SHIFT

The State of AI in Insurance

Large Language Models for Insurance — How Do They Compare?

www.shift-technology.com

Executive summary

- **Performance comparison** of six different Large Language Models (LLMs) applied to common insurance industry processes
- **LLMs featuring a larger context size** - the maximum number of tokens the model can remember when generating text - generally perform better, although there are exceptions
- **Larger context size** comes at a cost premium, but is **necessary to achieve desired performance** for certain use cases
- **Effective prompt engineering is key** to obtaining the best possible performance from LLMs
- Performance metrics for LLMs are **unique to the use case** and must be evaluated carefully to ensure business requirements are being met
- The choice of which LLM to use should be based on a combination of **use case, acceptable performance** and **cost**

From the editor

Generative Artificial Intelligence (Gen AI) applications and the underlying Large Language Models (LLMs) that support this technology have captured the attention of the insurance industry. This technology has the potential to significantly increase the efficiency and accuracy of underwriting, claims and fraud, and risk processes.

Yet, for as much interest as there is in Generative AI, there is also uncertainty and unanswered questions. Insurers face an unprecedented amount of information about where Generative AI can benefit, where it may fall flat, and which models may be best for a variety of use cases. These are just some of the issues insurers must evaluate when considering how to bring Generative AI into their technology stack and business processes.

Shift Technology has been a pioneer in AI for insurance since 2014. Over the past decade we have built one of the industry's largest data science teams dedicated to AI in insurance. This team is engaged in research and development to advance the state of AI for insurance use cases, as well as the application of that R&D to develop innovative solutions for our insurance customers.

This report is the first in a series that will periodically highlight the findings from research our data scientists have undertaken to better understand the performance of specific LLMs when applied to common insurance processes. The goal is to provide insurance professionals with a trusted source of information pertaining to AI to help them make the best decisions possible when evaluating this technology.

Thank you to the Shift data scientists and researchers who made this report possible.

LLM model comparison: Data extraction and document classification

Methodology

The data science and research teams devised four test scenarios to evaluate the performance of six different publicly available LLMs: GPT3.5, GPT4, Mistral Large, Llama2-70B, Llama2-13B, and Llama2-7B.

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 developed by the Shift data science team. For each individual scenario, the team engineered a single prompt that was utilized by all six of the tested LLMs.

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.

Results & analysis

English-language airline invoices

85 anonymized English-language airline invoices were used in this scenario.

The extraction prompt sought the following results:

- Provider Name
- Start Date
- End Date
- Document Date
- Booking Number
- Flight Number (for all associated flights)
- Last Four Credit Card Digits
- Currency
- Base Fare for all Passengers
- Taxes and Fees for all Passengers
- Additional Fees for all Passengers
- Payments - this is a complex field consisting of the following: Payment Date, Amount & Status
- Travellers - this is a complex field consisting of the following: Traveller Name, Basic Fare, Total Taxes & Total Amount

	GPT4		GPT3.5		Mistral Large		Llama2-70b		Llama2-13b		Llama2-7b	
MetricName	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc
ProviderName	98.5%	67.1%	98.5%	59.5%	100.0%	63.8%	43.3%	52.4%	53.7%	33.3%	6.0%	100.0%
StartDate	98.3%	83.1%	98.3%	76.1%	100.0%	79.7%	37.9%	47.6%	50.0%	44.4%	5.2%	50.0%
EndDate	100.0%	82.7%	97.9%	66.2%	100.0%	79.6%	36.2%	33.3%	42.6%	29.6%	4.3%	50.0%
DocumentDate	95.3%	80.3%	96.9%	67.5%	100.0%	70.5%	40.6%	45.2%	56.3%	44.4%	4.7%	50.0%
BookingNumber	96.7%	71.8%	98.3%	3.8%	98.3%	55.1%	36.7%	21.4%	50.0%	1.9%	5.0%	0.0%
FlightNumbers	98.5%	65.5%	98.5%	50.0%	100.0%	61.2%	41.8%	35.7%	53.7%	33.3%	6.0%	75.0%
CreditCard4LastDigits	98.0%	94.2%	98.0%	90.7%	100.0%	94.3%	41.2%	60.0%	54.9%	50.0%	3.9%	25.0%
Currency	98.3%	96.7%	98.3%	93.7%	100.0%	75.0%	38.3%	54.8%	55.0%	59.3%	3.3%	50.0%
BasicFareAllPassengers	97.0%	51.7%	100.0%	33.3%	100.0%	57.4%	39.4%	30.6%	57.6%	20.4%	0.0%	0.0%
TaxesAndFeesAllPassengers	96.9%	44.4%	100.0%	23.2%	96.9%	42.6%	34.4%	16.7%	50.0%	5.8%	0.0%	0.0%
AdditionalFeesAllPassengers	91.7%	28.0%	75.0%	12.5%	91.7%	13.8%	25.0%	11.8%	25.0%	10.5%	0.0%	0.0%
AdditionalFeeInsurance	100.0%	86.7%	69.2%	75.0%	100.0%	92.9%	38.5%	22.7%	53.8%	17.1%	0.0%	0.0%
TotalAmount	93.2%	91.4%	91.5%	89.7%	96.6%	91.5%	35.6%	52.8%	50.8%	42.0%	3.4%	25.0%
TotalPaidAmount	92.5%	90.6%	69.8%	81.4%	90.6%	84.6%	37.7%	42.9%	54.7%	33.3%	3.8%	25.0%
PaymentDate	95.7%	38.2%	82.6%	34.7%	73.9%	32.0%	21.7%	9.8%	43.5%	4.4%	4.3%	0.0%
PaymentStatus	89.3%	76.6%	58.9%	66.0%	73.2%	70.7%	26.8%	36.6%	42.9%	14.8%	3.6%	25.0%
PaymentAmount	88.7%	96.9%	60.6%	86.0%	74.6%	91.4%	29.6%	46.3%	42.3%	17.0%	2.8%	25.0%
TravellerBasicFare	82.8%	75.4%	50.0%	32.4%	70.7%	60.3%	17.2%	24.4%	25.9%	12.3%	3.4%	28.6%
TravellerTotalTaxes	83.0%	60.7%	53.2%	36.8%	68.1%	43.3%	19.1%	21.6%	31.9%	14.2%	0.0%	0.0%
TravellerTotalAmount	83.1%	80.3%	49.2%	41.2%	71.2%	54.2%	15.3%	19.5%	22.0%	11.3%	0.0%	0.0%

Analysis

GPT4, GPT3.5 and Mistral Large proved most adept in this scenario. The Llama models proved to be significantly behind, especially when it comes to Coverage. We may be experiencing a situation where the Llama models simply have a harder time finding the relevant information or formatting the output. The results may also be influenced by Llama's established context size of only 4k, which is smaller than any of the other models tested. In this situation, any document that was larger than the context size would simply not be processed and the model would not return any result, thus impacting its Coverage score.

GPT4 and Mistral Large performed well when dealing with complex fields. These LLMs can not only extract nested information but also output the result in a usable format.

While adequate, performance related to list fields may have been negatively affected by the complexity associated with these extractions.

In the case of Payment Date, we did witness lower accuracy, which can be attributed to the models' tendency to substitute the document date for payment date if the payment date is unavailable.

Japanese-language property repair quotes

In this scenario we applied each of the test LLMs against 100 anonymized Japanese-language property repair quotes. Documents represented quotes from multiple different providers in non-standard formats. These documents would not be considered templated.

The extraction prompt sought the following results:

- Provider Name
- Provider Address
- Post Code
- Provider email
- Tax Amount
- Total Amount with Tax
- Discount Amount

	GPT4		GPT3.5		Mistral Large		Llama2-70b		Llama2-13b		Llama2-7b	
MetricName	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc
providerName	98.9%	73.1%	98.9%	68.1%	98.9%	72.3%	78.3%	68.8%	66.3%	66.3%	23.9%	23.9%
providerAddress	91.2%	70.2%	93.4%	62.8%	90.1%	69.9%	74.7%	55.7%	64.8%	64.8%	20.9%	20.9%
postCode	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	32.9%	96.4%	51.2%	51.2%	12.2%	12.2%
providerEmail	100.0%	63.6%	100.0%	63.6%	100.0%	63.6%	87.5%	42.9%	75.0%	75.0%	0.0%	0.0%
taxAmount	100.0%	86.0%	96.4%	77.5%	100.0%	85.7%	78.3%	74.0%	65.1%	65.1%	24.1%	24.1%
totalAmountWithTax	100.0%	97.0%	97.9%	95.9%	99.0%	92.9%	78.4%	90.9%	67.0%	67.0%	25.8%	25.8%
discountAmount	100.0%	9.6%	73.3%	10.0%	93.3%	4.9%	73.3%	0.0%	53.3%	53.3%	33.3%	33.3%

Analysis

Overall, GPT4, GPT3, and Mistral Large performed best in both Coverage and Accuracy, with some exceptions. While Llama70 and Llama13 showed only slightly worse in Accuracy, Coverage is clearly lacking. This may be due to similar characteristics identified for underperformance in the previously described airline invoices scenario.

French-language dental invoices

In this scenario, each LLM was applied against a dataset of 119 French-language dental invoices. 79 of the invoices are considered to have a strong layout, meaning they could be described as templated documents. The remaining 60 were selected at random to mimic what may be experienced in an insurer's data.

The extraction prompt sought the following results:

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

	GPT4		GPT3.5		Mistral Large		Llama2-70b		Llama2-13b		Llama2-7b	
MetricName	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc
DocumentDate	100.0%	95.7%	99.3%	93.5%	100.0%	96.4%	98.5%	89.8%	89.7%	87.8%	90.4%	82.5%
ProviderName	100.0%	95.0%	100.0%	93.5%	100.0%	94.2%	98.5%	93.4%	89.1%	90.3%	90.5%	88.9%
RawProviderFiness	65.1%	61.9%	58.1%	53.7%	69.8%	64.4%	67.4%	70.0%	53.5%	69.7%	60.5%	56.5%
ProviderRpps	97.1%	92.5%	92.3%	94.9%	91.3%	92.9%	95.2%	93.2%	80.8%	94.2%	81.7%	94.3%
ProviderPostCode	99.3%	99.3%	99.3%	99.3%	100.0%	97.8%	90.5%	96.8%	85.4%	100.0%	81.0%	98.2%
TotalIncurredAmount	100.0%	97.8%	100.0%	96.4%	100.0%	98.6%	98.5%	97.1%	86.8%	86.8%	89.0%	88.6%
PaidAmount	100.0%	69.2%	95.6%	74.4%	98.5%	55.8%	95.6%	44.2%	79.4%	32.7%	55.9%	26.1%

Analysis

For this scenario, GPT4, GPT3.5 and Mistral Large performed well in both Coverage and Accuracy. Of the remaining models, Llama70 performed well, but not at the same levels of the best performers.

We did note that for both Coverage and Accuracy the Provider FINESS identifier underperformed across the board with all models. This may be attributed to a unique feature of French health invoices. The Provider FINESS identifier is not always clearly indicated or may be easily confused with other provider identifiers such as SIRET (Système d'identification du répertoire des établissements). This could impact the models' ability to accurately identify what should be extracted as well as the content ultimately extracted.

The witnessed underperformance could also be the result of general confusion. The field is confusing for the LLM because it is actually confusing in and of itself, even for a human. What this means is that the labels we use to evaluate the LLM may not be as accurate as the labels for the other fields. While additional prompt engineering could potentially help improve performance, if the ground truth itself is inherently unreliable, it would be hard to improve the performance. This demonstrates the importance of establishing good quality labels when evaluating LLMs.

English-language documents for travel claims

This dataset consisted of 405 anonymized English-language documents provided to support travel insurance claims.

The extraction prompt sought the following results:

- A classification for each page
- A group of pages related to the same document

The expected output would be a list of segmented documents including the document type and its span of pages (indicating start and end page).

In addition to metrics for individual document type, we also compute an aggregated performance at the file level, as defined below as PerfectClassif and PerfectTypes. We consider the outputs for the models correct when all the segmented documents in a file (PerfectCalssif) are correct or when all the document types in a file are correct (PerfectTypes).

	GPT4		GPT3.5		Mistral Large		Llama2-70b		Llama2-13b		Llama2-7b	
MetricName	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov	Acc
Receipt - Airplane	91.6%	82.1%	77.1%	77.1%	75.9%	90.9%	1.2%	0.0%	2.4%	66.7%	2.4%	40.0%
Receipt - Hotel / Rental reservations	90.0%	66.7%	70.0%	54.5%	65.0%	80.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Receipt - Activities reservations	50.0%	33.3%	33.3%	40.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Receipt - Cruises	95.8%	67.9%	70.8%	68.2%	75.0%	69.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Receipt - Train	100.0%	100.0%	100.0%	66.7%	66.7%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Bank / Credit card statement	96.9%	85.3%	87.5%	82.8%	84.4%	76.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Cancellation policy	33.3%	16.7%	50.0%	37.5%	16.7%	33.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Cancellation proof	83.3%	71.4%	58.3%	58.8%	54.2%	70.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Medical bills	88.9%	60.0%	77.8%	38.5%	77.8%	55.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Medical report	95.0%	89.2%	79.2%	94.8%	69.2%	86.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Proof of payment	41.7%	25.0%	50.0%	50.0%	33.3%	66.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Claim form 1 - Front page	100.0%	100.0%	100.0%	28.6%	66.7%	22.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Claim form 1 - Claimant information / Agency details	100.0%	100.0%	25.0%	50.0%	50.0%	16.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Claim form 1 - Details of loss / Incident description	100.0%	100.0%	50.0%	66.7%	50.0%	14.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Claim form 1 - Claimed expenses / Payment	100.0%	100.0%	50.0%	50.0%	50.0%	16.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Claim form 1 - Required documentation	100.0%	80.0%	75.0%	50.0%	50.0%	14.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Claim form 1 - Authorization and assignment	100.0%	50.0%	75.0%	40.0%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Claim form 2 - General information	100.0%	50.0%	0.0%	0.0%	100.0%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Claim form 2 - Details of trip cancellation / trip interruption / trip delay	100.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Claim form 2 - Claimed expenses and authorization	100.0%	100.0%	0.0%	0.0%	100.0%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Claim form 3 - Front page and summary information	100.0%	100.0%	60.0%	66.7%	80.0%	57.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Claim form 3 - Disclosures	100.0%	100.0%	50.0%	100.0%	100.0%	40.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Claim form 3 - Assignment and authorization	42.9%	66.7%	14.3%	100.0%	42.9%	33.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Physician statement form 1 - Insured and physician information	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Physician statement form 1 - Patient's diagnosis	100.0%	33.3%	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	87.5%	93.3%	50.0%	88.9%	18.8%	50.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Other	42.1%	54.5%	47.4%	38.8%	45.6%	62.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
UsPhysicianStatementFormWholeDoc	100.0%	100.0%	41.7%	80.0%	8.3%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
PerfectClassif	94.6%	80.7%	90.9%	71.7%	84.7%	68.8%	0.2%	0.0%	0.7%	66.7%	1.5%	33.3%
PerfectTypes	94.6%	82.5%	90.9%	72.8%	84.7%	70.3%	0.2%	0.0%	0.7%	66.7%	1.5%	33.3%

Analysis (cont'd)

GPT4 is clearly ahead of all other LLMs tested when it comes to Coverage and Accuracy for document types individually as well as for our aggregated categories. The Llama models did not fare well in this particular test which may be related to the context size (4k) associated with these models or because they were unable to deliver results in the output format mandated in the prompt.

We believe what could be considered underperformance for certain documents types (e.g. Receipt - Activities Reservations, Cancellation Policy, Proof of Payment) may be due to either slightly vague document type descriptions included in the prompt or document types that closely resemble others. Additional prompt engineering could resolve some of the witnessed underperformance.

Cost comparison

For the LLMs we tested, there are essentially two price ranges. GPT3.5 and the Llama models are relatively inexpensive, while GPT4 and Mistral Large cost more to use. Perhaps not surprisingly, our analysis shows that overall, the more expensive LLMs performed better. However, it is interesting to see that GPT3.5, while less expensive, delivers performance levels closer to that of the expensive models. We may surmise that when analyzing cost to performance for GPT3.5 when compared to the Llama models the key to this performance discrepancy lies in context size. We see that although the cost of GPT3.5 is similar to that of the Llama models, it features a context size four times that of any of the Llama models, providing a measurable performance edge.

Model	Input/1M tokens	Output/1M tokens	100k documents	Context size
llama-2-7b-chat	€0.63	€0.49	€ 301.00	4k
llama-2-13b-chat	€0.89	€0.77	€ 433.00	4k
llama-2-70b-chat	€1.67	€1.46	€ 814.00	4k
gpt-3.5-turbo-0125	€0.46	€1.37	€ 321.00	16k
gpt-4-0125-preview	€9.14	€27.41	€ 6,397.00	128k
mistral-large	€7.41	€22.22	€ 5,186.00	32k

Conclusion

In the world of Generative AI and LLMs, it is important to remember that one size does not fit all. You must first determine the job that needs to be done and apply the right tool to accomplish that. Evaluating performance compared to cost is also important. As we have seen GPT4 and Mistral Large consistently outperform the other LLMs in this comparison, with GPT4 performing exceptionally well for classification tasks.

The performance witnessed by GPT3.5 is close behind that of the leading LLMs, and may perform well enough for many use cases. This is especially true when its price point is taken into consideration.

The Llama models we tested, specifically when compared to the pricing and performance for GPT3.5, were simply not competitive, especially so in the classification scenario.

Our data science team is consistently testing LLMs and we will continue to report on their results in future editions of this report.