SHIFT

保険業界におけるAIの現状

LLM の比較 (Vol. II)

www.shift-technology.com/ja

エグゼクティブサマリー

- **LLM 市場は急速に進化しており**、現在ではさまざまなユースケースに適した多様なモデルが提供されています
- **どの LLM がどのユースケースに最適**であるかを判断するには、コンテキストのサイズ、全体的なコスト、パフォーマンスを比較する必要があります
- **焦点を絞ったプロンプトエンジニアリングと調整**が、卓越したパフォーマンスと期待に沿わないパフォーマンスを分ける要因となります

編集者より

人工知能 (AI) と生成人工知能 (GenAI) の活用が浸透することで保険の重要なプロセスが改善されており、継続的に保険業界を魅了しています。その一方で、急速に変化する状況を乗り切り、これらのイノベーションをどのように導入すれば最良の結果が得られるかについて最善の決断を下すことは、非常に難しくなっています。

前回のレポート保険業界における AI の現状では、保険に特化したいくつかのユースケースに適用した 6 つの異なる大規模言語モデル (LLM) のパフォーマンスを調査しました。シフトのデータサイエンティストと研究者は、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 に関するコストとパフォーマンスの比較を試みました。

Vol.II では、新たに 8 つの LLM をテストし、前回のレポートに登場した 2 つの LLM は外しました。新たにテストされたモデルは、Llama3-8b、Llama3-70b、GPT4o、Command r、Command r+、Claude3 Opus、Claude3 Sonnet、Claude3 Haiku です。前回のレポートに掲載した Llama2 モデルを比較対象から外し、Llama3 モデルに変更しました。Llama3 は、現在利用可能な LLM の中でも代表的な最先端のモデルです。

さらに、このレポートでは各モデルで生成された F1 スコアをハイライトした新しい表を追加しています。このレポートでは、F1 スコアは特定のユースケースを表わす 2 つの軸 (例 : フランス語の歯科請求書) とこれに関連する個々のフィールドに対するカバー率と精度をまとめています。このアプローチにより、ユースケースごとに 1 つのパフォーマンス指標を生成できるだけでなく、10 万もの文書の分析に関連するコストを含んだ総合的なスコアも生成できます。F1 スコアの生成に使用した式 : 2 x Cov x Acc / (Cov + Acc)

本レポートの作成にご協力いただいたシフトのデータサイエンス チームと研究チームに皆様に感謝いたします。

LLM モデルの比較 : 情報の抽出と保険医関する書類の選択

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、11 の公開されている大規模言語モデルのパフォーマンスを評価しました。このモデルには次が含まれます : GPT3.5、GPT4、GPT4o、Mistral Large、Llama3-8b、Llama3-70b、Command r、Command r+、Claude3 Opus、Claude3 Sonnet、Claude3 Haiku

実施したシナリオ :

- 英語の航空会社の請求書からの情報抽出
- 日本語の不動産修繕見積書
- フランス語の歯科請求書からの情報抽出
- 旅行保険請求に関連する英語の文書の文類

LLM に実施したテストの内容 :

カバレッジ - LLM の出した結果。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、抽出したデータ。

正確性 - LLM は、何かが抽出され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と研究チームが実施しました。個々のシナリオごとに単一のプロンプトを設計し、テスト対象のすべての LLM で使用しました。すべてのプロンプトは GPT LLM に最適化されているため、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ります。

表の読み方

LLM パフォーマンスの評価は、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行われます。本レポートに含まれる表は、このような現実を反映し、ユースケースに適用された LLM 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。青の濃淡は相対的なパフォーマンスレベルが最も高いことを表し、赤の濃淡はユースケースの相対的なパフォーマンスが低いことを表し、白の濃淡は平均的な相対的パフォーマンスを表しています。

そのため、この表記ではパフォーマンスの評価が 90% であっても、特定のユースケースに関する性能評価の最低値が 90% である場合は、赤色で表示されます。パフォーマンス評価としての 90% はユースケースを考えれば許容範囲である可能性もありますが、当該タスクに対する他の LLM のパフォーマンスと比較して、基準値未満と評価されます。

コンテキストサイズ

各 LLM はそれぞれコンテキストサイズを備えており、これはプロンプトの入力からレスポンスの出力までの間にモデルが扱えるトークン (テキストを単語や文字などの意味のある単位に分割したもの) の最大数を示します。基本的に、コンテキストサイズが小さいモデルは、長い文書の分析には適していません。そのため、パフォーマンスに影響を与える可能性のある要因のひとつとして、モデルのコンテキストサイズに着目しました。しかし、コンテキストサイズが大きいほどパフォーマンスが向上するとは言い切れません。たとえば、Llama3-70b のコンテキストサイズは 8k と比較的小さいですが、はるかに大きなコンテキストを持つモデル (Command r+ は 128k) と同等のパフォーマンスを発揮する場合が多くみられました。同様に、Claude3 Haiku と Gpt3.5-turbo は同等の結果を示しましたが、Claude3 のコンテキストサイズは Gpt3.5-turbo の約 10 倍です。

モデル	コンテキストサイズ
mistral-large	32k
llama3-70b-instruct	8k
llama3-8b-instruct	8k
gpt4o	128k
gpt-4-0125-preview	128k
gpt-3.5-turbo-0125	16k
command r+	128k
command r	128k
claude3-sonnet	200k
claude3-opus	200k
claude3-haiku	200k

結果と分析

英語の航空会社の請求書

このシナリオでは、テスト対象の LLM が匿名化された 85 件の英語の請求書を分析しました。

抽出プロンプトでは、次の結果を求めました。

- プロバイダー名
- 開始日
- 終了日
- 文書の日付
- 予約番号
- フライト番号 (関連するすべてのフライト番号)
- クレジットカードの最後の 4 桁
- 通貨
- 全乗客の基本運賃
- 全乗客の税金と手数料 (すべての税金と手数料のリスト)
- 全乗客の追加料金 (追加料金のリスト)
- 合計金額
- 合計支払額
- 支払い (複合的なフィールド : 支払日、金額、ステータスのフィールドを含むオブジェクト)
- 旅行者 (複合的なフィールド : 旅行者のリスト、旅行者名、基本運賃、合計税金、合計金額を含むオブジェクト)

英文のフライト請求書	GPT4		GPT4o		GPT3.5		Claude3 Opus		Claude3 Sonnet		Claude3 Haiku	
メトリック名	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性
プロバイダー名	98.5%	67.1%	100.0%	68.8%	98.5%	59.5%	100.0%	63.0%	95.5%	61.1%	100.0%	61.2%
開始日	98.3%	83.1%	100.0%	84.8%	98.3%	76.1%	100.0%	83.6%	94.8%	85.7%	100.0%	63.5%
終了日	100.0%	82.7%	100.0%	82.1%	97.9%	66.2%	100.0%	69.8%	100.0%	79.3%	100.0%	51.8%
文書の日付	95.3%	80.3%	95.3%	79.1%	96.9%	67.5%	93.8%	73.2%	96.9%	71.6%	100.0%	62.4%
予約番号	96.7%	71.8%	96.7%	82.8%	98.3%	3.8%	88.3%	70.3%	93.3%	29.3%	100.0%	2.4%
フライト番号	98.5%	65.5%	100.0%	61.2%	98.5%	50.0%	100.0%	62.4%	94.0%	53.8%	100.0%	48.2%
クレジットカードの最後の 4 桁	98.0%	94.2%	100.0%	96.2%	98.0%	90.7%	100.0%	96.2%	94.1%	95.9%	100.0%	89.3%
通貨	98.3%	96.7%	98.3%	95.2%	98.3%	93.7%	91.7%	96.5%	90.0%	88.5%	100.0%	71.4%
全乗客の基本運賃	97.0%	51.7%	100.0%	54.5%	100.0%	33.3%	100.0%	61.5%	90.9%	57.7%	100.0%	49.2%
全乗客の税金と料金	96.9%	44.4%	100.0%	45.3%	100.0%	23.2%	100.0%	51.9%	90.6%	48.0%	100.0%	41.8%
すべての乗客の追加料金	91.7%	28.0%	91.7%	34.8%	75.0%	12.5%	91.7%	33.3%	83.3%	43.8%	91.7%	30.4%
保険の追加料金	100.0%	86.7%	100.0%	92.9%	69.2%	75.0%	100.0%	100.0%	92.3%	100.0%	100.0%	81.3%
合計金額	93.2%	91.4%	98.3%	98.3%	91.5%	89.7%	96.6%	93.2%	88.1%	92.6%	96.6%	88.7%
合計支払額	92.5%	90.6%	62.3%	100.0%	69.8%	81.4%	83.0%	87.5%	83.0%	93.6%	96.2%	73.3%
支払日	95.7%	38.2%	39.1%	32.0%	82.6%	34.7%	73.9%	42.1%	82.6%	32.1%	100.0%	31.3%
支払状況	89.3%	76.6%	46.4%	86.7%	58.9%	66.0%	75.0%	79.2%	78.6%	74.6%	76.8%	65.2%
支払金額	88.7%	96.9%	42.3%	100.0%	60.6%	86.0%	73.2%	98.1%	78.9%	91.5%	80.3%	83.3%
旅行者基本料金	82.8%	75.4%	58.6%	87.2%	50.0%	32.4%	74.1%	72.9%	67.2%	56.7%	89.7%	57.5%
旅行者合計税金	83.0%	60.7%	63.8%	71.8%	53.2%	36.8%	74.5%	56.9%	63.8%	42.4%	89.4%	47.1%
旅行者合計金額	83.1%	80.3%	57.6%	84.6%	49.2%	41.2%	72.9%	69.0%	66.1%	43.9%	86.4%	55.2%

(次のページに続く)

(前ページからの続き)

英文のフライト請求書	Mistral Large		Command r+		Command r		Llama3-70b		Llama3-8b	
メトリック名	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性
プロバイダー名	100.0%	63.8%	100.0%	61.4%	92.5%	56.3%	98.5%	64.1%	95.5%	56.1%
開始日	100.0%	79.7%	100.0%	73.6%	96.6%	65.8%	98.3%	78.3%	96.6%	63.9%
終了日	100.0%	79.6%	97.9%	75.9%	93.6%	58.6%	100.0%	67.7%	95.7%	58.8%
文書の日付	100.0%	70.5%	100.0%	59.5%	95.3%	60.0%	95.3%	68.4%	96.9%	50.6%
予約番号	98.3%	55.1%	100.0%	42.5%	93.3%	6.3%	98.3%	4.9%	96.7%	10.8%
フライト番号	100.0%	61.2%	100.0%	55.3%	94.0%	52.5%	98.5%	64.6%	97.0%	44.6%
クレジットカードの最後の 4 桁	100.0%	94.3%	98.0%	85.2%	94.1%	69.8%	92.2%	90.0%	92.2%	75.4%
通貨	100.0%	75.0%	100.0%	70.6%	95.0%	72.2%	95.0%	89.1%	96.7%	71.6%
全乗客の基本運賃	100.0%	57.4%	97.0%	45.8%	97.0%	34.2%	90.9%	54.7%	93.9%	40.3%
全乗客の税金と料金	96.9%	42.6%	96.9%	36.5%	96.9%	23.7%	87.5%	37.3%	90.6%	31.4%
すべての乗客の追加料金	91.7%	13.8%	91.7%	26.1%	91.7%	9.5%	66.7%	19.0%	83.3%	8.0%
保険の追加料金	100.0%	92.9%	100.0%	92.9%	84.6%	19.3%	92.3%	92.3%	100.0%	65.0%
合計金額	96.6%	91.5%	96.6%	84.7%	89.8%	68.0%	91.5%	93.0%	93.2%	86.4%
合計支払額	90.6%	84.6%	92.5%	87.0%	88.7%	58.3%	83.0%	82.0%	83.0%	66.0%
支払日	73.9%	32.0%	69.6%	42.9%	100.0%	24.7%	91.3%	40.8%	52.2%	40.0%
支払状況	73.2%	70.7%	89.3%	74.6%	71.4%	47.6%	78.6%	72.1%	69.6%	79.6%
支払金額	74.6%	91.4%	87.3%	93.9%	74.6%	61.0%	78.9%	88.5%	66.2%	90.0%
旅行者基本料金	70.7%	60.3%	60.3%	44.9%	60.3%	22.5%	51.7%	63.6%	41.4%	35.0%
旅行者合計税金	68.1%	43.3%	72.3%	35.7%	59.6%	23.4%	46.8%	51.2%	36.2%	32.0%
旅行者合計金額	71.2%	54.2%	67.8%	34.9%	59.3%	29.7%	50.8%	65.9%	37.3%	29.8%

分析

新しい LLM をテスト環境に導入した結果、興味深い結果が得られました。「シンプルなフィールド」(日付、型、金額などの単純なデータ型を含むフィールド、および値が一意であり、リスト形式ではないその他のフィールド) と分類されるものについては、GPT4o、GPT4、Claude3 Opus が他のすべてのモデルを凌駕し、GPT4o が最も優秀な結果を出しました。ただし、ひとつ例外があります。総支払額のカバレッジでは、GPT4o は他の 2 つの主要 LLM と比較して下回る結果となりました。この例外の原因となった理由はまだ判明しておらず、原因の特定するにはさらなる調査と実験が必要です。

Claude3 Sonnet、Mistral Large、Command r+、Command r は、上位の LLM に近いパフォーマンスを示しましたが、若干低い結果となっています。Llama3 、GPT3.5、Claude3 Haiku はそれぞれ類似する結果を収めましたが、上位 7 つのモデルには及びませんでした。前回のレポートでは Llama2 と他のモデルとの間のパフォーマンスの差が大きく目立っていましたが、この差は Llama3 の導入で大幅に縮まりました。この要因は、Llama3 の方がベースコンテキストのサイズが大きい (4k > 8k) ためと思われます。

「複雑なフィールド」(複雑なオブジェクトを表すフィールド、またはオブジェクトのリストを値とするフィールド) については、GPT4 と Claude3 Opus がこのようなタスクに最適なモデルであり、GPT4 が Claude3 Opus をわずかに上回っていることがわかりました。

Claude3 (Sonnet と Haiku の両方)、Mistral Large、Llama3-70b は、上位のモデルほどのパフォーマンスを発揮できませんでした。興味深い点として、これらのモデルの性能は分析対象のフィールドによって大きく異なり、要求されるデータによっては各モデルが他のモデルをわずかに上回るということがわかりました。

最後に、このシナリオでは、GPT4o は精度の点では非常に優れていますが、カバレッジの点では驚くほど悪いことがわかりました。この結果は、モデルが複雑な情報を取得できていないこと、あるいは単にこれらのプロンプトエンジニアリングやフィールドの調整を促すことで、問題が解決する可能性が高いことを示唆しているとも言えるでしょう。

日本語の不動産修繕見積書

私たちは、さまざまなサービスプロバイダーに関連する 100 件の匿名化された日本語の不動産修繕の見積もりを評価しました。これら文書の形式は標準化されていませんでした。

抽出プロンプトは、以下のフィールドからのデータを要求しました :

- プロバイダー名
- プロバイダーの住所
- 郵便番号
- プロバイダーのメール
- 課税額
- 税込総額
- 割引額

日本語の不動産修繕見積	GPT4		GPT4o		GPT3.5		Claude3 Opus		Claude3 Sonnet		Claude3 Haiku	
メトリック名	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性
プロバイダー名	98.9%	73.1%	98.9%	74.2%	98.9%	68.1%	98.9%	74.5%	100.0%	76.8%	98.9%	68.7%
プロバイダーの住所	91.2%	70.2%	89.0%	69.5%	93.4%	62.8%	92.3%	68.2%	93.4%	58.1%	93.4%	66.7%
郵便番号	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
プロバイダーのメール	100.0%	63.6%	100.0%	70.0%	100.0%	63.6%	100.0%	70.0%	100.0%	70.0%	100.0%	43.8%
課税額	100.0%	86.0%	100.0%	85.9%	96.4%	77.5%	100.0%	81.8%	100.0%	83.5%	100.0%	78.8%
税込総額	100.0%	97.0%	100.0%	93.9%	97.9%	95.9%	100.0%	96.0%	100.0%	94.0%	99.0%	94.9%
割引額	100.0%	9.6%	100.0%	5.5%	73.3%	10.0%	100.0%	9.1%	100.0%	9.1%	100.0%	10.3%

日本語の不動産修繕見積	Mistral Large		Command r+		Command r		Llama3-70b		Llama3-8b	
メトリック名	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性
プロバイダー名	98.9%	72.3%	100.0%	65.6%	98.9%	65.7%	100.0%	72.6%	98.9%	66.3%
プロバイダーの住所	90.1%	69.9%	90.1%	46.4%	92.3%	50.0%	92.3%	67.1%	94.5%	66.7%
郵便番号	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	15.9%	100.0%
プロバイダーのメール	100.0%	63.6%	100.0%	53.8%	100.0%	26.9%	100.0%	58.3%	100.0%	38.9%
課税額	100.0%	85.7%	95.2%	77.5%	98.8%	60.4%	100.0%	70.7%	100.0%	51.0%
税込総額	99.0%	92.9%	100.0%	96.0%	99.0%	76.8%	100.0%	92.9%	99.0%	88.9%
割引額	93.3%	4.9%	53.3%	18.5%	93.3%	3.7%	100.0%	3.0%	100.0%	2.1%

分析

全体的に GPT4o、GPT4、Mistral、Claude3 (Opus と Sonnet) が、このシナリオで最も優れた結果を示し、これらすべてが同様の結果を導き出しました。

GPT3.5、Claude3 Haiku、Llama3-70b に関してはパフォーマンスにわずかな低下を観測しました。また、これらのモデルはほとんどのフィールドで同等のパフォーマンスを達成しましたが、その他のフィールドではパフォーマンス不足が評価に影響しました。

カバレッジの観点からは、Command r、Command r+、Llama3-8b の各モデルは他のモデルと比較して十分なパフォーマンスを示しました。しかし、精度の面では明らかに期待値を下回る評価となっています。この結果は、これらのモデルが日本語の処理が苦手であることを示唆していると思われるため、この事象をさらに調査する予定でした。

最初のテストでは、テキストフィールド (プロバイダー名、プロバイダーの住所、プロバイダーのメールなど) の精度が多少低く感じられるかもしれません。しかし、金額や日付の場合と同様に、構造化された形式でモデルの出力を検証することが不可能であることを考えると、まったく想定外というわけではありません。これらのメトリクスは非常に厳密であるため、グランドトゥルースと予測がまったく同じである必要があります。

すべてのモデルにおいて、「郵便番号」と「割引額」のフィールドで一貫して低いパフォーマンスを示していることには驚きました。何が原因であるかは現時点では不明です。さらなる調査を実施する予定です。

フランス語の歯科請求書

このシナリオでは、119 件の匿名化されたフランス語の歯科請求書に対して LLM のテストを実施しました。そのうち 79 件はテンプレート化されたもので、残りの 60 件は無作為に選ばれたものです。できあがったデータセットの約 60 パーセントがテンプレート化されたものであると言えるでしょう。この方法は、典型的な歯科保険業者のデータセットを反映する目的で選択しました。

各 LLM には以下のような抽出を依頼しました :

- 文書の日付
- プロバイダー名
- 原本データ提供者 FINESS (Fichier National des Établissements Sanitaires et Sociaux - 医療・社会福祉施設の全国ファイル)
- データ提供者 RPPS (Répertoire Partagé des Professionnels de Santé - フランスの医療専門家共有ディレクトリ)
- プロバイダーの郵便番号
- 発生額合計
- 支払金額

	GPT4		GPT4o		GPT3.5		Claude3 Opus		Claude3 Sonnet		Claude3 Haiku	
メトリック名	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性
文書の日付	100.0%	95.7%	100.0%	95.7%	99.3%	93.5%	100.0%	97.1%	100.0%	95.7%	100.0%	97.1%
プロバイダー名	100.0%	95.0%	100.0%	95.0%	100.0%	93.5%	100.0%	94.2%	100.0%	94.2%	100.0%	95.0%
原本データ提供者 Finess	65.1%	61.9%	74.4%	50.0%	58.1%	53.7%	67.4%	71.8%	62.8%	65.0%	65.1%	61.9%
原本データ提供者 Rpps	97.1%	92.5%	96.2%	95.1%	92.3%	94.9%	98.1%	93.4%	89.4%	92.7%	96.2%	95.1%
プロバイダーの郵便番号	99.3%	99.3%	100.0%	98.6%	99.3%	99.3%	99.3%	98.6%	99.3%	98.6%	99.3%	99.3%
発生額合計	100.0%	97.8%	100.0%	97.1%	100.0%	96.4%	100.0%	97.1%	100.0%	97.8%	100.0%	97.8%
支払金額	100.0%	69.2%	98.5%	81.8%	95.6%	74.4%	98.5%	73.6%	98.5%	83.8%	98.5%	71.8%

	Mistral Large		Command r+		Command r		Llama3-70b		Llama3-8b	
メトリック名	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性
文書の日付	100.0%	96.4%	100.0%	95.7%	95.0%	97.1%	100.0%	95.7%	99.3%	88.4%
プロバイダー名	100.0%	94.2%	100.0%	92.8%	95.0%	96.4%	100.0%	94.2%	99.3%	91.3%
原本データ提供者 Finess	69.8%	64.4%	72.1%	76.9%	82.1%	100.0%	62.8%	61.9%	65.1%	65.9%
原本データ提供者 Rpps	91.3%	92.9%	89.4%	92.8%	93.0%	95.9%	94.2%	94.1%	95.2%	94.1%
プロバイダーの郵便番号	100.0%	97.8%	99.3%	98.6%	97.1%	97.8%	99.3%	99.3%	89.1%	97.6%
発生額合計	100.0%	98.6%	100.0%	98.6%	96.4%	97.8%	100.0%	97.8%	99.3%	94.9%
支払金額	98.5%	55.8%	98.5%	50.4%	45.0%	85.3%	98.5%	49.6%	95.6%	50.8%

分析

全体的に GPT4o、GPT4、Mistral、Claude3 (すべてのバージョン) が、このシナリオで最も優れた結果を示し、これらすべてが同様の結果を導き出しました。しかし、GPT4o と Claude3 Opus のパフォーマンスは非常に近く、評価も他のモデルをわずかに上回っています。

Command r+、Llama3-70b、Mistral Large は GPT3.5 と同等であるのに対し、他の Llama モデルと Command r は下回っていることがわかりました。

FINESS (デジタルヘルス機関が管理する全国のディレクトリに関連付けられた識別子) フィールドで観測された期待値未満のパフォーマンスは、フランスの医療請求書が常に FINESS または他のデータ提供者の識別子 (AM または SIRET) を明確に識別していないという事実に起因している可能性があります。この混乱は、モデルの情報検索能力に影響を与えるだけでなく、このフィールドのグランドトゥルースの質にも影響を与える可能性があります。

旅費請求の英語文書

このシナリオでは、旅費申請時に提出された匿名化された英文の文書 405 件を使用しています。

プロンプトではモデルに次のように尋ねます :

- 各ページの分類
- 同じ文書に関連するページをグループ化します (同じファイルに複数の文書を配置できるため)。
- 各ファイルは、分割された文書のリストとして出力されます。各要素は、文書の種類と、ファイル内の開始ページと終了ページを示すスパンを含んでいます

他のシナリオと同様に、各文書タイプの典型的なカバレッジと精度の指標を個別にレポートします。加えて、2 つの集約された指標も含んでいます :

- 完全な分類 : ここでは、ファイル内のすべてのセグメント化された文書にエラーがない場合 (ドキュメント タイプとページのスパン) に、モデルの出力が正しいとみなします
- 完全なタイプ : ここでは、ファイル内のすべての文書タイプにエラーがない (ページ スパンのエラーは考慮しない) 場合に、モデルの出力が正しいとみなします (PerfectTypes)。

他のシナリオと同様、初回のレポートから Vol.2 までの間に、このシナリオのプロンプトが若干調整されていることに留意してください。GPT4o で期待値を下回る想定外のパフォーマンスを観測した後、私たちは調査の結果プロンプトのエラーを発見しました。具体的には、モデルに要求する出力に「JSON」ではなく、「markdown JSON」を指定していました。その結果、想定しないセパレータが追加され、正しい分析がされませんでした。これを修正するため、プロンプトを調整しました。

分類	GPT4		GPT4o		GPT3.5		Claude3 Opus		Claude3 Sonnet		Claude3 Haiku		Mistral Large		Command r+		Command r		Llama3-70b		Llama3-8b	
メトリック名	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性
領収書 - 飛行機	92.8%	85.2%	95.2%	88.2%	85.5%	77.1%	96.4%	80.9%	95.2%	74.7%	88.0%	54.1%	95.2%	86.4%	66.3%	69.1%	60.2%	69.5%	85.5%	79.3%	15.7%	45.0%
領収書 - ホテル / レンタル予約	85.0%	80.0%	85.0%	80.0%	75.0%	68.4%	90.0%	78.9%	90.0%	76.2%	85.0%	73.7%	85.0%	71.4%	75.0%	54.2%	60.0%	66.7%	75.0%	73.7%	0.0%	0.0%
領収書 - アクティビティ予約	83.3%	36.4%	16.7%	0.0%	33.3%	50.0%	33.3%	16.7%	16.7%	20.0%	16.7%	33.3%	33.3%	25.0%	0.0%	0.0%	16.7%	16.7%	50.0%	30.0%	0.0%	0.0%
領収書 - クルーズ	91.7%	77.8%	91.7%	80.8%	79.2%	66.7%	79.2%	72.7%	87.5%	70.0%	66.7%	76.2%	91.7%	63.3%	75.0%	69.2%	33.3%	53.8%	75.0%	64.0%	4.2%	50.0%
領収書 - 電車	100.0%	100.0%	66.7%	100.0%	100.0%	33.3%	66.7%	100.0%	100.0%	66.7%	100.0%	66.7%	66.7%	25.0%	33.3%	100.0%	66.7%	50.0%	100.0%	66.7%	0.0%	0.0%
銀行 / クレジットカードの明細書	93.8%	90.3%	100.0%	88.2%	43.8%	93.3%	90.6%	89.7%	93.8%	84.4%	71.9%	95.8%	81.3%	92.9%	68.8%	74.1%	37.5%	69.2%	75.0%	82.1%	3.1%	33.3%
キャンセル ポリシー	33.3%	50.0%	33.3%	33.3%	0.0%	0.0%	66.7%	44.4%	16.7%	100.0%	16.7%	50.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	33.3%	100.0%	0.0%	0.0%
キャンセル証明	83.3%	69.6%	87.5%	70.8%	41.7%	81.8%	83.3%	70.8%	62.5%	76.5%	33.3%	75.0%	50.0%	62.5%	29.2%	87.5%	29.2%	85.7%	37.5%	70.0%	0.0%	0.0%
医療費請求書	88.9%	50.0%	88.9%	87.5%	66.7%	71.4%	66.7%	71.4%	77.8%	85.7%	44.4%	30.0%	66.7%	55.6%	66.7%	71.4%	55.6%	71.4%	66.7%	71.4%	0.0%	0.0%
医療報告書	86.7%	90.9%	98.3%	88.9%	89.2%	91.2%	97.5%	92.6%	96.7%	88.0%	80.0%	90.1%	27.5%	85.7%	85.0%	94.3%	60.8%	92.0%	77.5%	87.3%	6.7%	87.5%
支払証明	66.7%	53.8%	41.7%	44.4%	16.7%	66.7%	33.3%	44.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	33.3%	80.0%	25.0%	50.0%	8.3%	100.0%	0.0%	0.0%
英文の旅行書類請求フォーム 1 - 表示	100.0%	100.0%	100.0%	75.0%	66.7%	28.6%	100.0%	20.0%	100.0%	42.9%	100.0%	25.0%	33.3%	0.0%	66.7%	14.3%	66.7%	22.2%	33.3%	33.3%	0.0%	0.0%
英文の旅行書類請求フォーム 1 - 申請の情報 / 代理店の詳細	100.0%	100.0%	100.0%	80.0%	25.0%	100.0%	75.0%	50.0%	100.0%	66.7%	100.0%	50.0%	25.0%	100.0%	25.0%	11.1%	25.0%	16.7%	25.0%	25.0%	0.0%	0.0%
英文の旅行書類請求フォーム 1 - 損失の詳細 / インシデントの内容	100.0%	100.0%	100.0%	100.0%	50.0%	100.0%	100.0%	44.4%	75.0%	60.0%	100.0%	50.0%	25.0%	100.0%	50.0%	20.0%	25.0%	14.3%	25.0%	100.0%	0.0%	0.0%
英文の旅行書類請求フォーム 1 - 申請した経費 / 支払い	100.0%	100.0%	100.0%	100.0%	50.0%	100.0%	100.0%	57.1%	100.0%	80.0%	100.0%	50.0%	25.0%	100.0%	25.0%	11.1%	25.0%	20.0%	25.0%	100.0%	0.0%	0.0%
英文の旅行書類請求フォーム 1 - 必要書類	100.0%	80.0%	100.0%	100.0%	50.0%	100.0%	100.0%	57.1%	100.0%	80.0%	100.0%	50.0%	0.0%	0.0%	25.0%	11.1%	50.0%	28.6%	25.0%	100.0%	0.0%	0.0%
英文の旅行書類請求フォーム 1 - 承認と割り当て	100.0%	100.0%	100.0%	50.0%	50.0%	40.0%	100.0%	36.4%	100.0%	66.7%	100.0%	36.4%	25.0%	100.0%	50.0%	16.7%	25.0%	20.0%	25.0%	20.0%	0.0%	0.0%
英文の旅行書類請求フォーム 2 - 一般的な情報	100.0%	100.0%	100.0%	100.0%	0.0%	0.0%	100.0%	100.0%	0.0%	0.0%	100.0%	20.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
英文の旅行書類請求フォーム 2 - 旅行キャンセル / 旅行中断 / 旅行遅延の詳細	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
英文の旅行書類請求フォーム 2 - 申請した経費と承認	100.0%	100.0%	100.0%	100.0%	0.0%	0.0%	100.0%	100.0%	0.0%	0.0%	100.0%	20.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

(次のページに続く)

(前ページからの続き)

分類	GPT4		GPT4o		GPT3.5		Claude3 Opus		Claude3 Sonnet		Claude3 Haiku		Mistral Large		Command r+		Command r		Llama3-70b		Llama3-8b	
メトリック名	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性	カバレッジ	正確性
英文の旅行書類請求フォーム 3 - 表紙と概要情報	100.0%	100.0%	100.0%	100.0%	40.0%	33.3%	100.0%	100.0%	80.0%	50.0%	60.0%	42.9%	40.0%	50.0%	80.0%	40.0%	60.0%	42.9%	40.0%	50.0%	0.0%	0.0%
英文の旅行書類請求フォーム 3 - 明細	100.0%	100.0%	100.0%	100.0%	50.0%	100.0%	100.0%	100.0%	50.0%	100.0%	50.0%	16.7%	0.0%	0.0%	100.0%	25.0%	100.0%	33.3%	0.0%	0.0%	0.0%	0.0%
英文の旅行書類請求フォーム 3 - 割り当てと承認	85.7%	100.0%	42.9%	100.0%	14.3%	100.0%	42.9%	66.7%	57.1%	75.0%	14.3%	20.0%	42.9%	100.0%	42.9%	33.3%	28.6%	33.3%	0.0%	0.0%	0.0%	0.0%
英文の旅行書類 診断書フォーム 1 - 被保険者と医師の情報	100.0%	100.0%	100.0%	50.0%	0.0%	0.0%	100.0%	100.0%	100.0%	100.0%	100.0%	14.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
英文の旅行書類 診断書フォーム 1 - 患者の診断	100.0%	50.0%	100.0%	50.0%	0.0%	0.0%	100.0%	50.0%	100.0%	33.3%	100.0%	20.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
死亡証明 - 死亡証明書または訃報	62.5%	90.9%	81.3%	92.9%	81.3%	92.9%	68.8%	91.7%	81.3%	92.9%	50.0%	80.0%	18.8%	75.0%	75.0%	92.3%	18.8%	100.0%	37.5%	100.0%	6.3%	100.0%
その他	38.6%	66.7%	52.6%	78.4%	21.1%	42.9%	59.6%	40.4%	24.6%	52.0%	31.6%	38.7%	50.9%	50.0%	38.6%	53.6%	35.1%	38.1%	19.3%	68.8%	1.8%	16.7%
英文の旅行書類	100.0%	100.0%	91.7%	100.0%	83.3%	90.0%	100.0%	100.0%	100.0%	100.0%	83.3%	71.4%	33.3%	100.0%	100.0%	70.6%	33.3%	37.5%	50.0%	100.0%	0.0%	0.0%
完全な分類	**95.6%**	**80.1%**	**97.8%**	**82.6%**	**89.4%**	**74.9%**	**98.3%**	**78.1%**	**94.6%**	**78.1%**	**92.8%**	**67.6%**	**95.3%**	**56.2%**	**83.5%**	**76.6%**	**65.2%**	**64.8%**	**87.2%**	**71.4%**	**14.6%**	**33.9%**
完全なタイプ	95.6%	82.2%	97.8%	84.6%	89.4%	76.5%	98.3%	79.9%	94.6%	79.9%	92.8%	68.4%	95.3%	56.5%	83.5%	78.4%	65.2%	67.4%	87.2%	72.8%	14.6%	33.9%

分析

プロンプトエンジニアリングが洗練されたことで、GPT4o のパフォーマンスが低下する問題を解決しました。Mistral Large と Llam3-70b を除き、テストした他の LLM の性能指標に大きな影響はありませんでした。

GPT4 と GPT3.5 のパフォーマンスは、このプロンプトの修正後も安定しており、GPT4 が GPT4o をわずかに下回り、GPT3.5 がその後に続くという、優れた性能を発揮しました。

Llama3-70b はカバレッジが 10% 向上し、GPT3.5 に匹敵する安定した精度を示しました。

プロンプトの調整後、Command r+ と Command r のカバレッジと精度のパフォーマンスは、フィールド全体でさらに高い一貫性を発揮しました。Command r+ の性能は GPT3.5 と Llama3-70b に匹敵し、Command r はこれらをわずかに下回っています。

プロンプトエンジニアリングは、Mistral Large に関する予想外の結果を生み出しました。カバレッジは GPT4 や GPT4o とほぼ同じですが、精度は期待値を下回ったと言えるでしょう。詳細な調査を実施したことで、この低パフォーマンスは、モデルが想定するネーミングに従っていない保険形態に起因していることが判明しました。実際、後続の処理で文書を簡単に正しく再分類することもできますが、モデルがこれらの文書の正しい名前を出力できないことは予期せぬ結果でした。

F1 スコアと結論

	GPT4	GPT4o	GPT3.5	Claude3 Opus	Claude3 Sonnet	Claude3 Haiku
メトリック名	F1	F1	F1	F1	F1	F1
10 万件のドキュメントでの価格	€6,397	€3,227	€321	€12,519	€2,503	€208
フランス語の歯科請求書	92.9%	93.7%	91.8%	93.7%	93.3%	93.2%
日本語の不動産修繕見積	82.9%	83.0%	79.2%	83.0%	82.2%	78.4%
英文のフライト請求書	83.8%	82.6%	69.9%	82.2%	77.8%	72.9%
分類	87.1%	89.5%	81.5%	87.1%	85.5%	78.2%
すべてのユースケースの集約	**86.7%**	**87.2%**	**80.6%**	**86.5%**	**84.7%**	**80.7%**

	Mistral Large	Command r+	Command r	Llama3-70b	Llama3-8b
メトリック名	F1	F1	F1	F1	F1
10 万件のドキュメントでの価格	€5,186	€2,493	€323	€2,443	€238
フランス語の歯科請求書	91.5%	91.4%	90.7%	90.9%	89.0%
日本語の不動産修繕見積	81.6%	75.4%	67.1%	79.1%	72.1%
英文のフライト請求書	78.7%	75.5%	61.0%	75.0%	65.5%
分類	70.7%	79.9%	65.0%	78.5%	20.4%
すべてのユースケースの集約	**80.6%**	**80.6%**	**71.0%**	**80.9%**	**61.8%**

私たちが実施したテストに基づき、以下の分析と結論を導き出すことができます。

シンプルなフィールドとみなされる情報抽出タスクに関しては、GPT4o、GPT4、Claude3 (Opus、Sonnet、Haiku) が最も優れたパフォーマンスを発揮できるモデルであり、すべてのフィールドで同等のパフォーマンスを発揮できることが明らかになりました。

GPT3.5、Mistral Large、Llama3-70b は、最もパフォーマンスの良いモデルと同程度の成果を確認できており、主な違いはフィールド間のパフォーマンスが一貫していない点です。また、Llama3-8b、Command r+、Command r は、全体的には GPT3.5 と同等ですが、特定のフィールドでは期待値を下回ることもわかりました。

複雑なフィールドに関連するタスクでは、GPT4 と Claude3 Opus が最良のモデルであり、GPT4 が Claude3 Opus をわずかに上回っていることがわかりました。

Claude3 (Sonnet and Haiku)、Mistral Large と Llama3-70b は前述の最良のモデルにやや遅れをとっていますが、分析するフィールドによっては各モデルが他をわずかに上回るケースも観測されています。

分類タスクに関しては GPT4o、GPT4、Claude3 (Opus と Sonnet) が最良のパフォーマンスを発揮しており、GPT4o が他をわずかに上回っています。

GPT3.5、Llama3-70b、Command r+、Claude3 Haiku は、上位のモデルと比較すると若干パフォーマンスが下回っており、Llama3-8b と Command r は大きく下回っています。

モデル	入力 /100 万トークン	出力 /100 万トークン	10 万件のドキュメント	コンテキストサイズ
mistral-large	€7.41	€22.22	€ 5,186.00	32k
llama3-70b-instruct	€3.49	€10.47	€ 2,443.00	8k
llama3-8b-instruct	€0.34	€1.02	€ 238.00	8k
gpt4o	€4.61	€13.83	€ 3,227.00	128k
gpt-4-0125-preview	€9.14	€27.41	€ 6,397.00	128k
gpt-3.5-turbo-0125	€0.46	€1.37	€ 321.00	16k
command r+	€2.77	€13.85	€ 2,493.00	128k
command r	€0.46	€1.39	€ 323.00	128k
claude3-sonnet	€2.78	€13.91	€ 2,503.00	200k
claude3-opus	€13.91	€69.55	€ 12,519.00	200k
claude3-haiku	€0.23	€1.16	€ 208.00	200k

ただし、LLM を評価する際には、総合的なパフォーマンスだけでなく、コストに関するパフォーマンスも評価することが非常に重要です。上記のコスト比較表や F1 の表で提示しているとおり、最も高性能なモデルには、最も高いコストを伴うのが一般的です。しかし、ユースケースによっては経済性のために性能を犠牲にすることが許容される場合もあります。たとえば、Claude3 Opus は高性能ですが、法外な費用が発生します。GPT4o と Claude3 Sonnet は、多少お手頃な価格帯で優れたパフォーマンスを実現することで知られていますが、私たちの分析では、価格とパフォーマンスの見事なバランスが取れている組み合わせは GPT3.5-turbo と Claude3 Haiku であることがわかっています。

GenAI と LLM の分野は急速に進化しています。本レポートは、この技術を技術戦略の一環としてどのように活用していくかについて、可能な限り最善の決断を下せるよう、公平な評価をお届けすることを目的としています。