SHIFT

保険AIの現状

保険向けLLMの性能比較レポート（第三弾）

www.shift-technology.com/ja

エグゼクティブサマリー

- **LLM技術は急速に進化しており、**既存モデルの新バージョンや全く新しいモデルが導入されています。
- 複数のモデルが非常に類似したパフォーマンスを達成する**接合点に達しています。**
- 新たなパフォーマンス向上が確立されるまで、LLM選定において**価格/パフォーマンスの比較**が重要な要因となる可能性があります。

編集者からのメッセージ

レポート第二弾でのF1スコアの導入により、特定の保険利用ケースにおける大規模言語モデル（LLM）のパフォーマンス評価に関する報告の仕方を少し異なる視点で考えることができるようになりました。この報告書には、その進化した考え方が反映されています。

私たちは、シナリオごとの集計F1スコアが、テスト対象のLLMが一般的な保険業界の利用ケースに対してどのようなパフォーマンスを発揮するか、またその展開に関連するコストがパフォーマンスと一致しているかを理解するために必要なインサイトを提供できると信じています。このアプローチにより、各モデルが「単純なシナリオ」（テキストフィールドからの情報抽出、金額、日付など）に対してどのようにパフォーマンスを発揮するか、または「複雑なシナリオ」（複数のステップを含むタスクや、複雑なオブジェクトからの情報抽出を含むタスク）に対してどのようなパフォーマンスを発揮するかを理解しやすくなります。

レポート第二弾以降に市場に導入されたLLM技術の進展に基づき、私たちのデータサイエンスおよび研究チームは、GPT4o-Mini、Claude3.5 Sonnet、Mistral Large 2407、Llama3.1-405b、Llama3.1-70b、Llama3.1-8bの6つの新しいLLMをテストに追加しました。Command rおよびCommand r+は評価から除外されました。

この報告書は、シフトのデータサイエンスおよび研究チームのサポートがなければ実現できませんでした。

情報抽出のためのLLMモデル比較 及び 保険文書の選択

方法論

シフトテクノロジーのデータサイエンスおよび研究チームは、16種類の公に利用可能なLLMのパフォーマンスを評価するために4つのテストシナリオを考案しました：GPT3.5、GPT4、GPT4o、GPT4o-Mini、Mistral Large、Mistral Large 2407、Llama3.1-8b、Llama3.1-70b、Llama3.1-405bおよびそれに対応するLlama3バリアント、Claude3 Opus、Claude3、3.5 Sonnet、Claude3 Haiku

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）
- 日本語の住宅修理見積書からの情報抽出（単純）
- 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に実施したテストの内容：

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される値）が示した時にLLMが実際にデータを抽出したか。

正確性 - LLMが何かを抽出したときに正しい情報を提示したか。

すべてのシナリオに対するプロンプトエンジニアリングは、シフトのデータサイエンスおよび研究チームによって行われました。各シナリオごとに、単一のプロンプトが設計され、すべてのテストされたLLMで使用されました。すべてのプロンプトはGPT LLM向けに調整されており、これが測定されたパフォーマンスに影響を与える場合があります。

表の読み方

LLMのパフォーマンス評価は、特定のユースケースと達成された相対的なパフォーマンスに基づいています。このレポートに含まれる表は、その現実を反映しており、LLMがユースケースに適用された際の相対的なパフォーマンスに基づいて色分けされています。青の濃淡は最高の相対的パフォーマンスレベルを示し、赤の濃淡はユースケースに対して劣った相対的パフォーマンスを示し、白の濃淡は平均的な相対的パフォーマンスを示します。

そのため、90%のパフォーマンス評価は、特定のユースケースに関連する範囲の中で最低のパフォーマンス評価であれば赤色で表示されることがあります。90%のパフォーマンスはユースケースに対して許容されるかもしれませんが、他のLLMが定義されたタスクを実行した際のパフォーマンスと比較すると、依然として劣っていると評価されます。

コンテキストサイズ

コンテキストサイズは依然として関連性がありますが、LLMの評価においては、もはや当たり前のことと見なされることが急速に進行しています。ほとんどの主要モデルは128kトークンのコンテキストを特徴としており、これは大多数の保険利用ケースをサポートするのに十分です。

結果と分析

LLM定量比較

F1スコア	GPT4o	GPT4o-Mini	GPT3.5	Claude3.5 Sonnet	Claude3 Opus	Claude3 Sonnet	Claude3 Haiku	Mistral Large 2407	Mistral Large	Llama 3.1-405b	Llama 3.1-70b	Llama 3.1-8b	Llama 3-70b	Llama 3-8b
10万件の文書の価格	€1,840	€112	€321	€2,503	€12,519	€2,503	€208	€2,604	€5,186	€3,471	€1,326	€168	€2,443	€238
フランス語の歯科請求書	93.7%	90.0%	91.8%	94.3%	93.7%	93.3%	93.2%	93.5%	91.5%	93.1%	91.8%	90.4%	90.9%	89.0%
日本語の住宅修理見積書	83.0%	78.4%	79.2%	83.0%	83.0%	82.2%	78.4%	82.5%	81.6%	83.3%	79.8%	71.0%	79.1%	72.1%
英語の航空会社請求書	82.6%	75.3%	69.9%	81.5%	82.2%	77.8%	72.9%	82.1%	78.7%	83.5%	79.2%	65.0%	75.0%	65.5%
文書分類	89.5%	84.8%	81.5%	88.2%	87.1%	85.5%	78.2%	87.2%	70.7%	83.7%	68.3%	62.2%	78.5%	20.4%
文書分類ID有	91.6%	85.6%		88.7%	86.7%	88.7%	79.3%	87.8%		87.2%	87.9%	67.5%		
文書分類ID無	91.3%	85.8%		88.1%	88.0%	86.2%	79.3%	88.0%		88.2%	87.7%	70.4%		
すべてのユースケースの集約	87.2%	82.1%	80.6%	86.7%	86.5%	84.7%	80.7%	86.3%	80.6%	85.9%	79.7%	72.2%	80.9%	61.8%

上記の表を参照すると、テストされたLLMに関して以下の結論を引き出すことができます：

- **GPT4o**は、87.2%の集計パフォーマンススコアで他のモデルを上回りました。また、レポート第二弾がリリースされて以来、GPT4oの価格は実質的に半分に削減されました。このモデルは単純なタスクと複雑なタスクの両方で良好に機能し、多くの組織にとって堅実な選択肢となります。
- **GPT4o-Mini**は、82.1%の尊敬すべき集計パフォーマンススコアを達成しました。GPT3.5の後継として意図されており、単純なタスクにおいてはわずかなパフォーマンス低下（約2%）が見られましたが、複雑なタスクにおいては顕著なパフォーマンス向上（約3-5%）が見られました。価格/パフォーマンスメトリクスに基づくと、GPT3.5の約3倍のコストであるGPT4o-Miniは、適切な後継モデルであると考えています。
- 86.7%のスコアを持つClaude3.5 Sonnetは、2番目に高い集計パフォーマンススコアを達成しました。この新しいモデルは、Claude3 Opusをわずかに上回るパフォーマンスを示しましたが、Claude3 Sonnetと同等の価格です。その全体的なパフォーマンスはGPT4oと比較可能ですが、依然として高価です。
- **Mistral Large 2407**は、Mistral Largeに対して約6%のパフォーマンス向上を示し、価格はほぼ半分です。86.3%の集計スコアは、GPT4oおよびClaude3.5 Sonnetと近接した能力を示しています。ここでも、GPT4oの価格削減が価格/パフォーマンス比に大きな影響を与えています。
- Llama3の以前のバージョンと比較してパフォーマンスが大幅に改善されたものの、Llama3.1 70bおよびLlama3.1 8bは集計において劣っており、複雑なタスクを完了する際に期待を下回りました。他のLLMと比較して、特定の保険シナリオに適用された際に単に成果を上げていません。
- 85.9%の集計パフォーマンススコアを持つLlama3.1 405bは、評価された主要モデルに近いですが、英語の旅行文書の「分類」におけるリーダーとの比較での劣ったパフォーマンスと、相対的に高いコストがこのモデルを他のいくつかのモデルよりも魅力を欠くものにしています。

結論

LLMの最先端技術は急速に進化しており、新しいバージョンや新しいモデルが短期間でリリースされています。これに追いつくのは非常に難しい場合があります。同時に、テストしたいくつかのモデルが非常に似た結果を達成する興味深い接合点に達しているようです。今後、価格がどのLLMが組織に最適かを決定する重要な要因となる可能性があります。現時点では、特にGPT4oおよびGPT4o-Miniが、これらの4つの保険業界シナリオに対してテストされた際に最良の価格/パフォーマンス比較を示しています。.