GDPxベンチマーク徹底分析：フロンティアAIエージェントの実力と限界

GDPxは11領域・206タスクでAIエージェントの実業務能力を測定。首位モデルでも成功率は16.4%に留まり、
完全自律化の難しさとコスト対効果のトレードオフが明らかに。

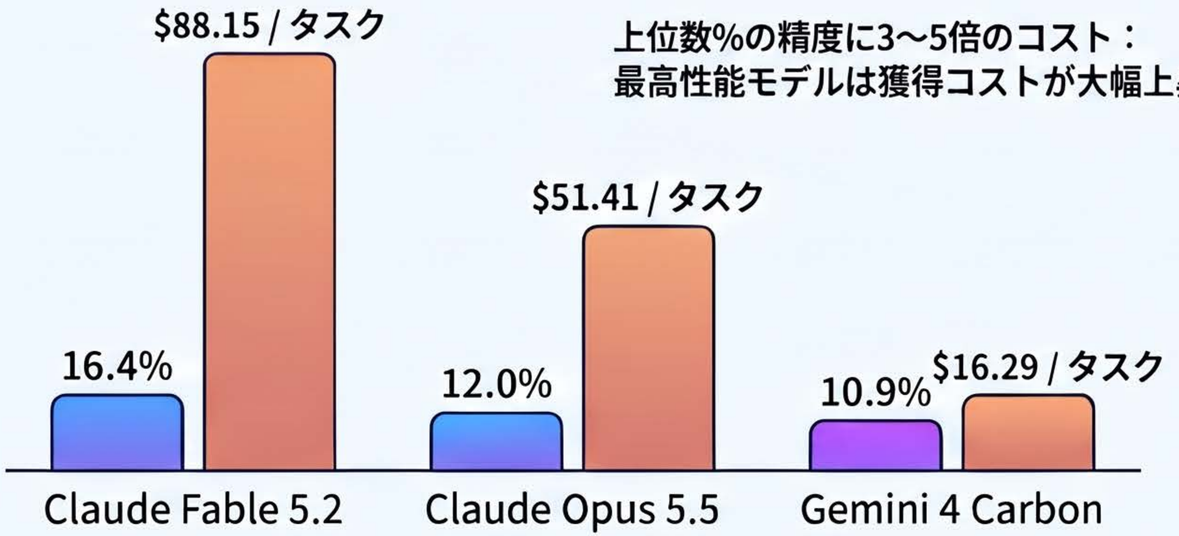
GDPxリーダーボードとコスト構造



最高達成率はわずか16.4%

2026年11月時点でClaude Fable 5.2が首位
だが実業務の1割強。

主要フロンティアモデル比較



上位数%の精度に3〜5倍のコスト：
最高性能モデルは獲得コストが大幅上昇。

領域別傾向と実用化への示唆

金融・法務：完全自律化不可



二値採点では全モデルが0点。人間による確認が不可欠。

安全フィルターによるスコア変動



サイバーセキュリティ　あたの敏感なほど、

特定領域では安全分類器の発動が評価に直結。

中位モデルと人間検証の併用



費用対効果に優れた中価格帯モデルをドラフト作成に活用するのが現実的。

Gemini Notebook