SHIFT

The State of AI in Insurance (Vol. VIII)

推論（Reasoning）モデル

www.shift-technology.com/ja

エグゼクティブサマリー

- 「推論（reasoning）」とは、LLMが応答を返す前に内部で“思考”する能力を指します。
- 推論能力は、より複雑な業務に適用した場合にとくに有用です。
- これまでの知見と同様に、本調査でも推論モデルの性能は適用するユースケースに大きく依存することが示されました。
- 推論モデルの選択や「推論の深さ（reasoning effort）」の設定は、求められる性能、コスト、応答遅延（レイテンシ）などの要因によって決まります。

編集部より

本号の「The State of AI in Insurance」では、データサイエンスおよびリサーチチームが OpenAI と Anthropic が提供する推論モデルを詳細に検証しました。推論モデルは、LLMが応答を出す前に内部で“考える”ことを可能にし、理論的には複雑な業務での精度や効率を高めることを狙っています。

私たちは以前の調査（The State of AI in Insurance: Vol. VI — Claims Decisioning and Liability Determination in Subrogation）で、推論モデルは、単純なケースでも複雑なケースでも高い性能を示すことが確認されました。ただし、高性能であるがゆえにコストや応答遅延（レイテンシ）が増える傾向があり、そのため単純なデータ抽出や分類には必ずしも最適ではないという課題も明らかになりました。

本調査では推論モデルの性能検証を続ける一方で、今回新たに「推論の深さ（reasoning effort）」という概念に注目しました。これはモデル内部の“思考”の深さを調整できる機能で、OpenAI の初期 o-series 推論モデルでは選択可能でしたが、我々の従来検証では中（medium）設定のみで評価していました。本レポートでは、最小（minimal）・低（low）・中（medium）・高（high）の各推論深度設定が、主要ユースケースに与える影響を比較しています。

保険関連文書からの情報抽出・分類に関する LLM 推論モデル比較

方法論

データサイエンスおよびリサーチチームは、6つのテストシナリオを用い、公開されている推論モデル10種の性能を評価しました。うち4シナリオは抽出・分類タスクに分類され、残る2シナリオ（Claims Decisions と Motor Liability）は明確に推論タスクとして定義しています。

対象シナリオ（概要）

- 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）
- 旅行保険に関する英語文書の分類（複雑）
- 保険金請求の有無責判定（推論タスク）
- 自動車事故の損害賠償責任（Motor Liability：推論タスク）

保険金請求判定シナリオの定義 — モデルが以下の判断を行えるかどうかを検証する

- 申告は保険の有効期間内か？
- 請求書は申告内容と整合しているか？
- 申告内容に基づいて却下すべき理由はあるか？
- 補償や支払判断に必要な情報が欠けていないか？
- 保険金請求のステータス判定（例：保留、承認、却下）
- もし補償対象であれば、総額でいくら支払うべきか？

自動車事故の損害賠償責任シナリオの定義

- 保険金請求に含まれる情報から、代位求償（サブロゲーション）判断に必要な責任の所在を特定できるか？

評価指標

テスト対象の LLM について以下を評価しました

- Coverage（抽出対象の網羅性） — 期待される値に抽出対象が存在する場合、モデルはそれを抽出できたか。
- Accuracy（正確性） — 抽出した情報が正しかったか。

なお、プロンプト設計はすべて Shift のデータサイエンス／リサーチチームが担当し、各シナリオにつき1つのプロンプトを作成してすべてのテストモデルで共通して使用しました。すべてのプロンプトは GPT 系 LLM 向けに調整しているため、モデルごとの性能評価に影響を与える可能性がある点は留意が必要です。

結果

本レポートの評価表は、各ユースケースにおける相対的な性能差を示すよう色分けしています。それぞれのユースケースに対して、青系が高い相対性能、赤系が低い相対性能、白系が平均的な性能を示します。そのため、たとえば「90%」というスコアでも、同ユースケースにおける他モデルの性能と比べて最低値であれば赤で示される場合があります。これは、あるパーセンテージ自体はユースケースにとって許容できる値であっても、相対評価では劣ることを意味します。

Model	gpt-5-chat	gpt-5 high	gpt-5 medium	gpt-5 low	gpt-5 minimal	gpt-5-mini high	gpt-5-mini medium	gpt-5-mini low	gpt-5-mini minimal	gpt-5-nano high	gpt-5-nano medium	gpt-5-nano low
コスト($)	**$1,500**	**$3,500**	**$2,500**	**$2,000**	**$1,500**	**$700**	**$500**	**$400**	**$300**	**$140**	**$100**	**$80**
フランス語の歯科請求書	93.0%	94.6%	94.1%	94.3%	92.2%	95.0%	94.7%	94.2%	91.4%	93.5%	93.9%	93.2%
日本語の住宅修理見積書	85.0%	83.5%	77.8%	74.9%	71.8%	82.9%	83.0%	82.2%	82.1%	80.3%	80.8%	79.2%
英語の航空会社請求書	84.0%	82.2%	81.5%	82.0%	81.5%	82.0%	81.4%	81.1%	82.2%	71.7%	72.4%	68.4%
分類	91.4%	93.1%	92.6%	92.6%	90.2%	91.4%	91.4%	90.6%	87.7%	87.2%	86.9%	84.2%
代表的な簡易ユースケースのまとめ	**88.3%**	**88.4%**	**86.5%**	**86.0%**	**83.9%**	**87.8%**	**87.6%**	**87.0%**	**85.8%**	**83.2%**	**83.5%**	**81.3%**
フライト請求書の審査	86.9%	100.0%	100.0%	100.0%	85.1%	94.3%	93.8%	97.0%	79.8%	90.8%	90.5%	89.0%
自動車事故の損害賠償責任	80.2%	78.8%	79.9%	80.2%	79.1%	79.5%	78.8%	80.1%	76.7%	75.6%	74.4%	72.9%
推論を用いたユースケースのまとめ	**83.5%**	**89.4%**	**89.9%**	**90.1%**	**82.1%**	**86.9%**	**86.3%**	**88.5%**	**78.3%**	**83.2%**	**82.5%**	**81.0%**
全てのユースケースの集約	86.7%	88.7%	87.6%	87.3%	83.3%	87.5%	87.2%	87.5%	83.3%	83.2%	83.2%	81.2%

Model	gpt-5-nano minimal	gpt-4.1	gpt-4.1-mini	gpt-4.1-nano	o4-mini	claude41-opus	claude4-opus	claude4-sonnet
コスト($)	**$60**	**$1,600**	**$320**	**$80**	**$880**	**$13,500**	**$13,500**	**$2,700**
フランス語の歯科請求書	55.7%	91.0%	91.0%	90.9%	95.2%	93.5%	94.4%	93.9%
日本語の住宅修理見積書	73.5%	85.6%	81.9%	76.6%	82.2%	85.2%	84.5%	85.0%
英語の航空会社請求書	66.7%	85.4%	81.0%	68.0%	80.7%	85.6%	85.3%	82.6%
分類	65.2%	91.7%	88.7%	66.2%	90.3%	90.7%	91.1%	90.1%
代表的な簡易ユースケースのまとめ	**65.3%**	**88.4%**	**85.7%**	**75.4%**	**87.1%**	**88.8%**	**88.8%**	**87.9%**
フライト請求書の審査	33.9%	83.7%	82.7%	59.7%	96.8%	90.5%	86.8%	84.4%
自動車事故の損害賠償責任	52.1%	81.4%	76.0%	58.4%	79.6%	79.4%	79.1%	78.9%
推論を用いたユースケースのまとめ	**43.0%**	**82.5%**	**79.4%**	**59.1%**	**88.2%**	**85.0%**	**82.9%**	**81.7%**
全てのユースケースの集約	57.9%	86.5%	83.6%	70.0%	**87.5%**	**87.5%**	**86.9%**	**85.8%**

分析

GPT-5 ファミリー：本検証では、GPT-5 は推論タスクで優れた性能を示す一方、より単純あるいは構造化されたユースケースではやや劣後する傾向が見られました。興味深いことに、これは以前 o4-mini（純粋な推論モデル）を評価した際の振る舞いと類似しています。GPT-5 はときに「汎用モデル」と見なされますが、今回の検証ではしばしば推論モデル的な振る舞いを示しました。したがって、GPT-5 は複雑な推論ワークフロー向けに理想的ですが、単純な抽出や分類を行う場合はより適した代替モデルがあると考えられます。

GPT-5-mini の性能は、ほとんどのユースケースにおける魅力的なベースラインとして評価できます。出力品質は優れた汎用モデルと同等のレンジに位置しつつ、メインラインの GPT-5 よりもコスト効率が高い点が利点です。ただし、デフォルトでは推論深度（reasoning effort）を「low（低）」に設定することを推奨し、レイテンシやコストに余裕がある場合にのみ「medium（中）」を検討するとよいでしょう。「high（高）」の設定は通常は推奨されません。

コスト面で魅力的な GPT-5-nano については、構造化抽出や意思決定タスクにおいて明確なトレードオフ（品質低下）があるため、コスト最優先で品質要件が低い場合に限定しての採用を推奨します。

OpenAI の既存ライン（GPT-4.1 等）と o4-mini：我々の検証では、GPT-4.1 は単純タスクにおいて引き続き信頼できる選択肢であり、コストパフォーマンスにも優れています。また o4-mini は依然として強力な推論モデルの選択肢であると評価します。しかし、GPT-5-mini のバランスの良さを踏まえると、まずは GPT-5-mini（推論深度を low に設定）を評価対象とし、それでも要件を満たさない場合により重い推論モデルへ進む、という順序が現実的と考えます。

Anthropic（クロード）について：Claude 4.1 Opus は以前の Claude バージョンから改善が見られますが、今回のユースケースの組み合わせにおいては、そのプレミアム価格を正当化するほどの上乗せ効果は限定的でした。Anthropic ラインナップ内では Sonnet がコスト面で現実的な選択肢として残ります。

結論

Shift の研究は、LLM の性能およびどの LLM を活用するのが適切かは、その適用ユースケースに本質的に依存することを改めて示しています。LLM を採用する組織は、望ましい性能レベル、許容可能なレイテンシ、およびコストを十分に理解したうえで最適な選択を行う必要があります。一見すると「より大きいモデルがより良い」と考えがちですが、本調査の結果はその仮定が多くの場合において誤りであることを示しています。