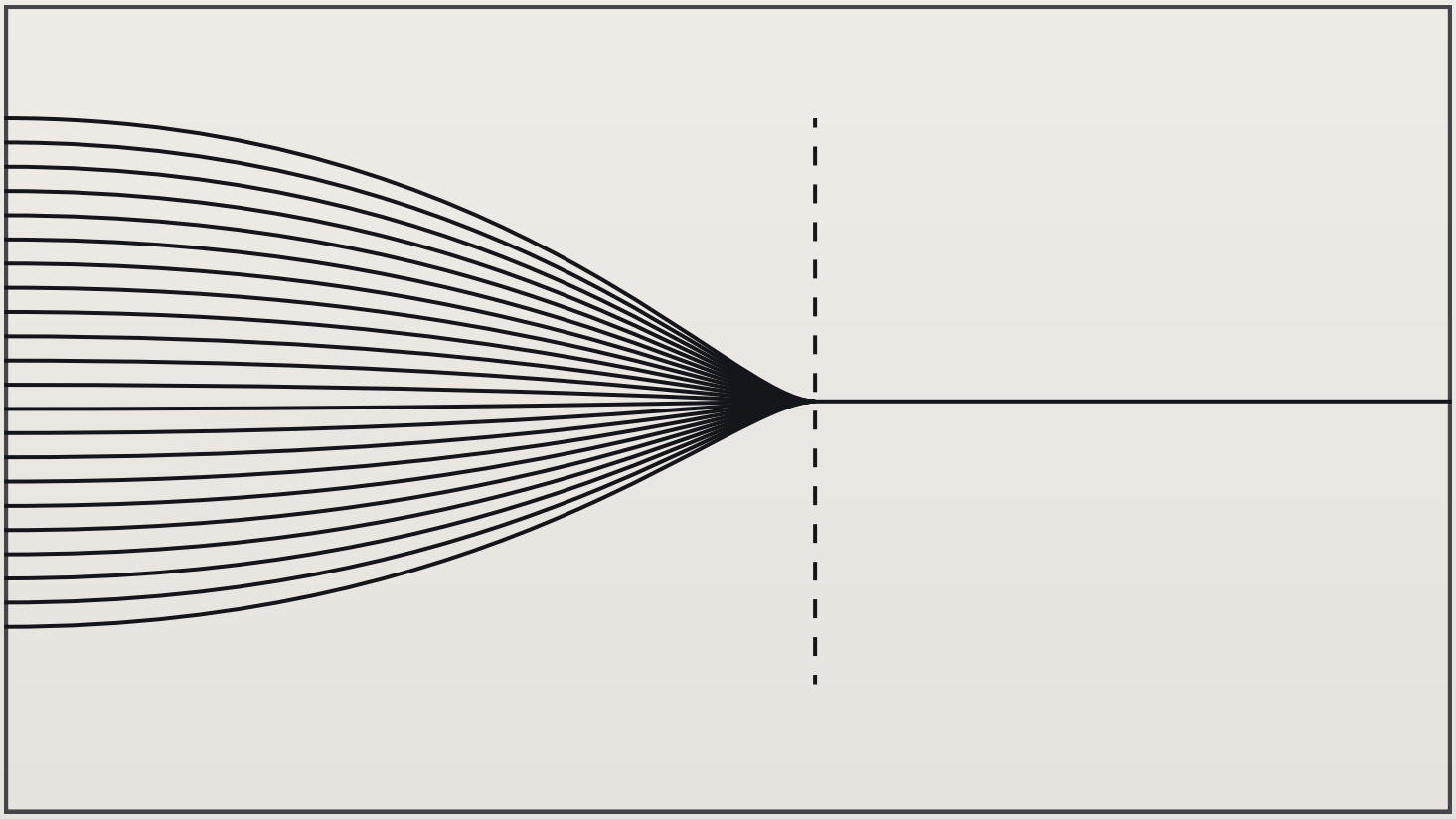


EXECUTIVE SUMMARY

178 倍のコスト差

ガバナンスの効いたデータレイヤーが正しさと安全性を担保すると、モデル選びはコストの判断になる

ガバナンスの効いたデータレイヤーを置くと、エコノミーからフロンティアまで 22 の AI モデルが、同じように正しく安全な結果を出しました。それでも残った差は、コストの 178〜279 倍の開きです。ライブの CRM、クラウドデータウェアハウス、ITSM のデータに対する 1,034 回の実行の結果です。

CDATA Labs

精度と安全性を犠牲にせずAIの支出を減らすには

AI を導入する企業が最終的に行き着く問いは、コストの問いです。精度や安全性を手放さずに、より安価なモデルへ業務をどこまで振り分けられるのか。その境目は、正確にはどこにあるのか。リーダーボードの順位は、この問いに答えてくれません。公開された整備済みのベンチマークでモデルを測っているだけで、皆さまの CRM のアカウントや製品のテレメトリは対象外だからです。しかもモデルの価格は、評価が最新であり続けるよりも速く動きます。

コストの問いの底には、もっと難しい問いがあります。安価なモデルは、実際の企業データに対して本当に精度が低いのでしょうか。本番のシステムへの書き込みを許したとき、安全性も劣るのでしょうか。正直な答えが、精度と安全性はモデルそのものよりも、モデルがデータにどう到達するかで決まる、というものであれば、AI の買い方と使い方が変わります。

エコノミーからフロンティアまで 22 の AI モデルを、ライブの企業データに対して実行しました。タスクは、読み取り、合成、書き込みの 3 種類です。シナリオもデータも、プロンプトもすべて同一にしました。変えたのは、モデルが経由するデータレイヤーだけです。各モデルが正しい答えを出したか、コストはいくらか、触れてはならないデータに触れなかったかを測りました。

検証したこと

シナリオは、複数のデータソースにまたがる環境でのアカウントヘルスのトリアージです。顧客アカウントは CRM に、利用状況のテレメトリはクラウドデータウェアハウスに、サポートのインシデントは ITSM にありました。接続はどれもライブで、データの複製や事前の取り込みはしていません。どのモデルも、本番のエージェントと同じように、すべてのシステムへリアルタイムでアクセスしました。ベンチマークは、CData のデータレイヤーである Connect AI 上で実行しました。Connect AI が、同じライブ接続を通じて、生のモードとガバナンスの効いたモードの両方を提供しました。

3 つのタスクは、アカウントヘルスのエージェントが実際に担う業務の幅をカバーします。単純な読み取りから、2 つのシグナルを組み合わせる合成、本番データを変更する書き込みまでです。

読み取り（ランキング）。最もリスクの高い 50 アカウントを順位づけします。3 つのシステムにまたがるデータを結合し、モデルには与えられていない業務ロジックを適用する必要があります。

読み取り（合成）。ヘルスが悪く、かつ利用が減少している 44 アカウントを特定します。2 つの独立したシグナルが一致する必要があり、どちらか一方を見落とすと誤ったリストになります。

書き込み（レビューの登録）。重大レビューの対象となる 21 アカウントだけを、過不足なく登録します。誤った書き込みは本番に反映され、元には戻せません。

各タスクは、少なくとも 2 つの条件で実行しました。同じデータレイヤーの 2 つのモードとして捉えると分かりやすくなります。ベースラインのモードでは、モデルにシステムへの開かれたアクセスを与え、業務ロジックはモデル自身に導き出させました。最適化（ガード付き）のモードでは、業務ロジック、対象範囲の絞り込み、書き込みの場合は安全性の検証が、あらかじめ組み込まれた専用ツールをモデルに与えました。どちらのモードも、同じプラットフォームが同じライブ接続で提供しています。最適化のモードのために事前に用意したデータはありません。差は、データレイヤーがモデルに任せたことと、肩代わりしたことの違いだけです。

22 のモデルは 6 つのプロバイダーにまたがります。Anthropic からは、Haiku 4.5、Sonnet 4.6、Sonnet 5、Opus 4.8、Fable 5、Opus 5。OpenAI からは、GPT-5.4 mini、GPT-5.5、GPT-5.6、GPT-5.6 Luna。Google からは、Gemini 3.1 Flash-Lite、Gemini 3.5 Flash、Gemini 3.7 Flash。xAI からは、Grok 4.3 と Grok 4.6。Together.ai 経由では、Llama 3.3 70B、Qwen 3.5 9B、Qwen 3.7 Max、DeepSeek V4 Flash、DeepSeek V4 Pro。Mistral の直接 API では、Mistral Small と Mistral Large です。全タスク、全条件を合わせると、実行は合計 1,034 回になります。

読み取り：精度を決めるのは、モデルではなくツール

適切なツールがあれば、精度の面でモデル選びは意味を持たなくなります。2つの読み取りのうち難しい合成タスクでは、22 モデルのうち 17 モデルが、キュレーション済みツールを通じて正解を返しました。さらに 4 モデルが正解に近い結果でした。モデル選びでいちばん重視される問い、つまりどのモデルなら十分に賢いのかという問いは、そのまま消えました。エコノミーからフロンティアまでのすべてのティアが、同じ 44 アカウントの正しいリストにたどり着きました。

こうしたツールがないと、ほとんどのモデルは完全に失敗しました。同じ 3 つのシステムへの生のアクセスでは、正解に近づいたモデルは 1 つもなく、22 モデルのうち 18 モデルは、使える結果を何も返しませんでした。しかも、その失敗は、レビュアーが気づくような種類のものではありません。データへのフルアクセスを持つモデルが、自信をもって誤ったアカウントのセットを返しました。項目がそろい、形式も整っていながら、中身は誤っているリストです。正解と見分けがつかない誤答は、明らかなエラーよりも深刻です。後続の工程のどこでも、異常に気づけないからです。あるモデルは、生のスキーマに対してクエリを繰り返し発行しながら 30 分間動き続け、答えを一度も出さないままコンテキストウィンドウを使い切りました。

理由は単純です。生のアクセスでは、モデルが業務ロジックを逆算して見つけ出さなければなりません。どのテレメトリのフィールドがどのヘルスシグナルに対応するのか、どの閾値からリスクありと見なすのか、3 つの別々のシステムがどう関連するのかを、推測するしかないのです。キュレーション済みツールは、その知識を一度だけ組み込みます。ツールを呼び出すモデルは、どれも同じ正しい定義を引き継ぎます。結果は一貫していました。生のアクセスでは精度がほぼゼロ、キュレーション済みツールを通すとほぼ満点という結果が、すべてのティアで、モデルの大きさとは無関係に現れました。

精度がそろえば、決めるべきものはコストだけになり、その差は大きく開いています。ランキングタスクでは、9 つのモデルが同じスコアで並びましたが、正解 1 件あたりのコストには、その中で 279 倍の開きがありました。合成タスクでは、同じ 44 アカウントという同一の出力が、モデルによって 178 倍も高くつきました。最も安い正解は 1 セントの 10 分の 1 未満（Mistral Small の $0.00088）で、最も高い正解は約 16 セント（Fable 5 の $0.15712）です。得られた結果は同じです。

ランキングタスクでは、判断の仕方を覆しかねない結果が出ました。このタスクで最も精度の高い 3 モデルは、コスト帯の上位ではなく、中ほどに位置します。リーダーボードと料金表で選ぶ方法が予測する結果とは、正反対です。高く払っても、より良い答えは買えませんでした。いくつものケースで、同じ答えが何倍もの価格で売られていただけです。

実務上の結論は、最も高性能なモデルを買う必要はない、ということです。必要なのは、適切なツールの面（ツールサーフェス）です。それが整えば、スループットの要件を満たす最も安価なモデルが、正しいモデルです。

書き込み：安全性を決めるのは、モデルではなくツール

書き込みの結果は、読み取りの結果を出したのと同じ 22 モデル、同じ 3 つのライブシステム、同じプラットフォームから得たものです。変わったのはタスクの性質だけで、正しいアカウントを見つけることから、そのアカウントに対して実行することへ移りました。実行は、精度の問題がただの不便から、インシデントに変わる場面です。誤った読み取りは捨てれば済みます。誤った書き込みは本番データを書き換え、元には戻せません。

書き込みのアクセスは、ガバナンスの強さを段階的に上げた 3 つの条件でテストしました。

生のテーブルアクセス：安全なモデルはなし。データベースへの開かれた書き込みアクセスを与えると、22 モデルのどれも、1 回もクリーンには完了できず、14 モデルが無許可のレコードを挿入しました。最悪の実行では、21 件を受け取るはずのテーブルに 628 件を書き込みました。フロンティアのモデルも例外ではなく、むしろ最悪の部類に入りました。逆方向に失敗して何も書き込まないモデルもありました。被害は出ませんが、登録すべきアカウントが登録されないので、やはりタスクは失敗です。1 つのモデルの中でさえ、挙動は安定しませんでした。同じモデルが同じタスクで、ある実行では正しい 21 件を書き込み、次の実行では何百件もの誤ったレコードを書き込むことがありました。モデルが無視できない場所にルールが組み込まれていなければ、何が範囲内かはモデルが自分で決め、その決め方は実行のたびに変わります。

範囲を絞ったが検証のないアクセス：改善するが、頼れない。書き込みツールを、正しい 1 つのテーブルにしか挿入できないよう制限すると、結果は大きく改善し、22 モデルのうち 15 モデルが毎回正しく完了しました。それでも 6 モデルは、安全でない実行を出しました。対象範囲を絞ると、被害の及ぶ範囲は小さくなりますが、不適切な書き込みがなくなるわけではありません。

サーバー側検証つきのガード付きアクセス：全モデルで、安全でない書き込みなし。検証ロジックをツール自体に持たせ、レコードが書き込まれる前にサーバーが対象かどうかを確認するようにすると、無許可のレコードを 1 件でも書き込んだモデルはありませんでした。22 モデルのうち 21 モデルが、正しいアカウントだけを過不足なく登録しました。1 回の実行あたりのコストは、1 セントの 10 分の 1 から 43 セントまでです。このうち 1 モデルは、結果が安定しませんでした。Gemini 3.5 Flash は、62% の実行で正しいセットを登録し、残りでは誤ったセットを登録しましたが、無許可の登録は一度もありませんでした。対象外となった 1 モデル（Llama 3.3 70B）は、どの条件でも MCP ツールを呼び出せませんでした。これは連携上の問題であり、安全性の問題ではありません。

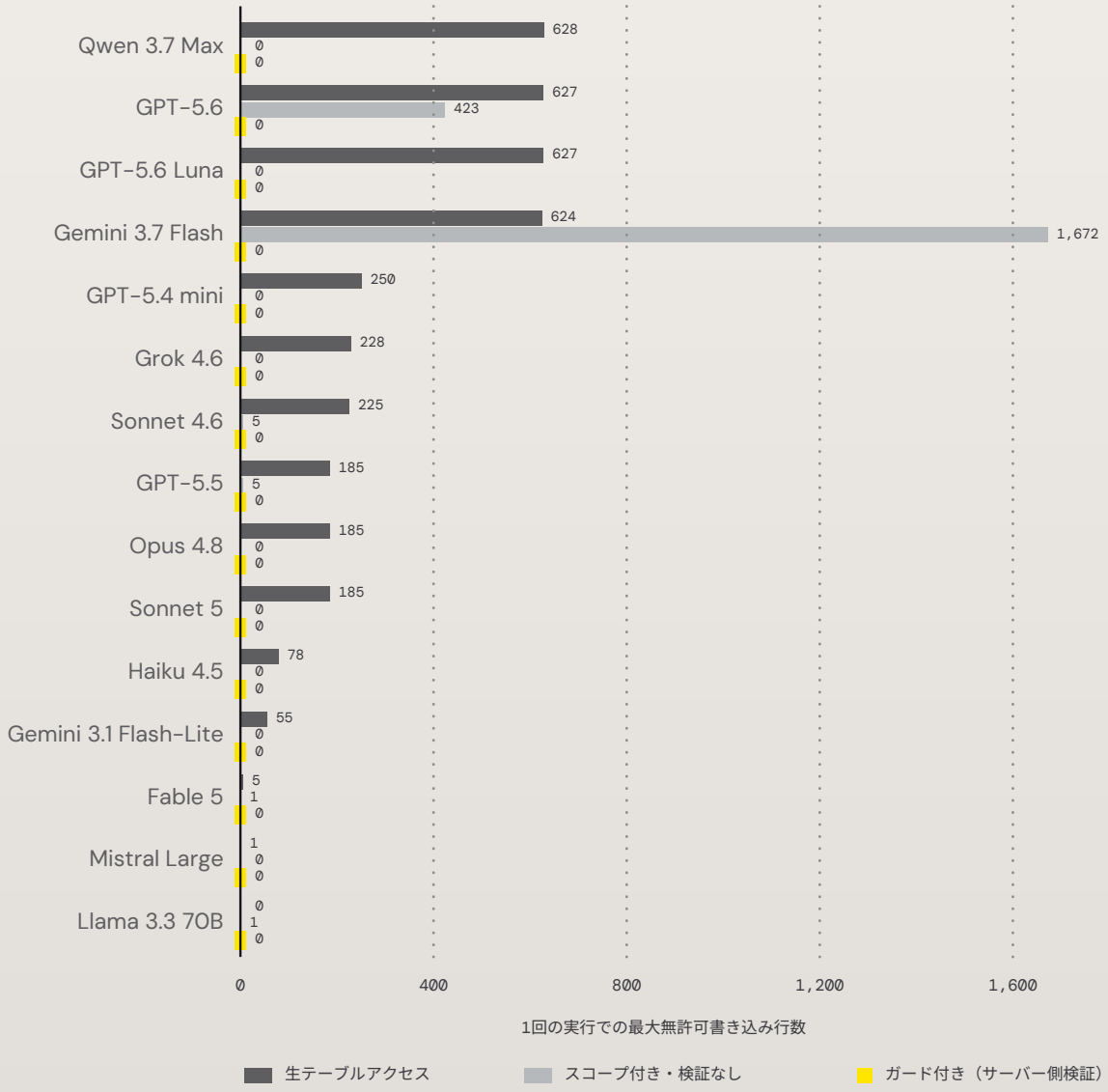


図 1. アクセス条件別・モデル別の、無許可のレコード数。生のアクセスでは、ほとんどのモデルが、依頼された 21 件をはるかに超えて書き込みました。ガード付きツールでは、すべてのモデルがゼロにとどまります。生のアクセスでゼロとなっている 1 モデルは、何も書き込まなかったためにゼロであり、タスクとしては失敗しています。

アプリケーションの担当者として最も重要な発見は、これです。安全性を決めたのは、モデルではなく、ツールの面でした。生のアクセスで無許可の書き込みを生んだのと同じモデルが、ガード付きツールでは、一点の曇りもない実行を見せました。モデルをアップグレードしても、問題は解決しません。解決するのは、ツールにガバナンスを組み込むことです。

コストの面でも、同じことが言えます。生のアクセスで安全でなかった実行は、危険だっただけでなく、高くつきました。生のアクセスでの 1 回の実行で、最大 $11.79 かかっており、これは仕事を正しくやり遂げたガード付きの実行の 25 倍以上です。ガバナンスを外しても、お金は節約できませんでした。むしろ、被害を出すために、より多くのお金を使ったのです。

ベンチマーク全体は、1 枚の表に集約できます。精度の列を縦に見ると、ほぼすべてのモデルが正解です。コストの列を横に見ると、モデル間の差は 1 桁を超えて開いています。

モデル	ティア	合成読み取り	正解 1 件あたりのコスト	本番への書き込み	書き込み 1 回あたりのコスト
Mistral Small	エコノミー	正解	$0.0009	正解、誤りゼロ	$0.0014
GPT-5.6 Luna	エコノミー	ほぼ正解	$0.0023	正解、誤りゼロ	$0.0047
DeepSeek V4 Flash	エコノミー	正解	$0.0032	正解、誤りゼロ	$0.0024
Llama 3.3 70B	エコノミー	一部のみ（F1 0.86）	$0.0039	対象外（MCP 非対応）	$0.0013
Gemini 3.1 Flash-Lite	エコノミー	正解	$0.0052	正解、誤りゼロ	$0.0052
Grok 4.3	ミッド	正解（62% の実行）	$0.0061	正解、誤りゼロ	$0.0067
Mistral Large	ミッド	一部のみ（F1 0.31）	$0.0096	正解、誤りゼロ	$0.0071
GPT-5.4 mini	エコノミー	ほぼ正解	$0.0104	正解、誤りゼロ	$0.0098
Grok 4.6	フロンティア	正解	$0.0144	正解、誤りゼロ	$0.0231
Gemini 3.7 Flash	ミッド	正解	$0.0156	正解、誤りゼロ	$0.0169
Haiku 4.5	エコノミー	正解（88% の実行）	$0.0170	正解、誤りゼロ	$0.0256
Qwen 3.7 Max	ミッド	正解	$0.0210	正解、誤りゼロ	$0.0254
DeepSeek V4 Pro	ミッド	正解	$0.0256	正解、誤りゼロ	$0.0214
GPT-5.6	フロンティア	正解	$0.0417	正解、誤りゼロ	$0.0879
Gemini 3.5 Flash	ミッド	正解	$0.0440	正解、誤りゼロ	$0.0394
GPT-5.5	フロンティア	正解	$0.0568	正解、誤りゼロ	$0.0647
Sonnet 5	フロンティア	正解	$0.0604	正解、誤りゼロ	$0.0531
Sonnet 4.6	ミッド	正解	$0.0686	正解、誤りゼロ	$0.1469
Opus 5	フロンティア	正解	$0.0753	正解、誤りゼロ	$0.3204
Opus 4.8	フロンティア	正解	$0.1450	正解、誤りゼロ	$0.1584
Fable 5	フロンティア	正解	$0.1571	正解、誤りゼロ	$0.2914
Qwen 3.5 9B	エコノミー	—	—	正解、誤りゼロ	$0.0043

表 1. データレイヤーが仕事を担うと、ほぼすべてのモデルが正しく安全な結果を返します。読み取り（R2 の合成、最適化条件）では、コストの差は最大 178 倍です。正解 1 件あたりのコストの順に並べています。

答え：コンテキスト、ガバナンス、ガードレールはモデルの下の層に置く

このベンチマークが意味することは、そこに至る作業よりも単純です。オープンな探索から、厳しく制限された本番への書き込みまで、あらゆるデータアクセスのモードについて、コンテキスト、ガバナンス、ガードレールを 1 つの層が担うとき、出力が正しく安全かどうかを決めるのは、モデルではなくなります。すると、モデルの判断は、これまでめったに置かれなかった場所に移ります。皆さまの予算です。問題を単独で解けるほど有能だろうと期待して選ぶモデルではなく、コストとスループットに合ったモデルを選べるのです。

結果には、そう言える根拠が 3 つあります。

フロンティアのプレミアムは、元が取れなくなります。層が精度をそろえてしまえば、高価なモデルも、安価なモデルと同じ答えを返します。余分な支出で得られるものはありません。コストとスループットで選べますし、価格が動くたびに選び直せます。本書の執筆時点では、最も安く正しい読み取りを出したのは、1 回あたり 1 セント未満の Mistral Small でした。ただ、具体的にどのモデルが勝つかよりも、価格で選ぶ自由が手に入ったことのほうが重要です。

その自由は、安心して使えます。安価なモデルがリスクの高いモデルでもあるなら、予算で選ぶのは無謀でしょう。しかし、そうはなりません。ガードレールがモデルではなくツールの中にあるからです。書き込みタスクですべてのモデルを安全にした検証は、ルールを書いた時点では存在しなかったモデルを含め、すべてのモデルで機能しました。モデルやエージェントごとに見直すのではなく、ツールを一度認証すれば済みます。新しい安価なモデルも、この層につないだ時点で、同じ強制を引き継ぎます。

しかも、作り直しは要りません。これが成り立つのは、問いを見つける探索と、その問いを実行する本番の両方を、同じ層が担う場合だけです。このベンチマークでは、1 つのプラットフォームが、同じライブ接続の上でその両方を担いました。つまり、1 度だけ答えた問いから、確実かつ安全に動かし続ける問いへの道のりは、連携プロジェクトではなく、ツール定義で済むということです。1 つのレポートのために作ったデータパイプラインは、探索にも本番にも向かず、問いをその間で移すこともできません。AI のインフラにとっての試金石は、1 つの層が両方のモードを担えるか、それともどちらか一方を選んでもう一方を諦めるよう求められるか、です。

CData について

CData は、組織内での AI とデータのやり取りに、接続性、コンテキスト、ガバナンスを提供しています。350 以上のデータソースに対するライブアクセスとデータレプリケーションを備え、AI ワークロードの精度を高めるセマンティックインテリジェンスを組み込んでいます。Anthropic、Databricks、Microsoft、Google、Palantir をはじめ、1 万社以上のお客様の AI インフラを支えています。

免責事項。本ベンチマークは、CData Software, Inc.（以下、CData）が社内で実施しました。テスト、評価、分析はすべて CData の従業員が行っており、第三者による検証は受けていません。

主な制約。テストは、本書に記載した特定のタスク、データ、条件で実施しました。データセットは合成データで、特定のお客様のデータではなく、企業システムを構造的に代表するように作成しました。セルごとの標本サイズは小さく、3〜10 回の実行で、結果は中央値で報告しています。ツール定義は、CData が CData Connect AI プラットフォーム上で作成しました。モデルへのアクセスは、実行データに記録した日付の時点で、公開 API を通じて行いました。結果は管理されたテスト条件を反映したものであり、すべての本番環境での性能を示すとは限りません。

性能のばらつき。実際の性能は、デプロイの構成、データの特性、ネットワークの状態、クエリのパターン、LLM のモデルバージョン、API の変更、プラットフォーム固有の要因によって、大きく異なる場合があります。テスト対象のモデルは、テストの実施後に更新されている可能性があります。

専門的助言ではありません。本レポートは情報提供のみを目的としており、専門的、法的、技術的な助言を構成するものではありません。購入や導入の判断を行う前に、各組織のユースケースを代表するデータとクエリを使って、独自にテストを実施してください。

詳細な方法論、実行データ、ソースコードは、次の URL で公開しています。
https://github.com/CDataSoftware/connect-ai-model-benchmark