SHIFT

保険 AI の現状

保険向け LLM の性能比較レポート（第 5 弾）

www.shift-technology.com/ja

エグゼクティブサマリー

- **開発業者は引き続き LLM モデルを投入し続けています。**全く新しい LLM モデルも、既存の LLM モデルファミリーの拡張版も、コストやパフォーマンス、さまざまなユースケースへの適合性など新たな疑問をもたらしています。
- LLM の性能を比較する場合、**「最良」という言葉は相対的なものであり、**個々のユースケースに密接に関係していています。
- LLM の状況が多様化するにつれ、**LLM の意図する目的を理解すること**が重要な評価基準になります。
- どの LLM が各ユースケースに適しているかを評価する上で、**価格性能比は引き続き**重要な指標になります。
- オープンソースコミュニティから生まれた大規模モデルである **Deepseek R1 の実効性を検証しました。**

編集者からのメッセージ

LLM は追いつくのが困難なほど急速なスピードで進化を続けています。確立されたモデルが新バージョンを発表し、新たなプレーヤーも参入しています。つまり、このような変化が、重要な保険プロセスやユースケースをサポートする LLM の使用方法にどのような影響を与えるかを理解することが非常に重要になってきています。

本レポートは、保険に特化した様々なユースケースに 6 つの異なる大規模言語モデル（Large Language Models：LLM）を適用した際のパフォーマンスをまとめた「**保険 AI の現状（The State of AI in Insurance）**」レポートから始まりました。最初のレポート発行以来、テストされたモデルのいくつかはテスト対象から削除され、新しいモデルが追加されました。これは、本レポートが利用可能な LLM の現在の最先端を最もよく反映するとともに、技術コミュニティから大きな関心を集めているモデル（例えば Deepseek R1 など）にも注目し、保険に特化したユースケースへの展開が最も考慮される可能性の高いモデルも含まれています。本レポートは、事前に設定された一連のタスクに対する相対的なパフォーマンスを比較するだけでなく、テストされた各 LLM に関連するコスト／パフォーマンスの比較結果も示すことを目的としています。

第 5 弾では、Deepseek R1 を含む 11 個の新しい LLM、合計 19 の LLM をテストしています。性能の報告には、各モデルで生成された F1 スコアを引き続き使用しています。F1 スコアは、特定のユースケース（フランス語の歯科請求書など）と、ユースケースに関連する個々の フィールドという 2 つの軸に対するカバレッジと精度を集約したものです。このアプローチにより、ユースケースごとの単一のパフォーマンス指標と、10 万件の文書の分析に関連するコストを含む集計された総合スコアを生成することができます。F1 スコアの生成には次の式を使用しています：2 x Cov x Acc / (Cov + Acc)。

情報抽出と分類のための LLM モデル比較 及び 保険文書の選択

方法論

データサイエンスチームと研究チームは、4 つのテストシナリオを考案し、19 種類の一般公開されている LLM の性能を評価しました：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)、Microsoft Phi4*

実施したシナリオ：

- 英語の航空会社請求書からの情報抽出（複雑）[1]
- 日本語の住宅修理見積書からの情報抽出（単純）[2]
- フランス語の歯科請求書からの情報抽出 (単純)
- 旅行保険請求に関連する英語文書の分類（複雑）

LLM に実施したテスト内容：

カバレッジ ― グラウンドトゥルース（モデルに何かを予測させるときに期待する値）が、抽出すべきものがあることを示したときに、LLM が実際にデータを抽出したか。

正確性 ― LLM が何かを抽出した際に正しい情報を提示したか。

すべてのシナリオのプロンプトエンジニアリングは、Shift のデータサイエンスチームとリサーチチームによって行われました。各シナリオでは、単一のプロンプトが設計され、テストされたすべての LLM で使用されました。すべてのプロンプトは GPT LLM 用にチューニングされており、場合によっては測定されたパフォーマンスに影響を与える可能性があることに留意することが重要です。

結果の読み方

LLM の性能評価は、特定のユースケースと達成された相対的な性能に基づいて行われます。本レポートに含まれる表は、この現実を反映し、ユースケースに適用された LLM の相対的性能に基づいて色分けされています。青の濃淡は、相対的性能の最高レベルを表し、赤の濃淡は、ユースケースに対する相対的性能が劣ることを表し、白の濃淡は、相対的性能が平均的であることを表しています。

そのため、90% という性能評価は、90% が特定のユースケースに関連する範囲の最低性能評価である場合、赤でコード化されることがあります。また、90%のパフォーマンスは、ユースケースを考慮すれば許容できるかもしれませんが、他の LLM が定義されたタスクをどのように実行したかに比べれば、依然として劣っていると評価されます。

コストについて

本ベンチマークレポートの第 1 弾から、10 万件の文書の処理に関わるコスト見積もりは、500 トークンの出力を前提にしています。しかし、この仮定は、現在テストに含まれている推論モデルには当てはまりません。定義上、これらのモデルは、専用の追加推論トークンを出力します。そのため、推論モデルの出力に 1,500 トークンを仮定してコスト計算を更新しました。

---

*= 今回のレポートでの新たな点

[1] 複雑なオブジェクトであるリストやフィールドから、複数のステップや情報抽出を含むタスク

[2] テキストフィールド、金額、日付などからの情報抽出

結果と分析

LLM 指標比較

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
10 万件の文書の価格	€1,840	€112	€22,800	$1,672	€1,672	€0	€2,503	€698	€2,503	€12,519	€2,503
フランス語の歯科請求書	93.7%	90.0%	93.2%	94.7%	92.7%	93.9%	94.0%	91.6%	94.3%	93.7%	93.3%
日本語の住宅修理見積書	83.0%	78.4%	84.9%	82.2%	82.0%	82.2%	82.7%	83.7%	83.0%	83.0%	82.2%
英語の航空会社請求書	82.6%	75.3%	82.0%	78.3%	78.9%	78.9%	79.7%	78.7%	81.5%	82.2%	77.8%
文書分類 ID 無	91.3%	85.8%	91.0%	89.1%	88.0%	89.5%	89.5%	87.0%	88.1%	88.0%	86.2%
すべてのユースケースの集約	**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
10 万件の文書の価格	€208	€2,604	€2,604	€344	€989	€179	€3,471	€1,326	€168	€95
フランス語の歯科請求書	93.2%	94.1%	93.5%	90.9%	92.1%	90.1%	93.1%	91.8%	90.4%	92.1%
日本語の住宅修理見積書	78.4%	82.5%	82.5%	79.2%	79.8%	71.1%	83.3%	79.8%	71.0%	81.0%
英語の航空会社請求書	72.9%	82.1%	82.1%	77.2%	78.5%	65.6%	83.5%	79.2%	65.0%	78.8%
文書分類 ID 無	79.3%	88.0%	88.0%	86.9%	84.3%	69.5%	88.2%	87.7%	70.4%	81.5%
すべてのユースケースの集約	**80.9%**	**86.7%**	**86.5%**	**83.5%**	**83.7%**	**74.1%**	**87.0%**	**84.6%**	**74.2%**	**83.3%**

最新の OpenAI モデルである o1-preview、o1-mini、o3-mini は優れたパフォーマンスを発揮し、o1-preview は F1 スコアの合計で最高を記録しました。しかし、それぞれのコスト（o1-preview は GPT4o の 10 倍）と計算時間（o1-preview は 45 〜 60 秒、o1-mini と o3-mini は平均 15 秒）により、本ベンチマークにあるようなユースケースで本番環境で使用される可能性は低いと考えられます。これらの最新の OpenAI モデルは、情報抽出や分類とは対照的に、複雑な推論タスクのために設計されているため、これは驚くべきことではありません。

テストした OpenAI のモデルと同様、Deepseek R1 は良好な結果を示しました。その結果は、OpenAI o3-mini と同等であり、OpenAI o1-preview の結果をわずかに下回りました。この評価でおそらく最も興味深いのは、Deepseek R1 はテストした OpenAI モデルよりも学習コストが大幅に低いということです。さらに、オープンソースの LLM である Deepseek R1 は、基本的に無料で使用することができます。Deepseek R1 は、効果的な大規模モデルがオープンソースの世界から生まれ、商用オプションに匹敵する性能を提供できることを示しています。

(次ページへ続く)

（続き）

同時に、テストでは 2 つの懸念事項が明らかになりました。待ち時間が長く、一部のレスポンスで 10 分に達することがありました。テストを繰り返したものの、この待ち時間がモデル自体に起因するものなのか、導入されたインフラに起因するものなのかを明確に判断することはできませんでした。説明したユースケースの場合、レイテンシーが発生するため、このモデルは本番環境には適さないでしょう。さらに、Deepseek R1 からの出力は、自動解析の前に後処理が必要な場合が多いことがわかりました。推論モデルであるこの LLM は、最終的なものとして読み取られる中間的な結果を頻繁に出力し、下流の構文解析に影響を与えました。

Mistral と Claude の新バージョン（それぞれ Mistral Large2411 と Claude3.5 Sonnet v2）は、以前のバージョンと同等の性能を発揮しました。我々のテストでは、パフォーマンスの大幅な向上は見られませんでした。興味深いことに、Claude3.5 Haiku は、前バージョンと比較して、3 倍の価格上昇に相関して +5% 近い大幅な性能向上を示しました。現在、Claude3.5 Haiku は、コスト競争力を維持しながら、「ビッグモデル」に近いパフォーマンスを示しています。

Llama3.2-11b の性能は期待外れでした。全体的なテストでは、性能とコストの点で以前の 3.1 バージョンと同等でした。Llama3.3-70b と Llama3.2-90b の 2 つの中型モデルを見てみると、これらのモデルは互いに、また以前のバージョンである Llama3.1-70b と性能の点で類似していることがわかります。Llama3.3-70b は Llama3.2-90b より 3 倍、Llama3.1-70b より 4 倍安価です。こうした Llama3.3-70b に関する性能と価格の改善により、性能と価格の両面で GPT4o-mini に匹敵するモデルとなっています。

Phi4 は GPT4o-mini に匹敵する性能を発揮し、価格も同程度であることが判りましたが、コンテキスト・ウィンドウが小さい（16k）ため、コンテキスト・ウィンドウを超える可能性があるユースケースでは魅力的な選択肢ではなくなるかもしれません。しかし、コンテキスト・ウィンドウのサイズが考慮されないユースケースでは、Phi4 は GPT4o-mini の合理的な代替品となり得ます。

結論

本レポート第 1 弾の制作を開始した際、“大型モデル”と“小型モデル”のパフォーマンスには明確な違いがあることを目の当たりにしました。そして、その後のレポートでは、性能がある程度均一化され始めたものの、依然として、最高の性能は最高のコストと関連していることが確認されました。

しかし、今回のテストでは、かつては大型モデル（非常に優れたパフォーマンス／高いコスト）と小型モデル（優れたパフォーマンス／妥当なコスト）の明確な区別があったものが、かなり明確ではなくなりつつあることがわかりました。推論モデル（Deepseek R1、OpenAI モデル（o1-preview、o1-mini、o3-mini））が登場したことで、さらに考慮すべき点が増えました。我々のユースケースでは非常に良いパフォーマンスを発揮しましたが、これらの LLM が複雑な推論タスクに最も適していることは明らかです。また、Deepseek R1 では、オープンソースコミュニティが大規模なモデル開発のための実行可能なルートになり得ることも分かりました。

Claude3.5 Haiku は、3-4 倍のコストで大型モデルに匹敵する性能を発揮します。しかし、Claude3.5 Haiku はまだほとんどの小型モデルより 6 倍高価であることに注意しなければなりません。それでも、我々のテストは、Claude3.5 Haiku が、優れたコスト管理を行いながら優れたパフォーマンスを必要とするユースケースにとって、小型モデルの良い代替品になり得ることを示しています。