SHIFT

保険業界における AI の現状

保険向け大規模言語モデル — その比較方法

www.shift-technology.com/ja

エグゼクティブサマリー

- 一般的な保険業界のプロセスに適用された 6 つの異なる大規模言語モデル（LLM）の**パフォーマンス比較**
- **コンテキスト サイズ**（テキスト生成時にモデルが記憶できるトークンの最大数）**が大きい LLM** は、例外はあるものの、一般的にパフォーマンスが向上します。
- **コンテキスト サイズを大きく**するとコストが高くなりますが、特定のユース ケースでは**望ましいパフォーマンスを実現するために必要**です。
- **効果的なプロンプト エンジニアリング** は、LLM から可能な限り最高のパフォーマンスを得るための**鍵**です。
- LLM のパフォーマンス指標は**ユース ケースに特有**であり、ビジネス要件が満たされていることを確認するために慎重に評価する必要があります。
- どの LLM を使用するかの選択は、**ユース ケース、許容可能なパフォーマンス、 コスト**の組み合わせに基づいて行う必要があります。

編集者より

生成型人工知能 (生成 AI) アプリケーションと、このテクノロジーをサポートする基盤となる大規模言語モデル (LLM) は、保険業界の注目を集めています。このテクノロジーは、引受査定、保険金請求、不正対策、リスク管理の効率と正確性を大幅に向上させる可能性があります。

しかし、生成 AI に対する関心の高さにもかかわらず、不確実性や未解決の疑問も存在します。保険会社は、生成AI がどこでメリットを得られるか、どこで失敗する可能性があるか、そしてさまざまなユース ケースにどのモデルが最適かなど、これまでにない量の情報に直面しています。これらは、保険会社が生成AI をテクノロジー スタックとビジネス プロセスに導入する方法を検討する際に評価しなければならない問題のほんの一部です。

シフトテクノロジーは、2014 年以来、保険分野における AI のパイオニアとして活躍しています。過去 10 年間にわたり、当社は保険分野における AI に特化した業界最大規模のデータ サイエンス チームを構築してきました。このチームは、保険のユース ケースにおける AI の進歩に向けた研究開発と、その研究開発を応用して保険の顧客向けの革新的なソリューションを開発することに取り組んでいます。

このレポートは、一般的な保険プロセスに適用した場合の特定の大規模言語モデルのパフォーマンスをより深く理解するために、当社のデータサイエンティストが実施した調査の結果を定期的に紹介するシリーズの最初のものです。目標は、保険の専門家に AI に関する信頼できる情報源を提供し、このテクノロジーを評価する際に最善の決定を下せるように支援することです。

このレポートの作成に尽力してくれたシフトのデータサイエンティストおよび研究者に感謝します。

LLM モデルの比較: データ抽出とドキュメント分類

方法論

データサイエンス チームと研究チームは、GPT3.5、GPT4、Mistral Large、Llama2-70B、Llama2-13B、Llama2-7B という 6 つの異なる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するために、4 つのテスト シナリオを考案しました。

シナリオには以下が含まれます。

- 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文書の文書分類

LLM は以下の項目についてテストが実施されました:

カバレッジ - LLM は、基準値（モデルに何かを予測するように依頼したときに私たちが期待する値）が抽出すべきものがあることを示したときに、実際にデータを抽出したか。

正確性 - 何かが抽出されたときに、LLM は正しい情報を提示したか。

すべてのシナリオに対応するプロンプトの設計は、シフトのデータ サイエンス チームによって開発されました。チームは、個々のシナリオごとに、テスト対象の 6 つの LLM すべてで使用される単一のプロンプトを設計しました。

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。このレポートに含まれる表は、その現実を反映しており、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。青は最高の相対的パフォーマンス レベルを表し、赤はユース ケースの標準以下の相対的パフォーマンスを表し、白は平均的な相対的パフォーマンスを表します。そのため、特定のユース ケースに関連付けられた範囲での最低パフォーマンス評価が 90% である場合、パフォーマンス評価が 90% であっても赤でコード化されることがあります。ユース ケースを考慮すると、90% のパフォーマンスは許容範囲かもしれませんが、他の LLM のパフォーマンスと比較すると、依然として期待以下と評価されます。

結果と分析

英語の航空会社の請求書

このシナリオでは、匿名化された英語の航空会社の請求書 85 件が使用されました。

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）
- クレジットカード番号の下 4 桁
- 通貨
- 全乗客向けの基本運賃
- 全乗客向けの税金と料金
- すべての乗客に対する追加料金
- 支払い - 支払日、金額、ステータスを含む複合項目
- 旅行者 - 旅行者名、基本運賃、合計税金、合計金額を含む複合項目

	GPT4		GPT3.5		Mistral Large		Llama2-70b		Llama2-13b		Llama2-7b	
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%	98.5%	59.5%	100.0%	63.8%	43.3%	52.4%	53.7%	33.3%	6.0%	100.0%
開始日	98.3%	83.1%	98.3%	76.1%	100.0%	79.7%	37.9%	47.6%	50.0%	44.4%	5.2%	50.0%
終了日	100.0%	82.7%	97.9%	66.2%	100.0%	79.6%	36.2%	33.3%	42.6%	29.6%	4.3%	50.0%
文書日付	95.3%	80.3%	96.9%	67.5%	100.0%	70.5%	40.6%	45.2%	56.3%	44.4%	4.7%	50.0%
予約番号	96.7%	71.8%	98.3%	3.8%	98.3%	55.1%	36.7%	21.4%	50.0%	1.9%	5.0%	0.0%
フライト番号	98.5%	65.5%	98.5%	50.0%	100.0%	61.2%	41.8%	35.7%	53.7%	33.3%	6.0%	75.0%
クレジットカードの最後の 4 桁	98.0%	94.2%	98.0%	90.7%	100.0%	94.3%	41.2%	60.0%	54.9%	50.0%	3.9%	25.0%
通貨	98.3%	96.7%	98.3%	93.7%	100.0%	75.0%	38.3%	54.8%	55.0%	59.3%	3.3%	50.0%
全乗客基本運賃	97.0%	51.7%	100.0%	33.3%	100.0%	57.4%	39.4%	30.6%	57.6%	20.4%	0.0%	0.0%
すべての乗客の税金と料金	96.9%	44.4%	100.0%	23.2%	96.9%	42.6%	34.4%	16.7%	50.0%	5.8%	0.0%	0.0%
全乗客の追加料金	91.7%	28.0%	75.0%	12.5%	91.7%	13.8%	25.0%	11.8%	25.0%	10.5%	0.0%	0.0%
保険追加料金	100.0%	86.7%	69.2%	75.0%	100.0%	92.9%	38.5%	22.7%	53.8%	17.1%	0.0%	0.0%
合計金額	93.2%	91.4%	91.5%	89.7%	96.6%	91.5%	35.6%	52.8%	50.8%	42.0%	3.4%	25.0%
合計支払額	92.5%	90.6%	69.8%	81.4%	90.6%	84.6%	37.7%	42.9%	54.7%	33.3%	3.8%	25.0%
支払日	95.7%	38.2%	82.6%	34.7%	73.9%	32.0%	21.7%	9.8%	43.5%	4.4%	4.3%	0.0%
支払状況	89.3%	76.6%	58.9%	66.0%	73.2%	70.7%	26.8%	36.6%	42.9%	14.8%	3.6%	25.0%
支払金額	88.7%	96.9%	60.6%	86.0%	74.6%	91.4%	29.6%	46.3%	42.3%	17.0%	2.8%	25.0%
旅行者基本料金	82.8%	75.4%	50.0%	32.4%	70.7%	60.3%	17.2%	24.4%	25.9%	12.3%	3.4%	28.6%
旅行者合計税金	83.0%	60.7%	53.2%	36.8%	68.1%	43.3%	19.1%	21.6%	31.9%	14.2%	0.0%	0.0%
旅行者合計金額	83.1%	80.3%	49.2%	41.2%	71.2%	54.2%	15.3%	19.5%	22.0%	11.3%	0.0%	0.0%

分析

このシナリオでは、GPT4、GPT3.5、Mistral Large が最も優れていることが確認されました。Llama モデルは、特にカバレッジに関しては大幅に劣っていることが判明しました。Llama モデルが関連情報を見つけたり、出力をフォーマットしたりするのが単純に困難になっている状況が発生している可能性があります。結果は、Llama の設定されたコンテキスト サイズが 4k のみであること（これは、テストされた他のどのモデルよりも小さい）によっても影響を受ける可能性があります。この状況では、コンテキスト サイズより大きいドキュメントは全く処理されないため、モデルは結果を返さず、カバレッジ スコアに影響します。

GPT4 と Mistral Large は、複雑なフィールドを処理する際に優れたパフォーマンスを発揮しました。これらの LLM はネストされた情報を抽出できるだけでなく、結果を使用可能な形式で出力することもできます。

リスト フィールドに関連するパフォーマンスは適切なものの、これらの抽出に関連する複雑さによって悪影響を受ける可能性があります。

支払日の場合、精度が低いことが確認されましたが、これは、支払日が不明な場合に、モデルが支払日を文書の日付で代用する傾向があるためと考えられます。

日本語の不動産修繕見積書

このシナリオでは、匿名化された日本語の不動産修繕見積書 100 件に対して各テスト LLM を適用しました。ドキュメントは、複数の異なるプロバイダーからの見積を非標準形式で表したものです。これらのドキュメントはテンプレート化されているとは見なされません。

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

	GPT4		GPT3.5		Mistral Large		Llama2-70b		Llama2-13b		Llama2-7b	
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%	68.1%	98.9%	72.3%	78.3%	68.8%	66.3%	66.3%	23.9%	23.9%
プロバイダーの住所	91.2%	70.2%	93.4%	62.8%	90.1%	69.9%	74.7%	55.7%	64.8%	64.8%	20.9%	20.9%
郵便番号	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	32.9%	96.4%	51.2%	51.2%	12.2%	12.2%
プロバイダーのメールアドレス	100.0%	63.6%	100.0%	63.6%	100.0%	63.6%	87.5%	42.9%	75.0%	75.0%	0.0%	0.0%
課税額	100.0%	86.0%	96.4%	77.5%	100.0%	85.7%	78.3%	74.0%	65.1%	65.1%	24.1%	24.1%
税込総額	100.0%	97.0%	97.9%	95.9%	99.0%	92.9%	78.4%	90.9%	67.0%	67.0%	25.8%	25.8%
割引額	100.0%	9.6%	73.3%	10.0%	93.3%	4.9%	73.3%	0.0%	53.3%	53.3%	33.3%	33.3%

分析

総じて、GPT4、GPT3、Mistral Large は、一部の例外はありますが、カバレッジと正確性の両方で最高のパフォーマンスを発揮しました。Llama70 と Llama13 は正確性がわずかに劣っているものの、カバレッジは明らかに不足しています。これは、前述の航空会社の請求書のシナリオでパフォーマンス不足として特定された同様の特性によるものである可能性があります。

フランス語の歯科請求書

このシナリオでは、各 LLM は 119 件のフランス語の歯科請求書のデータセットに適用されました。請求書のうち 79 件は構造化されたレイアウトであると考えられ、テンプレート化された文書と言えます。残りの 60 件は、保険会社のデータで発生する可能性のある状況を模倣するためにランダムに選択されました。

抽出プロンプトでは、次の結果を求めました。

- 文書の日付
- プロバイダー名
- プロバイダー FINESS（Fichier National des Établissements Sanitaires et Sociaux - 医療・社会福祉施設の全国ファイル）
- プロバイダー 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		GPT3.5		Mistral Large		Llama2-70b		Llama2-13b		Llama2-7b	
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
文書日付	100.0%	95.7%	99.3%	93.5%	100.0%	96.4%	98.5%	89.8%	89.7%	87.8%	90.4%	82.5%
プロバイダー名	100.0%	95.0%	100.0%	93.5%	100.0%	94.2%	98.5%	93.4%	89.1%	90.3%	90.5%	88.9%
Raw プロバイダー Finess	65.1%	61.9%	58.1%	53.7%	69.8%	64.4%	67.4%	70.0%	53.5%	69.7%	60.5%	56.5%
プロバイダー Rpps	97.1%	92.5%	92.3%	94.9%	91.3%	92.9%	95.2%	93.2%	80.8%	94.2%	81.7%	94.3%
プロバイダーの郵便番号	99.3%	99.3%	99.3%	99.3%	100.0%	97.8%	90.5%	96.8%	85.4%	100.0%	81.0%	98.2%
合計発生額	100.0%	97.8%	100.0%	96.4%	100.0%	98.6%	98.5%	97.1%	86.8%	86.8%	89.0%	88.6%
支払金額	100.0%	69.2%	95.6%	74.4%	98.5%	55.8%	95.6%	44.2%	79.4%	32.7%	55.9%	26.1%

分析

このシナリオでは、GPT4、GPT3.5、Mistral Large がカバレッジと正確性の両方で優れたパフォーマンスを発揮しました。残りのモデルのうち、Llama70 は優れたパフォーマンスを発揮しましたが、最高のパフォーマンスを発揮したモデルと同じレベルではありませんでした。

カバレッジと正確性の両方において、プロバイダー FINESS 識別子はすべてのモデルで全体的にパフォーマンスが低かったことに気付きました。これはフランスの医療費請求書の独特な特徴によるものと考えられます。プロバイダー FINESS 識別子は必ずしも明確に示されるわけではなく、SIRET (Système d'identification du répertoire des établissements - ディレクトリ識別システムの構築) などの他のプロバイダー識別子と混同されやすい場合があります。これにより、抽出すべき内容や最終的に抽出されるコンテンツを正確に識別するモデルの能力に影響する可能性があります。

確認されたパフォーマンスの低さも、全体的な混乱の結果である可能性もあります。この分野は、LLM にとっても混乱を招くものです。なぜなら、人間にとっても、それ自体が混乱を招くからです。つまり、LLM を評価するために使用するラベルは、他のフィールドのラベルほど正確ではない可能性があります。追加のプロンプト エンジニアリングはパフォーマンスの向上に役立つ可能性がありますが、基準値自体が本質的に信頼できない場合、パフォーマンスを向上させることは困難です。これは、LLM を評価する際に高品質のラベルを確立することの重要性を示しています。

旅行請求のための英語文書

このデータセットは、旅行保険の請求をサポートするために提供された 405 件の匿名化された英語文書で構成されています。

抽出プロンプトでは、次の結果を求めました。

- 各ページの分類
- 同じ文書に関連するページのグループ

期待される出力は、ドキュメントの種類とページの範囲（開始ページと終了ページを示す）を含むセグメント化されたドキュメントのリストになります。

個々のドキュメント タイプのメトリックに加えて、以下で PerfectClassif および PerfectTypes として定義されているように、ファイル レベルでの集計パフォーマンスも計算します。ファイル内のすべてのセグメント化されたドキュメント (PerfectClassif) が正しい場合、またはファイル内のすべてのドキュメント タイプ (PerfectTypes) が正しい場合、モデルの出力は正しいとみなされます。

	GPT4		GPT3.5		Mistral Large		Llama2-70b		Llama2-13b		Llama2-7b	
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
領収書 - 飛行機	91.6%	82.1%	77.1%	77.1%	75.9%	90.9%	1.2%	0.0%	2.4%	66.7%	2.4%	40.0%
領収書 - ホテル/レンタル予約	90.0%	66.7%	70.0%	54.5%	65.0%	80.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
領収書 - アクティビティ予約	50.0%	33.3%	33.3%	40.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
領収書 - クルーズ	95.8%	67.9%	70.8%	68.2%	75.0%	69.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
領収書 - 電車	100.0%	100.0%	100.0%	66.7%	66.7%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
銀行 / クレジットカードの明細書	96.9%	85.3%	87.5%	82.8%	84.4%	76.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
キャンセル ポリシー	33.3%	16.7%	50.0%	37.5%	16.7%	33.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
キャンセル証明	83.3%	71.4%	58.3%	58.8%	54.2%	70.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
医療費請求書	88.9%	60.0%	77.8%	38.5%	77.8%	55.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
医療報告書	95.0%	89.2%	79.2%	94.8%	69.2%	86.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
支払証明	41.7%	25.0%	50.0%	50.0%	33.3%	66.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
請求書類 1 - 最初のページ	100.0%	100.0%	100.0%	28.6%	66.7%	22.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
請求書類 1 - 請求者情報 / 代理店の詳細	100.0%	100.0%	25.0%	50.0%	50.0%	16.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
請求書類 1 - 損害の詳細 / 事故の説明	100.0%	100.0%	50.0%	66.7%	50.0%	14.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
請求書類 1 - 請求費用 / 支払い	100.0%	100.0%	50.0%	50.0%	50.0%	16.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
請求書類 1 - 必要書類	100.0%	80.0%	75.0%	50.0%	50.0%	14.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
請求書類1 - 承認と譲渡	100.0%	50.0%	75.0%	40.0%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
請求書類 2 - 一般情報	100.0%	50.0%	0.0%	0.0%	100.0%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
請求書類 2 - 旅行キャンセル / 旅行中断 / 旅行遅延の詳細	100.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
請求書類 2 - 請求費用と承認	100.0%	100.0%	0.0%	0.0%	100.0%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
請求書3 - 表紙と概要情報	100.0%	100.0%	60.0%	66.7%	80.0%	57.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
請求書3 - 開示事項	100.0%	100.0%	50.0%	100.0%	100.0%	40.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
請求書類 3 - 譲渡と承認	42.9%	66.7%	14.3%	100.0%	42.9%	33.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
医師の声明書フォーム1 - 被保険者と医師の情報	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
医師のコメント フォーム 1 - 患者の診断	100.0%	33.3%	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	87.5%	93.3%	50.0%	88.9%	18.8%	50.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
その他	42.1%	54.5%	47.4%	38.8%	45.6%	62.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
UsPhysicianStatementFormWholeDoc	100.0%	100.0%	41.7%	80.0%	8.3%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
PerfectClassif	94.6%	80.7%	90.9%	71.7%	84.7%	68.8%	0.2%	0.0%	0.7%	66.7%	1.5%	33.3%
PerfectTypes	94.6%	82.5%	90.9%	72.8%	84.7%	70.3%	0.2%	0.0%	0.7%	66.7%	1.5%	33.3%

分析（続き）
GPT4 は、ドキュメント タイプごとのカバレッジと正確性、および集約されたカテゴリのカバレッジと正確性に関して、テストされた他のすべての LLM よりも明らかに優れています。Llama モデルはこの特定のテストでは良い結果を出せませんでしたが、これはこれらのモデルに関連付けられたコンテキスト サイズ (4k) に関係しているか、プロンプトで要求された出力形式で結果を生成できなかったためである可能性があります。

特定のドキュメント タイプ (領収書 - アクティビティ予約、キャンセル ポリシー、支払証明など) のパフォーマンスが低いと考えられる原因は、プロンプトに含まれるドキュメント タイプの説明がやや曖昧であるか、ドキュメント タイプが他のドキュメント タイプと非常に似ているためであると考えられます。追加のプロンプト エンジニアリングにより、目撃されたパフォーマンスの低下の一部を解決できる可能性があります。

コスト比較

私たちがテストした LLM には、基本的に 2 つの価格帯があります。GPT3.5 と Llama モデルは比較的安価ですが、GPT4 と Mistral Large は使用コストが高くなります。予想通り、私たちの分析では、全体的に、より高価な LLM の方がパフォーマンスが優れていることが示されています。しかし、GPT3.5 は安価であるにもかかわらず、高価なモデルに近いパフォーマンス レベルを実現しているのは興味深いことです。GPT3.5 と Llama モデルのコスト対パフォーマンスを分析すると、このパフォーマンスの不一致の鍵は、コンテキストのサイズにあると推測できます。GPT3.5 のコストは Llama モデルとほぼ同じですが、コンテキスト サイズが Llama モデルの 4 倍であるため、測定可能なパフォーマンス上の優位性が得られることがわかります。

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
llama-2-7b-chat	€0.63	€0.49	€ 301.00	4k
llama-2-13b-chat	€0.89	€0.77	€ 433.00	4k
llama-2-70b-chat	€1.67	€1.46	€ 814.00	4k
gpt-3.5-turbo-0125	€0.46	€1.37	€ 321.00	16k
gpt-4-0125-preview	€9.14	€27.41	€ 6,397.00	128k
mistral-large	€7.41	€22.22	€ 5,186.00	32k

結論

生成 AI と LLM の世界では、一概にすべてに適応するわけではないことを覚えておくことが重要です。まず、実行する必要がある作業を決定し、それを達成するために適切なツールを適用する必要があります。コストと比較したパフォーマンスを評価することも重要です。これまで見てきたように、この比較では GPT4 と Mistral Large が他の LLM を一貫して上回っており、GPT4 は分類タスクで非常に優れたパフォーマンスを発揮しています。

GPT3.5 のパフォーマンスは主要な LLM に迫っており、多くのユース ケースで十分なパフォーマンスを発揮する可能性があります。これは、価格を考慮すると特にその価値が見えます。

私たちがテストした Llama モデルは、特に GPT3.5 の価格とパフォーマンスと比較すると、分類シナリオでは競争力を持っていませんでした。

当社のデータ サイエンス チームは大規模言語モデルを継続的にテストしており、このレポートの今後のエディションでもその結果を報告し続けます。