

# AI評価の転換点：Artificial Analysis Intelligence Index v4.2 改定のコル心

2026年9月に実施された本改定は、単なるスコアの更新ではなく、評価の主眼を「学術的な問題解決」から「現実の業務遂行能力」へと大きく移したものです。非公開テストの比率を倍増させることで、ベンチマークへの過適合（カンニング）を防ぎ、より信頼性の高い実力判定を目指しています。



## 評価構造の抜本的刷新：学術から実務へ

### 1 実務遂行型ベンチマークの新規導入



複数週にわたる知識労働を根拠した「AA-Briefcase」と、4,000ページ超の専門文書を扱う「GDP.pdf」を統合しました。



### 2 非公開テスト比率を40%へ倍増

開発側による公開問題への通連合（Gaming）を抑制し、モデルの真の実力を測定します。



### 3 飽和した旧指標「GPQA Diamond」の除外

フロンティアモデルが満点に近付き、識別力がなくなった学術テストを評価から外しました。



## 改定がもたらす影響と利用者への注意点



### 1 評価ウェイトの再配分による順位の変動

汎用能力（General）とエージェント能力（Agents）の比重が変更され、実務性能の高いモデルが上位に浮上しました。



### 1 評価ウェイトの再配分することの変動



汎用能力の高業務遂行により実務能力の高いエージェント能力モデルと位浮化しました。

### 2 英語・テキスト中心の評価であることの認識



本指数は主に英語評価のため、日本市場では別途「Multilingual Index」との併用が不可欠です。

### 3 「知能」の定義が業務ROIへ接近



単なる正答率ではなく、分析品質やコスト、実行時間を含む「現場で役立つか」が重視されています。