

Artificial Analysis Intelligence Index v4.2 : AI評価体系の劇的転換

実務アーキテクチャの刷新

非公開データセットの比重を40%へ倍増
テストへの過学習を排除し、未知のタスク
に対する真の実力を見極める。



4つの重点評価領域への再編

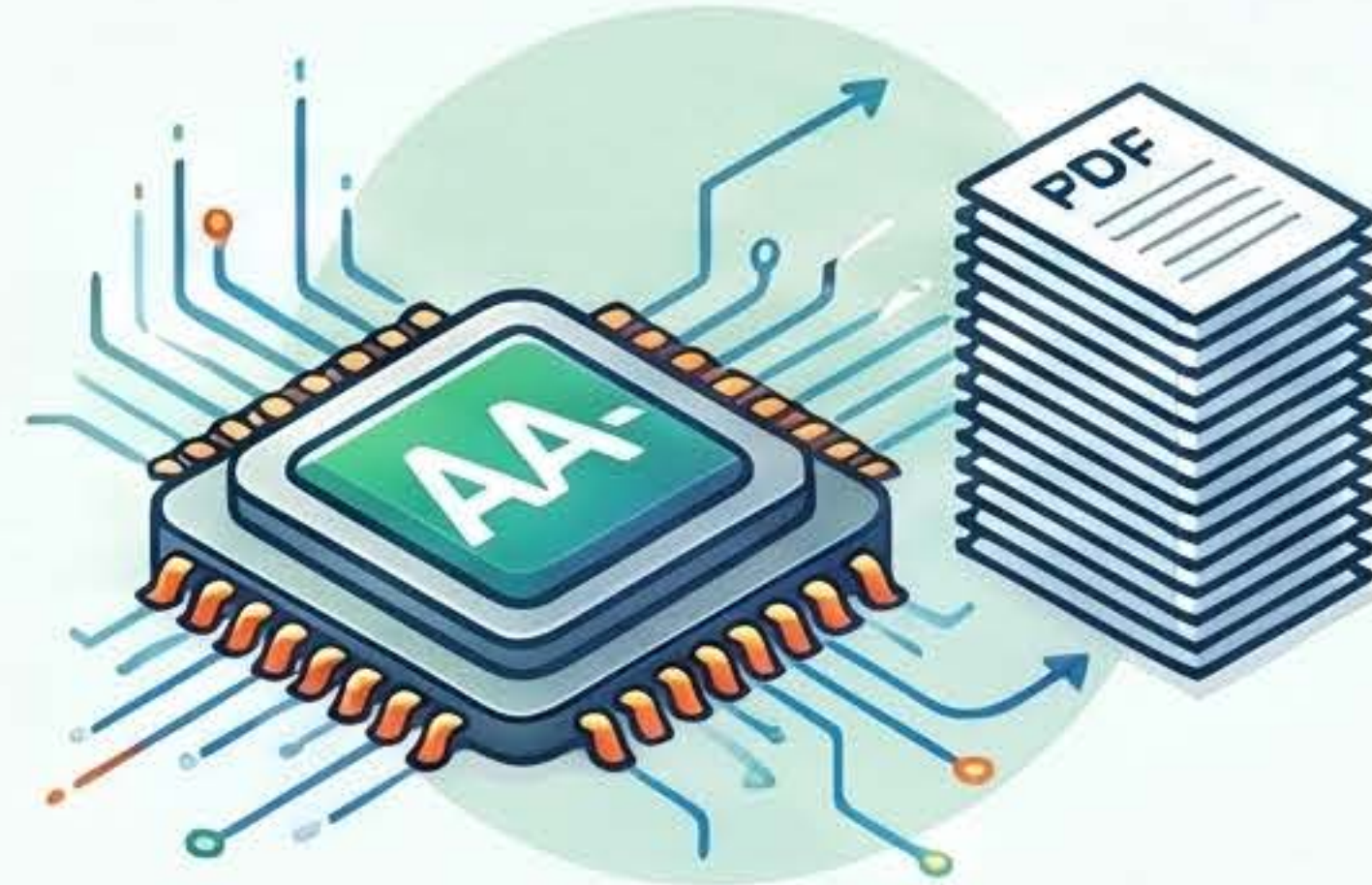


モデル別の得意分野と序列



実務を模した高難度 ベンチマークの導入

数週間に及ぶ実務を模した「
AA-Briefcase」などが評価の核となる。



GPT-6 Astraの卓越した トークン効率

他モデルより少ないトークン数で正解に
達し、コストと遅延を抑制する。

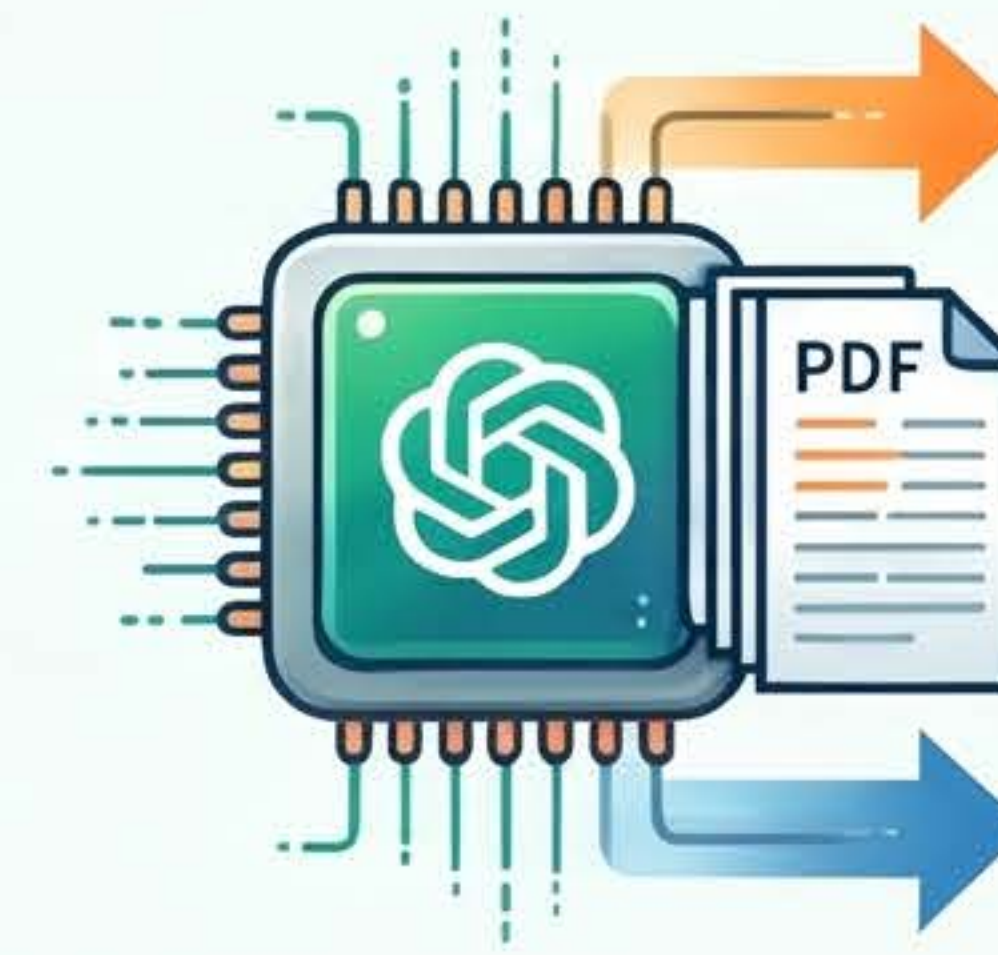
Claude Fable 5.1



総合首位Claudeと PDF解析の王者Astra

エージェント作業はClaude、
長大な文書解析はGPT-6
Astraが優れる。

GPT-6 Astra



GPT-6 Astraの卓越し たトークン効率

他モデルより少ないトークン
数で正解に達し、コストと
遅延を抑制する。

Gemini 3.8 Flash



Gemini 3.8 Flashが 示す高い費用対効果

低コストでありながら高い
知能指数を維持し、実務の
有力な選択肢に。

主要モデルの総合スコアと実務における適性の違い

	総合スコア	特徴的な強み
Claude Fable 5.1	57	長期的な文脈管理、自律エージェント作業
GPT-6 Astra	55	4,000枚超の長大PDF解析、推論効率
Gemini 3.8 Flash	59*	圧倒的なコストパフォーマンス

*Gemini 3.8 Flashのスコアは協定の条件下での測定値を含む場合があります。