SHIFT

The State of AI in Insurance

Claims Decisioning and Liability Determination in Subrogation

www.shift-technology.com

Executive summary

- The introduction and continued advancement of reasoning LLMs **creates new opportunities to apply Gen AI to important insurance use cases** such as subrogation liability assessment and claims decisions
- The wide variety of LLMs available **make assessment of pros and cons even more important** when determining which LLM to use for a specific use case
- In certain situations "standard" models **achieve comparable performance to "reasoning" models** on reasoning tasks

From the editor

Since Shift began publishing this report more than a year ago, the use of generative artificial intelligence (Gen AI) to drive efficiency, accuracy, and fairness in the claims process has become increasingly mainstream. And like most technologies, the large language models (LLMs) powering this important insurance transformation have continued to evolve. From its beginnings, this report was designed to provide insight into the intersection between LLMs and specific insurance use cases, and help provide some clarity around how specific LLMs performed when applied against specific tasks.

With the latest edition of the State of AI in Insurance Report we tested a total of 21 LLMs. As with subsequent reports, in an effort to best represent the current state-of-the-art as well as highlight those LLMs most likely to be in use in insurance environments we both retire older, and include newer, models to create an optimal testing environment. For this report we have added 10 new LLMS to the benchmark:

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).

LLMs New to This Report

- **GPT4.5:** the short-lived OpenAI flagship standard model to be retired soon
- **GPT4.1, GPT4.1-mini, and GPT4.1-nano:** the new suite of OpenAI's standard models
- **o4-mini:** the latest OpenAI reasoning model
- **Deepseek V3:** the latest version of Deepseek's standard model
- **MAI-DS-R1:** Microsoft's version of Deepseek R1, the Deepseek reasoning model
- **Claude3.7 Sonnet:** an updated version of Anthropic's flagship model
- **Mistral Small 2503:** the latest version of Mistral's small model
- **Llama4-maverick:** the latest available Llama model

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

Methodology

The data science and research teams devised six test scenarios to evaluate the performance of 21* different publicly available LLMs. Four of the scenarios are classified as extraction and classification tasks. Two scenarios, Claims Decisions and Motor Liability, are classified as reasoning tasks and are new to this report.

- 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)
- Claims Decisions (reasoning)3
- Motor Liability (reasoning)

Defining the Claims Decisions scenario — can the model make the following decisions:

- Is the claim declaration within the policy's effective date?
- Is the invoice consistent with the claim?
- Is there any reason to deny the claim based on the claim description?
- Is there any information missing that would be required to make a coverage or reimbursement decision?
- Claim status determination
- If the claim is covered, how much should be reimbursed in total?

Defining Motor Liability — can the model ascertain, based on the information contained in a claim, liability required for a subrogation determination.

*See table for all tested models

Definitions:

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.
3. Tasks requiring some type of decision to be made by the LLM

(Continued on next page)

(Continued)

The LLMs were tested for:

- **Coverage** - did the LLM 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.

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.

Results & analysis

LLM Metrics Comparison

F1 Score	GPT4.1 FL	GPT4.1	GPT4o	o1-preview	o4-mini	o3-mini	o1-mini	GPT4.1-mini	GPT4.1-nano	GPT4o-Mini	Deepseek V3	MAI-DS-R1	Deepseek R1
Price 100k docs	€2,816	€1,408	€1,840	€22,800	€1,672	€1,672	€1,672	€282	€70	€112	€0	€0	€0
French Dental	90.9%	91.0%	93.7%	93.2%	95.2%	94.7%	92.7%	91.0%	90.9%	90.0%	92.4%	94.9%	93.9%
Japanese Home	85.1%	85.6%	83.0%	84.9%	82.2%	82.2%	82.0%	81.9%	76.6%	78.4%	82.5%	81.2%	82.2%
English Flight	84.4%	85.4%	82.6%	82.0%	80.7%	78.3%	78.9%	81.0%	67.3%	75.3%	82.2%	79.1%	78.9%
Classification without ID	91.3%	91.7%	91.3%	91.0%	90.3%	89.1%	88.0%	88.7%	66.2%	85.8%	87.4%	89.2%	89.5%
Classification/ extraction aggregated	**87.9%**	**88.4%**	**87.6%**		**87.1%**	**86.1%**	**85.4%**	**85.7%**	**75.3%**	**82.4%**	**86.1%**	**86.1%**	**86.1%**
Claim Decisions	88.7%	83.7%	81.3%		96.8%	93.3%	92.0%	82.7%	57.5%	60.6%	80.0%	96.5%	96.5%
Motor Liability	80.6%	81.3%	80.9%		79.6%	78.7%	78.0%	76.0%	58.0%	71.6%	78.3%	80.1%	79.6%
Reasoning aggregated	**84.6%**	**82.5%**	**81.1%**		**88.2%**	**86.0%**	**85.0%**	**79.3%**	**57.7%**	**66.1%**	**79.2%**	**88.3%**	**88.0%**
All use cases aggregated	**87.0%**	**86.7%**	**85.8%**	**87.8%**	**87.4%**	**86.1%**	**85.3%**	**83.9%**	**70.3%**	**77.7%**	**84.1%**	**86.7%**	**86.7%**

F1 Score	Claude4 Opus	Claude4 Sonnet	Claude3.7 Sonnet	Claude3.5 Sonnet v2	Claude3.5 Haiku	Mistral Large 2411	Mistral Large 2407	Mistral Medium 2505	Mistral Small 2503	Llama4-maverick	Llama4-scout	Llama3.3-70b
Price 100k docs	€11,745	€2,503	€2,503	€2,503	€698	€2,604	€2,604	€313	€62	€247	€137	€344
French Dental	94.4%	93.9%	93.8%	94.0%	91.6%	94.1%	93.5%	93.3%	91.7%	91.7%	91.0%	90.9%
Japanese Home	84.5%	85.0%	83.3%	82.7%	83.7%	82.5%	82.5%	83.0%	81.9%	81.9%	82.3%	79.2%
English Flight	85.3%	82.6%	83.1%	79.7%	78.7%	82.1%	82.1%	81.4%	75.7%	81.4%	80.6%	77.2%
Classification without ID	91.1%	90.1%	88.9%	89.5%	87.0%	88.0%	88.0%	88.8%	85.7%	85.2%	85.7%	86.9%
Classification/ extraction aggregated	**88.8%**	**87.9%**	**87.3%**	**86.5%**	**85.3%**	**86.7%**	**86.5%**	**86.6%**	**83.8%**	**85.1%**	**84.9%**	**83.5%**
Claim Decisions	86.8%	81.0%	82.2%	83.4%	58.8%	83.1%	79.9%	78.6%	75.8%	69.1%	48.5%	70.2%
Motor Liability	79.0%	78.9%	79.5%	79.6%	78.5%	75.5%	77.5%	75.2%	76.6%	77.8%	76.2%	79.7%
Reasoning aggregated	**82.9%**	**79.9%**	**80.8%**	**81.5%**	**68.6%**	**79.3%**	**78.7%**	**76.9%**	**76.2%**	**73.5%**	**62.3%**	**75.0%**
All use cases aggregated	**87.1%**	**85.6%**	**85.4%**	**85.1%**	**80.5%**	**84.5%**	**84.3%**	**83.8%**	**81.6%**	**81.8%**	**78.5%**	**81.1%**

Results & analysis

Standard vs. reasoning models:
Although it may sound like we are simply stating the obvious, our testing continues to validate that reasoning models generally perform better than standard models at performing reasoning tasks. And in this benchmark we observed that MAI-DS-R1 and o4-mini performed the best against our scenarios.

While reasoning models will typically perform well across all use cases, they are generally more expensive and experience higher latency than standard models. This highlights the importance of choosing the correct model for a given use case, and taking all parameters (performance, price, latency) into account when making a decision about how and when to deploy LLMs.

Interestingly, we did find that some standard models including GPT4.1 and GPT4o slightly outperformed reasoning models on less complex use cases. However, that further supports the rationale that deciding which LLM to use and when must take all contributing factors into account.

And while GPT4.5, OpenAI's attempt at delivering a universal model that performs well across all use cases, did just that, it came at the cost of pricing and latency that made it impractical in production environments. That could explain its early retirement.

The OpenAI models:
For each version of the OpenAI o-mini models (the small option of their reasoning models) we recorded a slight performance increase over earlier models tested. OpenAI has shown the capability to improve performance with each generation while keeping pricing steady.

Regarding their new standard models suite, including GPT4.1, GPT4.1-min, and GPT4.1-nano, our testing revealed several interesting findings. GPT4.1 is a thoughtful evolution of GPT4o that features better performance at a slightly lower price. At the same time we found that GPT4.1-mini and nano are actually bookending GPT4o-mini in terms of both price and performance.

The Anthropic models:
In our testing, Claude3.7 Sonnet, managed to achieve a small performance increase compared to its previous version. In this benchmark we observed performance comparable to GPT4o on both extraction and reasoning tasks, yet still below OpenAI's new flagship model GPT4.1.

(Continued on next page)

(Continued)

The Mistral models:
Mistral released a new version of their small model, which is similar to GPT4.1-nano in price but gives much better performance. Our testing showed it performed better than GPT4o-mini and just below GPT4.1-mini—making its use quite compelling in the right circumstances.

The Deepseek models:
Microsoft released MAI-DS-R1, their version of a reasoning model based on Deepseek R1. As expected, the performance of both models are very similar but according to MS, their model gives more objective and neutral answers on some political or societal topics. However, this may or may not be relevant when considering specific insurance scenarios.

Deepseek V3, the flagship standard model, is just above GPT4.1-mini in terms of performance, continuing to demonstrate the viability of open source models.

The Meta models:
Meta released two new versions of its Llama model, Llama4-maverick (400B parameters) and Llama4-scout (109B parameters). We tested the former and observed performance similar to GPT4.1-mini on classification and extraction use cases, at a similar price to the OpenAI model. However, the model's performance was a bit disappointing on reasoning use cases.

Conclusion

Our research continues to show that in the world of insurance AI, one size does not fit all. With each round of testing and subsequent report, we observe that the requirements of specific use cases must be taken into careful consideration when determining which LLM to apply to each use case to achieve the desired result. As new LLMs are introduced and new use cases evolve, this level of evaluation will continue to be critically important.