SHIFT

The State of AI in Insurance

The Power of Fine-Tuning (Vol. IV)

www.shift-technology.com

Executive summary

- While prompt engineering is often thought of first, fine-tuning shows incredible promise in **boosting the performance of LLMs**
- Fine-tuning **can deliver "big model" performance** at "small model" prices despite incurring some additional costs
- Fine-tuning on complex data sets appears to **yield the best overall results**
- Fine-tuning is **just one strategy to achieve the best performance** in an LLM strategy

From the editor

As reported in **Vol. III of our comparison of LLMs**, performance gains across various available models and their variants are quickly reaching a point where current LLM performance differences are negligible depending on the use case. With performance equalizing, the cost differences of LLMs become even more critical to the decision making process.

At the same time, there are proven methods for improving LLM performance in production environments. For example, prompt engineering is a well documented approach to achieving incremental performance. The more precise the initial prompt, the more accurate the results.

Another approach generating interest is the concept of fine-tuning. In the case of LLMs, fine-tuning is defined as the process of further training a pre-trained model on smaller, more specific data sets to achieve greater performance. To better understand the kind of performance increases that are possible via fine-tuning, the Shift Data Science (DS) and Research Teams compared a fine-tuned GPT4o-Mini against a standard GPT4o and a standard GPT4o-Mini on an information extraction task.

Fine-Tuning LLM Models for Information Extraction: Select Insurance Documents

Methodology

The Shift Technology data science and research teams used three test scenarios to evaluate the performance of the three models: GPT4o, GPT4o-Mini, and GPT4o-Mini fine-tuned.

The scenarios include:

- Information extraction from English-language airline invoices (complex - e.g. fields that represent complex objects or whose value is a list of objects)
- Information extraction from Japanese-language property repair quotes (simple - e.g. basic extraction tasks)
- Information extraction from French-language dental invoices (simple)

Fine-tuning was accomplished via Azure OpenAI which required uploading instruction-formatted messages in the form of a json file. When fine-tuning, the available hyperparameters (batch size, learning rate multiplier, and number of epochs) are determined automatically or can be set manually. For the purposes of this report the Shift Data Science and Research teams used automatic settings.

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?

Performance is reported as an F1 score. 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).

GPT4o-Mini was fine-tuned against training sets derived from three different data sets:

CORD (Consolidated Receipt Dataset for Post-OCR Parsing) — 1,000 publicly available English-language receipts collected from shops and restaurants used to create a training set. The dataset contains the OCR output for each receipt and labels for given fields from the receipts.

(Continued on next page)

(Continued)

English-language Flight Invoice dataset — a dataset of 550 English-language flight invoices labeled using GPT4o-Mini. These invoices were selected from a production environment according to their classification and anonymized to ensure privacy standards were met.

Combined CORD and English-language Flight Invoice dataset — a combination of the two previously described datasets.

Name	Training Set	Validation Set
500	500	100
EnFlight-470	470	80
Cord-EnFlight-1000	1000	100

Reading the Results

Evaluating LLM performance is based on the specific use case and the relative performance achieved. The table included in this report reflects that reality and is 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.

Results & analysis

LLM Metrics Comparison

F1 Score	GPT4o	GPT4o-En-Flight470	GPT4o-Mini	o1-preview	o1-mini	GPT4o-Mini Cord-EnFlight1000	GPT4o-Mini En-Flight470	GP-T4o-Mini Cord500	Claude3.5 Sonnet	Claude3 Opus
Price 100k docs	€1,840	€2,685	€112	€12,276	€2,455	€214	€214	€214	€2,503	€12519
FrenchDentalInvoice	93.7%	93.7%	90.0%	93.2%	92.7%	91.0%	90.8%	86.9%	94.3%	93.1%
JapaneseHomeQuote	83.0%	83.5%	78.4%	84.9%	82.0%	79.4%	83.1%	78.4%	83.0%	83.3%
EnglishFlightInvoice	82.6%	78.4%	75.3%	82.0%	78.9%	79.1%	81.0%	74.5%	81.5%	82.2%
ClassifWithoutId	91.3%		85.8%	91.0%	88.0%	38.7%	30.9%		88.1%	83.7%
All use cases aggreg	**87.6%**		**82.4%**	**87.8%**	**85.4%**				**86.7%**	**87.2%**

F1 Score	Claude3 Sonnet	Claude3 Haiku	Mistral Large 2407	Mistral Large	Mistral Nemo	Minis-tral-3b	Lla-ma3.1-405b	Lla-ma3.1-70b	Llama3.1-8b
Price 100k docs	€2,503	€208	€2,604	€5,186	€75	€20	€3,471	€1,326	€168
FrenchDentalInvoice	93.3%	93.2%	93.5%	91.5%	91.1%	90.8%	93.1%	91.8%	90.4%
JapaneseHomeQuote	82.2%	78.4%	82.5%	81.6%	73.5%	72.8%	83.3%	79.8%	71.0%
EnglishFlightInvoice	77.8%	72.9%	82.1%	78.7%	69.1%	60.3%	83.5%	79.2%	65.0%
ClassifWithoutId	86.2%	79.3%	88.0%		80.0%	75.3%	88.2%	87.7%	70.4%
All use cases aggreg	**84.9%**	**80.9%**	**86.5%**		**78.4%**	**74.8%**	**87.0%**	**84.6%**	**74.2%**

Results Summary

	Fine-tuned on simple dataset	Fine-tuned on complex dataset	Fine-tuned on a mixed dataset
Same use case / Simple dataset	• Outperforms GPT-4o-Mini • Can almost compare to GPT-4o (on some datasets)	• Outperforms GPT-4o-Mini • Can compare to GPT-4o (on some datasets)	• Slightly outperforms GPT-4o-Mini • Underperforms GPT-4o • Does not bring value compared to mono dataset fine-tuning
Same use case / Complex dataset	• Slightly underperforms GPT-4o-Mini • Underperforms GPT-4o	• Outperforms GPT-4o-Mini • Can compare to GPT-4o	• Outperforms GPT-4o-Mini • Underperforms GPT-4o • Does not bring value compared to mono dataset fine-tuning
Other use case	• Delivers suboptimal performance		

Conclusion

Our research and testing demonstrates that in certain situations fine-tuning can be a cost-effective means of driving greater performance from an LLM. And while our testing at this time was limited to the GPT4o and GPT4o-Mini models we believe this may hold true for other available models as well. However, additional research will be required to ascertain if we are correct in that assumption.

In our testing we specifically demonstrated that GPT4o-Mini, fine-tuned on a complex dataset for a given use case, will deliver better performance than the base model and is even comparable to GPT4o. This means that for any given use case, GPT4o-Mini fine-tuned on a more complex dataset then used as a new base model would almost certainly deliver a performance gain. When taking cost into account, this becomes a compelling proposition.

However, we should remember that fine-tuning is not a "one size fits all" proposition. Our findings indicate that when a model is fine-tuned on a particular use case, information extraction for example in the case of this testing, the fine-tuning may cause performance degradation when applied to a different use case. We also observed that our testing seems to reveal that mixing datasets for fine-tuning does not necessarily result in better performance.

In further understanding the results, one must also take into account the testing environment. Here we labeled the English-language flight invoices data set using GPT4o versus manual labeling and correcting, resulting in some errors. This dataset is also known to contain a significant amount of noise. These factors likely contributed to the results showing that fine-tuning for GPT4o delivered no clear performance gains and in certain situations caused performance degradation. Since this was a relatively small study using limited available models, both these situations will need further investigation to make any lasting determinations.

As to the cost consideration of fine-tuning, the price of a fine-tuned model is double that of the base GPT4o-Mini. Given this, it is still 1.5x less expensive than the previous GPT-3.5 model and 10x less expensive than the base GPT4o. For many uses cases a fine-tuned GPT4o-Mini will meet or exceed price/performance requirements.

The world of LLMs is rapidly evolving and their impact on the world of insurance is already clear. The ability to generate measurable performance gains through fine-tuning creates even more opportunity to bring the power of GenAI to the challenges facing the global insurance industry.