, our team still observed that the highest performance was associated with the highest cost.

With this round of testing however we are seeing that what was once a clear distinction between big models (very good performance/high cost) and small models (good performance/reasonable cost) is becoming much less clear. The emergence of reasoning models — Deepseek R1 and the OpenAI models (o1-preview, o1-mini and o3-mini) — creates additional considerations. While we experienced very good performance on our use cases it is apparent that these LLMs are most appropriate for complex reasoning tasks. With Deepseek R1 we are also seeing that the open source community can be a viable route for large model development.

Claude3.5 Haiku demonstrates performance approximating big models while costing users 3-4x less. However, we must note that Claude3.5 Haiku is still 6x more expensive than most small models. Still, our testing demonstrates that Claude3.5 Haiku could be a good alternative to small models for those use cases requiring excellent performance while exercising good cost management.