SHIFT

The State of AI in Insurance

A Comparison of LLMs (Vol. II)

www.shift-technology.com

Executive summary

- **The LLM market is rapidly evolving,** with various models now available which are appropriate for a variety of use cases
- **Determining which LLM is best** for which use case involves comparing context size, overall cost and performance
- **Focused prompt engineering and tuning** can be the difference between exceptional and disappointing performance

From the editor

Advances in the use of artificial intelligence (AI) and generative artificial intelligence (GenAI) to improve critical insurance processes continue to captivate the industry. At the same time, it can be incredibly difficult to navigate the rapidly changing landscape and make the best decision about how to implement these innovations to reap the best results.

In the inaugural The State of AI in Insurance report, we explored the performance of six different Large Language Models (LLMs) when applied against several insurance-specific use cases. Shift data scientists and researchers sought not only to 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. II, we are testing eight new LLMs and have retired two that appeared in the previous report. The newly tested models include Llama3-8b, Llama3-70b, GPT4o, Command r, Command r+, Claude3 Opus, Claude3 Sonnet, and Claude3 Haiku. The Llama2 models which appeared in the inaugural report have been removed from the comparison and replaced with the Llama3 models. Llama3 models are more representative of the current state-of-the-art for available LLMs.

Further, the report now features a new table highlighting an F1 score generated for each model. For this report 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. The 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).

Thank you to the Shift data science and research teams that make this report possible.

LLM Model Comparison for Information Extraction, Select Insurance Documents

Methodology

The data science and research teams devised four test scenarios to evaluate the performance of 11 different publicly available LLMs: GPT3.5, GPT4, GPT4o, Mistral Large, Llama3-8b, Llama3-70b, Command r, Command r+, Claude3 Opus, Claude3 Sonnet, and Claude3 Haiku.

The scenarios include:

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

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 Tables

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.

Context Size

Each LLM features a specific context size, defined as the maximum number of tokens the model can handle in total between the input prompt and the output response. Fundamentally, a model with a small context size is not suitable to analyze long documents. As such, we look at the context size of the model as one of the factors potentially impacting performance. However, we cannot make the direct assumption that larger context size equals greater performance. For example, Llama3-70b has a relatively small context size of 8k but often delivers comparable performance to models with much larger context (e.g. Command r+ at 128k). Similarly Claude3 Haiku and Gpt3.5-turbo show comparable results while Claude3 context is approximately 10x that of Gpt3.5-turbo.

Model	Context size
mistral-large	32k
llama3-70b-instruct	8k
llama3-8b-instruct	8k
gpt4o	128k
gpt-4-0125-preview	128k
gpt-3.5-turbo-0125	16k
command r+	128k
command r	128k
claude3-sonnet	200k
claude3-opus	200k
claude3-haiku	200k

Results & analysis

English-language Airline Invoices

In this scenario, the LLMs being tested analyzed 85 anonymized English-language invoices.

The extraction prompt sought the following results:

- Provider Name
- Start Date
- End Date
- Document Date
- Booking Number
- Flight Numbers (list of all flight numbers)
- Credit Card 4 LastDigits
- Currency
- Basic Fare All Passengers
- Taxes And Fees All Passengers (list of all taxes and fees)
- Additional Fees All Passengers (list of additional fees)
- Total Amount
- Total Paid Amount
- Payments (complex field: list of Payment, object containing the following fields: Payment Date, Amount, Status)
- Travellers (complex field: list of Traveller, object containing the following fields: Traveller Name, Basic Fare, Total Taxes, Total Amount)

English Flight Invoice	GPT4		GPT4o		GPT3.5		Claude3 Opus		Claude3 Sonnet		Claude3 Haiku	
Metric Name	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc
Provider Name	98.5%	67.1%	100.0%	68.8%	98.5%	59.5%	100.0%	63.0%	95.5%	61.1%	100.0%	61.2%
Start Date	98.3%	83.1%	100.0%	84.8%	98.3%	76.1%	100.0%	83.6%	94.8%	85.7%	100.0%	63.5%
End Date	100.0%	82.7%	100.0%	82.1%	97.9%	66.2%	100.0%	69.8%	100.0%	79.3%	100.0%	51.8%
Document Date	95.3%	80.3%	95.3%	79.1%	96.9%	67.5%	93.8%	73.2%	96.9%	71.6%	100.0%	62.4%
Booking Number	96.7%	71.8%	96.7%	82.8%	98.3%	3.8%	88.3%	70.3%	93.3%	29.3%	100.0%	2.4%
Flight Numbers	98.5%	65.5%	100.0%	61.2%	98.5%	50.0%	100.0%	62.4%	94.0%	53.8%	100.0%	48.2%
Credit Card 4 Last Digits	98.0%	94.2%	100.0%	96.2%	98.0%	90.7%	100.0%	96.2%	94.1%	95.9%	100.0%	89.3%
Currency	98.3%	96.7%	98.3%	95.2%	98.3%	93.7%	91.7%	96.5%	90.0%	88.5%	100.0%	71.4%
Basic Fare All Passengers	97.0%	51.7%	100.0%	54.5%	100.0%	33.3%	100.0%	61.5%	90.9%	57.7%	100.0%	49.2%
Taxes And Fees All Passengers	96.9%	44.4%	100.0%	45.3%	100.0%	23.2%	100.0%	51.9%	90.6%	48.0%	100.0%	41.8%
Additional Fees All Passengers	91.7%	28.0%	91.7%	34.8%	75.0%	12.5%	91.7%	33.3%	83.3%	43.8%	91.7%	30.4%
Additional Fee Insurance	100.0%	86.7%	100.0%	92.9%	69.2%	75.0%	100.0%	100.0%	92.3%	100.0%	100.0%	81.3%
Tota lAmount	93.2%	91.4%	98.3%	98.3%	91.5%	89.7%	96.6%	93.2%	88.1%	92.6%	96.6%	88.7%
Total Paid Amount	92.5%	90.6%	62.3%	100.0%	69.8%	81.4%	83.0%	87.5%	83.0%	93.6%	96.2%	73.3%
Payment Date	95.7%	38.2%	39.1%	32.0%	82.6%	34.7%	73.9%	42.1%	82.6%	32.1%	100.0%	31.3%
Payment Status	89.3%	76.6%	46.4%	86.7%	58.9%	66.0%	75.0%	79.2%	78.6%	74.6%	76.8%	65.2%
Payment Amount	88.7%	96.9%	42.3%	100.0%	60.6%	86.0%	73.2%	98.1%	78.9%	91.5%	80.3%	83.3%
Traveller Basic Fare	82.8%	75.4%	58.6%	87.2%	50.0%	32.4%	74.1%	72.9%	67.2%	56.7%	89.7%	57.5%
Traveller Total Taxes	83.0%	60.7%	63.8%	71.8%	53.2%	36.8%	74.5%	56.9%	63.8%	42.4%	89.4%	47.1%
Traveller Total Amount	83.1%	80.3%	57.6%	84.6%	49.2%	41.2%	72.9%	69.0%	66.1%	43.9%	86.4%	55.2%

(Continued on next page)

(Continued)

English Flight Invoice	Mistral Large		Command r+		Command r		Llama3-70b		Llama3-8b	
Metric Name	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc
Provider Name	100.0%	63.8%	100.0%	61.4%	92.5%	56.3%	98.5%	64.1%	95.5%	56.1%
Start Date	100.0%	79.7%	100.0%	73.6%	96.6%	65.8%	98.3%	78.3%	96.6%	63.9%
End Date	100.0%	79.6%	97.9%	75.9%	93.6%	58.6%	100.0%	67.7%	95.7%	58.8%
Document Date	100.0%	70.5%	100.0%	59.5%	95.3%	60.0%	95.3%	68.4%	96.9%	50.6%
Booking Number	98.3%	55.1%	100.0%	42.5%	93.3%	6.3%	98.3%	4.9%	96.7%	10.8%
Flight Numbers	100.0%	61.2%	100.0%	55.3%	94.0%	52.5%	98.5%	64.6%	97.0%	44.6%
Credit Card 4 Last Digits	100.0%	94.3%	98.0%	85.2%	94.1%	69.8%	92.2%	90.0%	92.2%	75.4%
Currency	100.0%	75.0%	100.0%	70.6%	95.0%	72.2%	95.0%	89.1%	96.7%	71.6%
Basic Fare All Passengers	100.0%	57.4%	97.0%	45.8%	97.0%	34.2%	90.9%	54.7%	93.9%	40.3%
Taxes And Fees All Passengers	96.9%	42.6%	96.9%	36.5%	96.9%	23.7%	87.5%	37.3%	90.6%	31.4%
Additional Fees All Passengers	91.7%	13.8%	91.7%	26.1%	91.7%	9.5%	66.7%	19.0%	83.3%	8.0%
Additional Fee Insurance	100.0%	92.9%	100.0%	92.9%	84.6%	19.3%	92.3%	92.3%	100.0%	65.0%
Tota lAmount	96.6%	91.5%	96.6%	84.7%	89.8%	68.0%	91.5%	93.0%	93.2%	86.4%
Total Paid Amount	90.6%	84.6%	92.5%	87.0%	88.7%	58.3%	83.0%	82.0%	83.0%	66.0%
Payment Date	73.9%	32.0%	69.6%	42.9%	100.0%	24.7%	91.3%	40.8%	52.2%	40.0%
Payment Status	73.2%	70.7%	89.3%	74.6%	71.4%	47.6%	78.6%	72.1%	69.6%	79.6%
Payment Amount	74.6%	91.4%	87.3%	93.9%	74.6%	61.0%	78.9%	88.5%	66.2%	90.0%
Traveller Basic Fare	70.7%	60.3%	60.3%	44.9%	60.3%	22.5%	51.7%	63.6%	41.4%	35.0%
Traveller Total Taxes	68.1%	43.3%	72.3%	35.7%	59.6%	23.4%	46.8%	51.2%	36.2%	32.0%
Traveller Total Amount	71.2%	54.2%	67.8%	34.9%	59.3%	29.7%	50.8%	65.9%	37.3%	29.8%

Analysis

The introduction of the new LLMs into the testing environment produced several interesting results. On what have been classified as "simple fields," which we define as fields containing simple data types such as date, type, amount, and other fields whose value is unique and not a list of elements, GPT4o, GPT4, and Claude3 Opus outperformed all of the other models with GPT4o leading the top three contenders. There is however one exception. In the case of Total Paid Amount coverage, GPT4o underperformed relative to the other two leading LLMs. The reason for this exception is not immediately evident and will require additional research and experimentation to determine the cause.

Claude3 Sonnet, Mistral Large, Command r+ and Command r demonstrated performance close to, but still slightly behind the leading LLMs. And finally, the Llama3 models, GPT3.5 and Claude3 Haiku performed similarly, but behind the seven leading models. We do witness that the performance gap highlighted in the inaugural report between the Llama2 models and the other models has tightened significantly with the introduction of Llama3 into the testing. This may be due to the larger base context size (4k vs. 8k) for Llama3.

For what have been identified as "complex fields," which we define as fields that represent complex objects or whose value is a list of objects, we find that GPT4 and Claude3 Opus are the best models for these particular tasks, with GPT4 slightly outperforming Claude3 Opus.

Claude3 (both Sonnet and Haiku), Mistral Large and Llama3-70b did not perform as well as what would be considered the leading models. Interestingly we find that performance for these models is highly dependent on which field is being analyzed with our research showing that each model slightly outperforms the others depending on the particular data requested.

Finally, in this scenario we found GPT4o to be very good in terms of accuracy but surprisingly bad in terms of coverage. While this may imply that the model is not able to retrieve complex information, it may simply indicate that prompt engineering/tuning on these fields would most likely fix the issue.

Japanese-language Property Repair Quotes

We evaluated 100 anonymized Japanese property repair quotes associated with different service providers. The format of these documents were not standardized.

The extraction prompt requested data from the following fields:

- Provider Name
- Provider Address
- Postcode
- Provider email
- Tax Amount
- Total Amount with Tax
- Discount Amount

Japanese Home Quote	GPT4		GPT4o		GPT3.5		Claude3 Opus		Claude3 Sonnet		Claude3 Haiku	
Metric Name	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc
Provider Name	98.9%	73.1%	98.9%	74.2%	98.9%	68.1%	98.9%	74.5%	100.0%	76.8%	98.9%	68.7%
Provider Address	91.2%	70.2%	89.0%	69.5%	93.4%	62.8%	92.3%	68.2%	93.4%	58.1%	93.4%	66.7%
Post Code	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Provider Email	100.0%	63.6%	100.0%	70.0%	100.0%	63.6%	100.0%	70.0%	100.0%	70.0%	100.0%	43.8%
Tax Amount	100.0%	86.0%	100.0%	85.9%	96.4%	77.5%	100.0%	81.8%	100.0%	83.5%	100.0%	78.8%
Total Amount With Tax	100.0%	97.0%	100.0%	93.9%	97.9%	95.9%	100.0%	96.0%	100.0%	94.0%	99.0%	94.9%
Discount Amount	100.0%	9.6%	100.0%	5.5%	73.3%	10.0%	100.0%	9.1%	100.0%	9.1%	100.0%	10.3%

Japanese Home Quote	Mistral Large		Command r+		Command r		Llama3-70b		Llama3-8b	
Metric Name	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc
Provider Name	98.9%	72.3%	100.0%	65.6%	98.9%	65.7%	100.0%	72.6%	98.9%	66.3%
Provider Address	90.1%	69.9%	90.1%	46.4%	92.3%	50.0%	92.3%	67.1%	94.5%	66.7%
Post Code	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	15.9%	100.0%
Provider Email	100.0%	63.6%	100.0%	53.8%	100.0%	26.9%	100.0%	58.3%	100.0%	38.9%
Tax Amount	100.0%	85.7%	95.2%	77.5%	98.8%	60.4%	100.0%	70.7%	100.0%	51.0%
Total Amount With Tax	99.0%	92.9%	100.0%	96.0%	99.0%	76.8%	100.0%	92.9%	99.0%	88.9%
Discount Amount	93.3%	4.9%	53.3%	18.5%	93.3%	3.7%	100.0%	3.0%	100.0%	2.1%

Analysis

Overall, GPT4o, GPT4, Mistral and Claude3 (Opus and Sonnet) performed best in this scenario with all achieving similar results.

We witnessed a small degradation of performance related to GPT3.5, Claude3 Haiku and Llama3-70b. And while these models did achieve equivalent performances on most fields, underperformance on others impacted the assessment.

Command r, Command r+ and Llama3-8b models performed well enough in comparison to other models when considering coverage. However, they are clearly behind when assessing accuracy. That could indicate that these models have a difficult time with the Japanese language and we intended to further investigate this phenomenon.

On initial examination, the accuracy on textual fields (e.g. Provider Name, Provider Address, Provider Email) may seem a bit low. However, it is not entirely unexpected seeing that it is impossible to validate the output of the model with a structured format in the same way that you can for amounts or dates as those particular metrics are very strict and require the ground truth and prediction to be exactly the same.

We are surprised with the consistent underperformance of the fields Post Code and Discount Amount across all models. It is not apparent what would cause this. Further investigation will be conducted.

French-language dental invoices

For this scenario we applied the LLMs being tested against 119 anonymized French dental invoices. 79 of which would be considered templated with the remaining 60 invoices were randomly selected. The resulting dataset would be described as approximately 60 percent templated. This methodology was selected to best represent a typical dental insurance provider's dataset.

We asked each LLM to extract the following:

- Document Date
- Provider Name
- Raw Provider FINESS (Fichier National des Établissements Sanitaires et Sociaux)
- Provider RPPS (Répertoire Partagé des Professionnels de Santé)
- Provider Postcode
- Total Incurred Amount
- Paid Amount

	GPT4		GPT4o		GPT3.5		Claude3 Opus		Claude3 Sonnet		Claude3 Haiku	
Metric Name	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc
Document Date	100.0%	95.7%	100.0%	95.7%	99.3%	93.5%	100.0%	97.1%	100.0%	95.7%	100.0%	97.1%
Provider Name	100.0%	95.0%	100.0%	95.0%	100.0%	93.5%	100.0%	94.2%	100.0%	94.2%	100.0%	95.0%
Raw Provider Finess	65.1%	61.9%	74.4%	50.0%	58.1%	53.7%	67.4%	71.8%	62.8%	65.0%	65.1%	61.9%
Provider Rpps	97.1%	92.5%	96.2%	95.1%	92.3%	94.9%	98.1%	93.4%	89.4%	92.7%	96.2%	95.1%
Provider Post Code	99.3%	99.3%	100.0%	98.6%	99.3%	99.3%	99.3%	98.6%	99.3%	98.6%	99.3%	99.3%
Total Incurred Amount	100.0%	97.8%	100.0%	97.1%	100.0%	96.4%	100.0%	97.1%	100.0%	97.8%	100.0%	97.8%
Paid Amount	100.0%	69.2%	98.5%	81.8%	95.6%	74.4%	98.5%	73.6%	98.5%	83.8%	98.5%	71.8%

	Mistral Large		Command r+		Command r		Llama3-70b		Llama3-8b	
Metric Name	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc
Document Date	100.0%	96.4%	100.0%	95.7%	95.0%	97.1%	100.0%	95.7%	99.3%	88.4%
Provider Name	100.0%	94.2%	100.0%	92.8%	95.0%	96.4%	100.0%	94.2%	99.3%	91.3%
Raw Provider Finess	69.8%	64.4%	72.1%	76.9%	82.1%	100.0%	62.8%	61.9%	65.1%	65.9%
Provider Rpps	91.3%	92.9%	89.4%	92.8%	93.0%	95.9%	94.2%	94.1%	95.2%	94.1%
Provider Post Code	100.0%	97.8%	99.3%	98.6%	97.1%	97.8%	99.3%	99.3%	89.1%	97.6%
Total Incurred Amount	100.0%	98.6%	100.0%	98.6%	96.4%	97.8%	100.0%	97.8%	99.3%	94.9%
Paid Amount	98.5%	55.8%	98.5%	50.4%	45.0%	85.3%	98.5%	49.6%	95.6%	50.8%

Analysis

GPT4o, GPT4, and Claude3 (all versions) performed best overall in this scenario and all demonstrated similar performance. However, it should be noted that GPT4o and Claude3 Opus are extremely close in terms of performance, and slightly outperform the other models.

We find that Command r+, Llama3-70b and Mistral Large are comparable to GPT3.5 while the other Llama models and Command r lag.

The witnessed underperformance for the field Provider FINESS (an identifier associated with the national directory managed by the digital health agency) could be due to the fact that French health invoices do not always clearly identify the FINESS or other provider identifiers (AM or SIRET). This confusion could impact the models' ability to retrieve the information but also the quality of the ground truths for this field.

English-language documents for travel claims

This scenario used 405 anonymized English-language documents provided in support of travel claims.

The prompt will ask the model to:

- Classify each page
- Group the pages related to the same document (as several documents can be located in a same file)
- Output each file as a list of segmented documents, where each element contains the document type and a span indicating the start and end page in the file

As with the other scenarios, we report the typical coverage and accuracy metrics for each document type individually. In addition, we include two aggregated metrics:

- Perfect Classif: Here we consider an output of the model correct when all the segmented documents in a file are error-free (document type and page span)
- Perfect Types: Here we consider an output of the model correct when all the document types in a file are error-free (meaning there could be errors in the page spans)(PerfectTypes).

As with the other scenarios, we report the typical coverage and accuracy It is important to note that between the inaugural report and Vol. 2 the prompt for this scenario was tuned slightly. After generating surprising underperformance for GPT4o, our investigation uncovered an error in the prompt. Specifically we were asking the model to output a “markdown JSON" instead of a “JSON” which resulted in the addition of a separator that we were not expecting and were not parsing properly. The prompt was tuned to fix this and another slight ordering error.

Classif	GPT4		GPT4o		GPT3.5		Claude3 Opus		Claude3 Sonnet		Claude3 Haiku		Mistral Large		Command r+		Command r		Llama3-70b		Llama3-8b	
Metric Name	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc
Receipt - Airplane	92.8%	85.2%	95.2%	88.2%	85.5%	77.1%	96.4%	80.9%	95.2%	74.7%	88.0%	54.1%	95.2%	86.4%	66.3%	69.1%	60.2%	69.5%	85.5%	79.3%	15.7%	45.0%
Receipt - Hotel / Rental reservations	85.0%	80.0%	85.0%	80.0%	75.0%	68.4%	90.0%	78.9%	90.0%	76.2%	85.0%	73.7%	85.0%	71.4%	75.0%	54.2%	60.0%	66.7%	75.0%	73.7%	0.0%	0.0%
Receipt - Activities reservations	83.3%	36.4%	16.7%	0.0%	33.3%	50.0%	33.3%	16.7%	16.7%	20.0%	16.7%	33.3%	33.3%	25.0%	0.0%	0.0%	16.7%	16.7%	50.0%	30.0%	0.0%	0.0%
Receipt - Cruises	91.7%	77.8%	91.7%	80.8%	79.2%	66.7%	79.2%	72.7%	87.5%	70.0%	66.7%	76.2%	91.7%	63.3%	75.0%	69.2%	33.3%	53.8%	75.0%	64.0%	4.2%	50.0%
Receipt - Train	100.0%	100.0%	66.7%	100.0%	100.0%	33.3%	66.7%	100.0%	100.0%	66.7%	100.0%	66.7%	66.7%	25.0%	33.3%	100.0%	66.7%	50.0%	100.0%	66.7%	0.0%	0.0%
Bank / Credit card statement	93.8%	90.3%	100.0%	88.2%	43.8%	93.3%	90.6%	89.7%	93.8%	84.4%	71.9%	95.8%	81.3%	92.9%	68.8%	74.1%	37.5%	69.2%	75.0%	82.1%	3.1%	33.3%
Cancellation policy	33.3%	50.0%	33.3%	33.3%	0.0%	0.0%	66.7%	44.4%	16.7%	100.0%	16.7%	50.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	33.3%	100.0%	0.0%	0.0%
Cancellation proof	83.3%	69.6%	87.5%	70.8%	41.7%	81.8%	83.3%	70.8%	62.5%	76.5%	33.3%	75.0%	50.0%	62.5%	29.2%	87.5%	29.2%	85.7%	37.5%	70.0%	0.0%	0.0%
Medical bills	88.9%	50.0%	88.9%	87.5%	66.7%	71.4%	66.7%	71.4%	77.8%	85.7%	44.4%	30.0%	66.7%	55.6%	66.7%	71.4%	55.6%	71.4%	66.7%	71.4%	0.0%	0.0%
Medical report	86.7%	90.9%	98.3%	88.9%	89.2%	91.2%	97.5%	92.6%	96.7%	88.0%	80.0%	90.1%	27.5%	85.7%	85.0%	94.3%	60.8%	92.0%	77.5%	87.3%	6.7%	87.5%
Proof of payment	66.7%	53.8%	41.7%	44.4%	16.7%	66.7%	33.3%	44.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	33.3%	80.0%	25.0%	50.0%	8.3%	100.0%	0.0%	0.0%
English-language Travel Documents claim form 1 - Front page	100.0%	100.0%	100.0%	75.0%	66.7%	28.6%	100.0%	20.0%	100.0%	42.9%	100.0%	25.0%	33.3%	0.0%	66.7%	14.3%	66.7%	22.2%	33.3%	33.3%	0.0%	0.0%
English-language Travel Documents claim form 1 - Claimant information / Agency details	100.0%	100.0%	100.0%	80.0%	25.0%	100.0%	75.0%	50.0%	100.0%	66.7%	100.0%	50.0%	25.0%	100.0%	25.0%	11.1%	25.0%	16.7%	25.0%	25.0%	0.0%	0.0%
English-language Travel Documents claim form 1 - Details of loss / Incident description	100.0%	100.0%	100.0%	100.0%	50.0%	100.0%	100.0%	44.4%	75.0%	60.0%	100.0%	50.0%	25.0%	100.0%	50.0%	20.0%	25.0%	14.3%	25.0%	100.0%	0.0%	0.0%
English-language Travel Documents claim form 1 - Claimed expenses / Payment	100.0%	100.0%	100.0%	100.0%	50.0%	100.0%	100.0%	57.1%	100.0%	80.0%	100.0%	50.0%	25.0%	100.0%	25.0%	11.1%	25.0%	20.0%	25.0%	100.0%	0.0%	0.0%
English-language Travel Documents claim form 1 - Required documentation	100.0%	80.0%	100.0%	100.0%	50.0%	100.0%	100.0%	57.1%	100.0%	80.0%	100.0%	50.0%	0.0%	0.0%	25.0%	11.1%	50.0%	28.6%	25.0%	100.0%	0.0%	0.0%
English-language Travel Documents claim form 1 - Authorization and assignment	100.0%	100.0%	100.0%	50.0%	50.0%	40.0%	100.0%	36.4%	100.0%	66.7%	100.0%	36.4%	25.0%	100.0%	50.0%	16.7%	25.0%	20.0%	25.0%	20.0%	0.0%	0.0%
English-language Travel Documents claim form 2 - General information	100.0%	100.0%	100.0%	100.0%	0.0%	0.0%	100.0%	100.0%	0.0%	0.0%	100.0%	20.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
English-language Travel Documents claim form 2 - Details of trip cancellation / trip interruption / trip delay	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
English-language Travel Documents claim form 2 - Claimed expenses and authorization	100.0%	100.0%	100.0%	100.0%	0.0%	0.0%	100.0%	100.0%	0.0%	0.0%	100.0%	20.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

(Continued on next page)

(Continued)

Classif	GPT4		GPT4o		GPT3.5		Claude3 Opus		Claude3 Sonnet		Claude3 Haiku		Mistral Large		Command r+		Command r		Llama3-70b		Llama3-8b	
Metric Name	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc
English-language Travel Documents claim form 3 - Front page and summary information	100.0%	100.0%	100.0%	100.0%	40.0%	33.3%	100.0%	100.0%	80.0%	50.0%	60.0%	42.9%	40.0%	50.0%	80.0%	40.0%	60.0%	42.9%	40.0%	50.0%	0.0%	0.0%
English-language Travel Documents claim form 3 - Disclosures	100.0%	100.0%	100.0%	100.0%	50.0%	100.0%	100.0%	100.0%	50.0%	100.0%	50.0%	16.7%	0.0%	0.0%	100.0%	25.0%	100.0%	33.3%	0.0%	0.0%	0.0%	0.0%
English-language Travel Documents claim form 3 - Assignment and authorization	85.7%	100.0%	42.9%	100.0%	14.3%	100.0%	42.9%	66.7%	57.1%	75.0%	14.3%	20.0%	42.9%	100.0%	42.9%	33.3%	28.6%	33.3%	0.0%	0.0%	0.0%	0.0%
English-language Travel Documents physician statement form 1 - Insured and physician information	100.0%	100.0%	100.0%	50.0%	0.0%	0.0%	100.0%	100.0%	100.0%	100.0%	100.0%	14.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
English-language Travel Documents physician statement form 1 - Patient's diagnosis	100.0%	50.0%	100.0%	50.0%	0.0%	0.0%	100.0%	50.0%	100.0%	33.3%	100.0%	20.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Proof of death - Death certificate or obituary	62.5%	90.9%	81.3%	92.9%	81.3%	92.9%	68.8%	91.7%	81.3%	92.9%	50.0%	80.0%	18.8%	75.0%	75.0%	92.3%	18.8%	100.0%	37.5%	100.0%	6.3%	100.0%
Other	38.6%	66.7%	52.6%	78.4%	21.1%	42.9%	59.6%	40.4%	24.6%	52.0%	31.6%	38.7%	50.9%	50.0%	38.6%	53.6%	35.1%	38.1%	19.3%	68.8%	1.8%	16.7%
English-language Travel Documents	100.0%	100.0%	91.7%	100.0%	83.3%	90.0%	100.0%	100.0%	100.0%	100.0%	83.3%	71.4%	33.3%	100.0%	100.0%	70.6%	33.3%	37.5%	50.0%	100.0%	0.0%	0.0%
PerfectClassif	**95.6%**	**80.1%**	**97.8%**	**82.6%**	**89.4%**	**74.9%**	**98.3%**	**78.1%**	**94.6%**	**78.1%**	**92.8%**	**67.6%**	**95.3%**	**56.2%**	**83.5%**	**76.6%**	**65.2%**	**64.8%**	**87.2%**	**71.4%**	**14.6%**	**33.9%**
PerfectTypes	95.6%	82.2%	97.8%	84.6%	89.4%	76.5%	98.3%	79.9%	94.6%	79.9%	92.8%	68.4%	95.3%	56.5%	83.5%	78.4%	65.2%	67.4%	87.2%	72.8%	14.6%	33.9%

Analysis

The refined prompt engineering solved the GPT4o underperformance issue. It did not materially impact the performance metrics of the other LLMs tested, with the exception of Mistral Large and Llam3-70b.

GPT4 and GPT3.5 performances are stable after the prompt modification and demonstrated excellent performance, with GPT4 slightly below GPT4o and GPT3.5 behind that.

Llama3-70b benefited from a 10 percent coverage boost and demonstrated stable accuracy, which makes it comparable to GPT3.5.

Command r+ and Command r coverage and accuracy performance is more consistent across fields following prompt tuning. Command r+ performance is comparable with GPT3.5 and Llama3-70b with Command r slightly behind.

Prompt engineering did generate some unexpected results associated with Mistral Large. While coverage was similar to that of GPT4 and GPT4o, accuracy would be considered disappointing. Additional investigation revealed that the underperformance can be attributed to insurance forms for which the model does not follow the expected naming. And while this indicates the documents could easily be reclassified correctly in post processing it is also surprising that the model is not able to output the correct name for these documents.

The F1 Score and Conclusions

	GPT4	GPT4o	GPT3.5	Claude3 Opus	Claude3 Sonnet	Claude3 Haiku
MetricName	F1	F1	F1	F1	F1	F1
Price 100k docs	€6,397	€3,227	€321	€12,519	€2,503	€208
FrenchDentalInvoice	92.9%	93.7%	91.8%	93.7%	93.3%	93.2%
JapaneseHomeQuote	82.9%	83.0%	79.2%	83.0%	82.2%	78.4%
EnglishFlightInvoice	83.8%	82.6%	69.9%	82.2%	77.8%	72.9%
Classif	87.1%	89.5%	81.5%	87.1%	85.5%	78.2%
All use cases aggreg	**86.7%**	**87.2%**	**80.6%**	**86.5%**	**84.7%**	**80.7%**

	Mistral Large	Command r+	Command r	Llama3-70b	Llama3-8b
MetricName	F1	F1	F1	F1	F1
Price 100k docs	€5,186	€2,493	€323	€2,443	€238
FrenchDentalInvoice	91.5%	91.4%	90.7%	90.9%	89.0%
JapaneseHomeQuote	81.6%	75.4%	67.1%	79.1%	72.1%
EnglishFlightInvoice	78.7%	75.5%	61.0%	75.0%	65.5%
Classif	70.7%	79.9%	65.0%	78.5%	20.4%
All use cases aggreg	**80.6%**	**80.6%**	**71.0%**	**80.9%**	**61.8%**

Based on our testing we can offer the following analysis and conclusions.

Relating to information extraction tasks with what could be considered simple fields it is clear that GPT4o, GPT4, and Claude3 (Opus, Sonnet, Haiku) are the best performing models and fall within the same performance range on all fields.

GPT3.5, Mistral Large and Llama3-70b are comparable to the best performing models with the primary difference being inconsistency in performance between fields. We also found that Llama3-8b, Command r+ and Command r are overall comparable to GPT3.5 but also underperforms on some specific fields.

For those tasks associated with complex fields we see that GPT4 and Claude3 Opus are the best models, with GPT4 slightly outperforming Claude3 Opus.

Claude3 (Sonnet and Haiku), Mistral Large and Llama3-70b lag slightly behind the leaders with each model slightly outperforming the others depending on the field being analyzed.

For what are identified as classification tasks GPT4o, GPT4 and Claude3 (Opus and Sonnet) produced the best results with GPT4o slightly outperforming the others.

GPT3.5, Llama3-70b, Command r+ and Claude3 Haiku show some slightly diminished performance when compared with the leaders, while Llama3-8b and Command r clearly underperform.

Model	Input/1M tokens	Output/1M tokens	100k documents	Context size
mistral-large	€7.41	€22.22	€ 5,186.00	32k
llama3-70b-instruct	€3.49	€10.47	€ 2,443.00	8k
llama3-8b-instruct	€0.34	€1.02	€ 238.00	8k
gpt4o	€4.61	€13.83	€ 3,227.00	128k
gpt-4-0125-preview	€9.14	€27.41	€ 6,397.00	128k
gpt-3.5-turbo-0125	€0.46	€1.37	€ 321.00	16k
command r+	€2.77	€13.85	€ 2,493.00	128k
command r	€0.46	€1.39	€ 323.00	128k
claude3-sonnet	€2.78	€13.91	€ 2,503.00	200k
claude3-opus	€13.91	€69.55	€ 12,519.00	200k
claude3-haiku	€0.23	€1.16	€ 208.00	200k

However, when evaluating LLMs it is critically important to not only evaluate overall performance, but also performance related to cost. As we see in the above cost comparison chart as well as the F1 table, the highest performing models typically boast the highest costs. However, depending on the use case, it may be permissible to sacrifice some performance for economy. For example, Claude3 Opus is highly performant, but may also be prohibitively expensive. GPT4o and Claude3 Sonnet deliver excellent performance at a slightly more affordable price point while our analysis shows that GPT3.5-turbo and Claude3 Haiku delivers an admirable combination of price and performance.

The realm of GenAI and LLMs is rapidly evolving. This report is intended to provide an unbiased evaluation to help readers make the best decision possible about how to make this technology a part of their technology strategy.