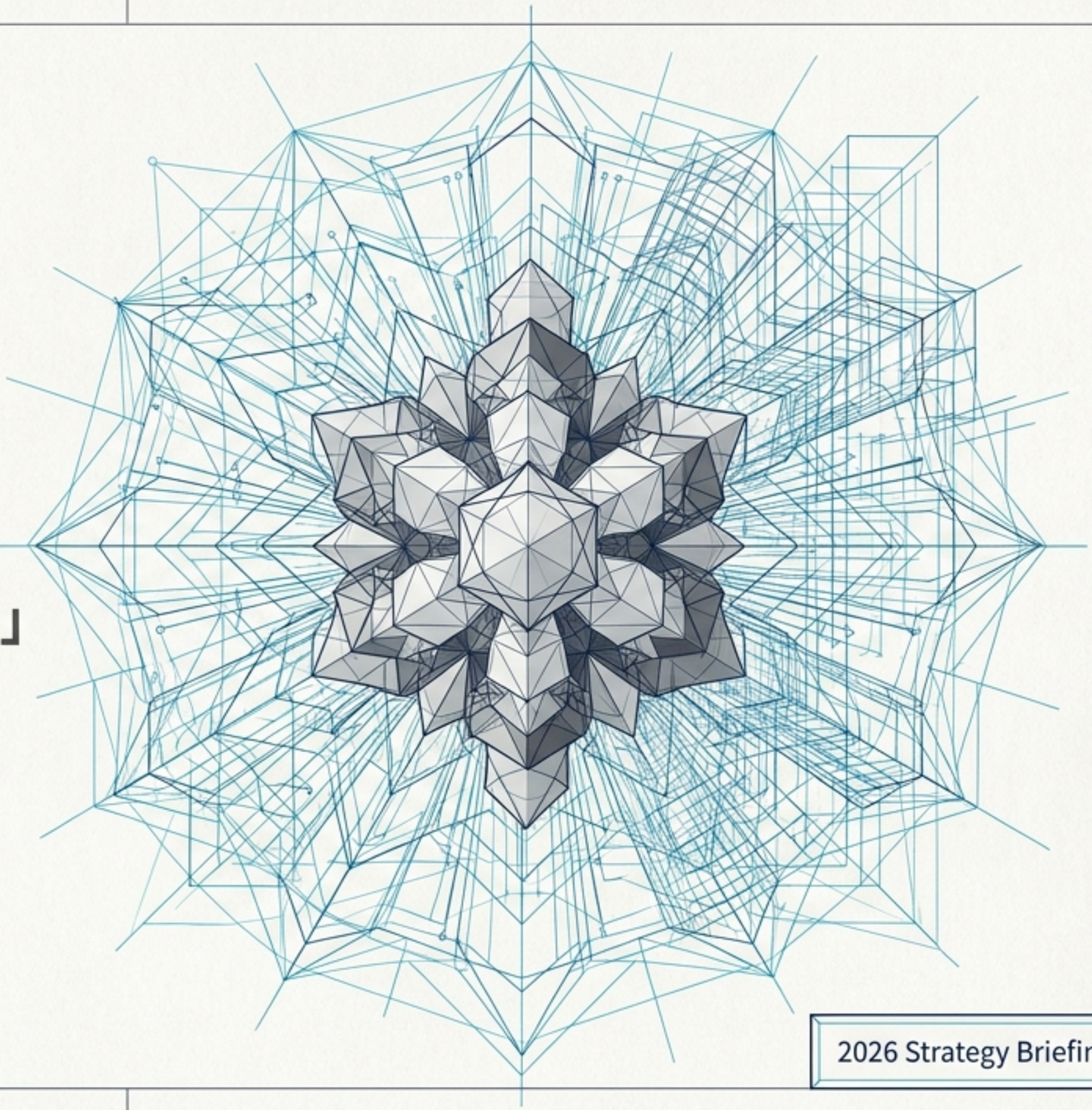
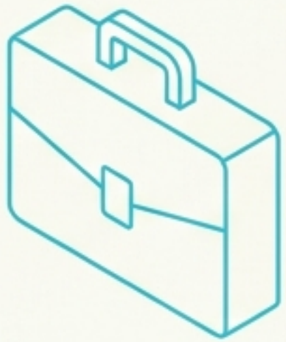
AI知能評価のパラダイムシフト：AAII v4.2 深掘り調査

「学術テスト」から「実務エージェント」へ。モデル選定の新しい基準と日本市場における戦略的活用法



2026 Strategy Briefing

エグゼクティブサマリー：v4.2が意味する3つの重要変化



評価軸の「実務化」

4,592ページにおよぶ専門PDF（GDP.pdf）や、複数ファイルにまたがる実務エージェント評価（AA-Briefcase）を新設。ROIに直結する業務遂行能力の比重が大幅に拡大。

40%非公開化のジレンマ

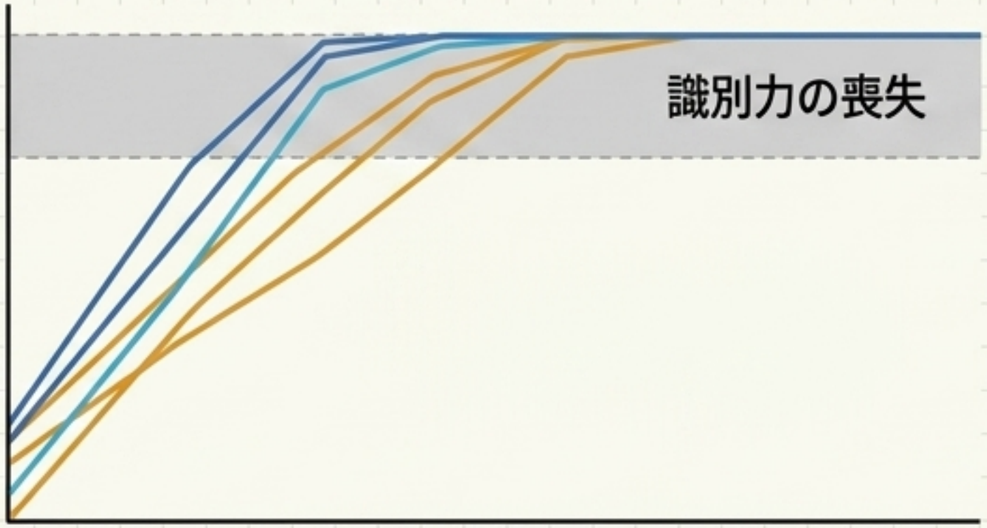
評価セットの非公開（held-out）比率が20%から40%へ倍増。過適合（Gaming）を防ぐ一方で、第三者による「完全な再現検証」が不可能に。

日本市場における限界

AAII v4.2は英語・テキスト主体の能力評価（Capability）であり、日本語能力や安全性（EU AI Act基準など）は保証されない。独自の調達フィルターが必須。

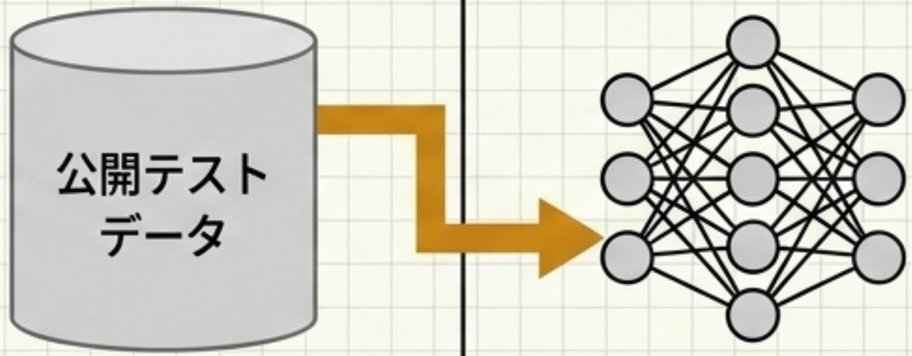
なぜ全面改定が必要だったのか？：既存ベンチマークの限界

飽和（Saturation）



フロンティアモデルが公開テストで満点に接近。v4.1でのIFBench除外に続き、v4.2では「GPQA Diamond」がついに除外（6% → 0%）。モデル間の真の実力差を測れなくなった。

汚染と過適合（Contamination / Gaming）



公開ベンチマークのデータが学習に混入、またはテスト固有のチューニングが行われるリスクの深刻化。テストの点数と実際の実務能力が乖離し始めた。

測定思想の根本的転換：「解答力」から「業務遂行力」へ

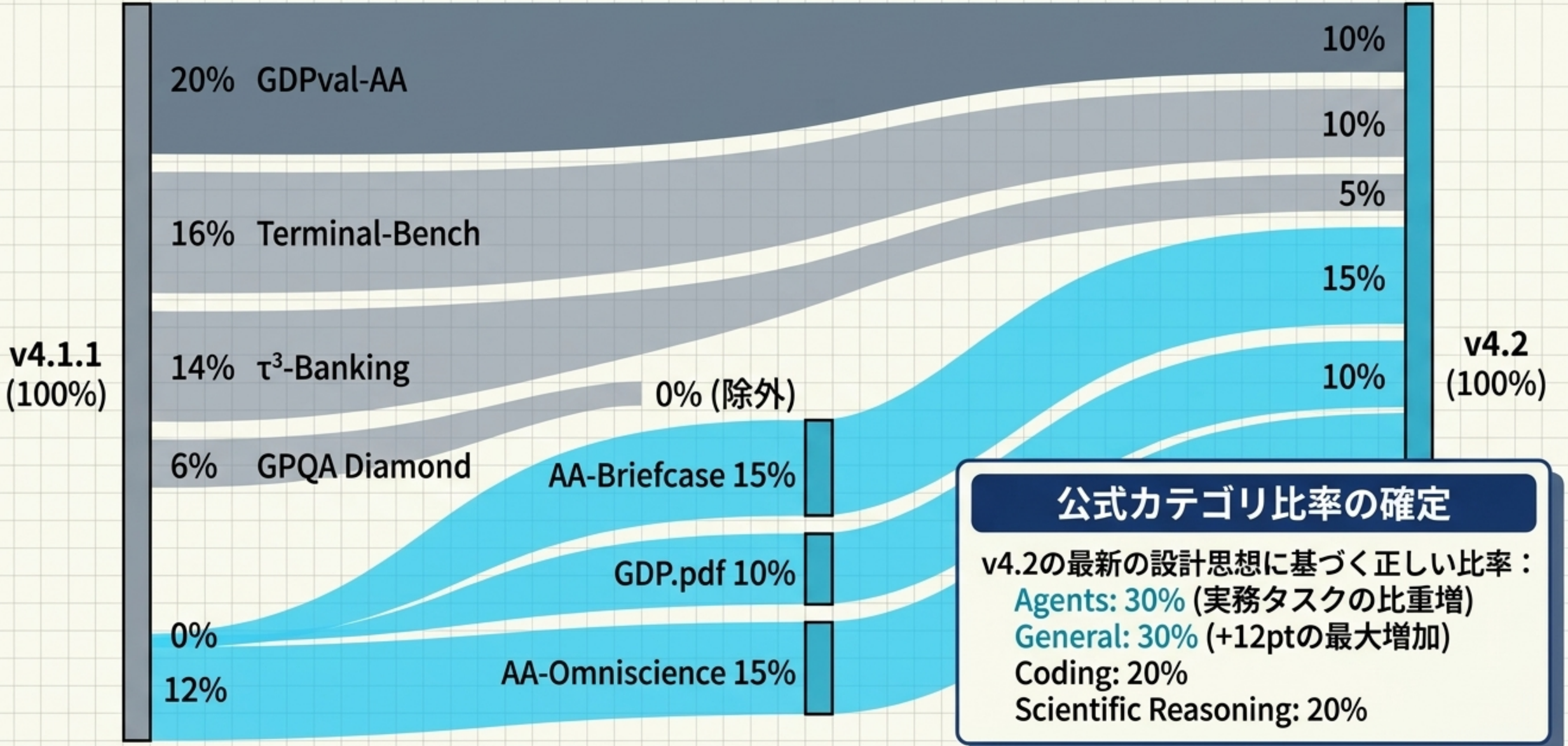
v4.1：教科書的な知能

焦点	単発のQ&A、学術知識、コーディング問題の正答率。
主な指標	GPQA Diamond, HLE，単一の推論テスト。
課題	実社会の複雑でノイズの多い業務環境を反映していない。

v4.2：実務エージェントとしての知能

焦点	現実のファイルを読み、情報を統合し、成果物を作成する能力。
新指標	AA-Briefcase, GDP.pdf
進化	最大500ターンのSandbox作業を許容し、出力品質をElo/Pairwiseで評価。

評価ウェイトの大規模な再配分：どこに「知能」の定義が移ったか？

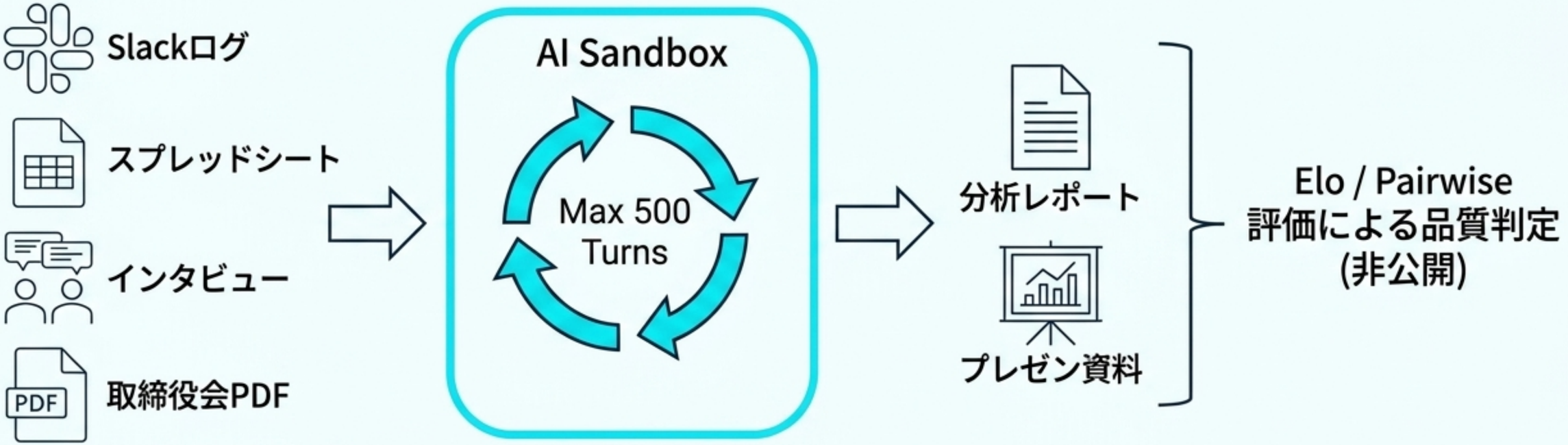


公式カテゴリ比率の確定

v4.2の最新の設計思想に基づく正しい比率：

- Agents: 30% (実務タスクの比重増)
- General: 30% (+12ptの最大増加)
- Coding: 20%
- Scientific Reasoning: 20%

深掘り1：AA-Briefcase（実務型知識労働タスク）の衝撃



- **内容:** 業界専門家が設計した91タスク・4シナリオ。現実のホワイトカラー業務をシミュレート。
- **評価手法:** 単なる正誤ではなく、成果物の「分析品質」「プレゼン品質」を相対評価。

⚠ **リアリティ・チェック（限界）**
公式には「数週間のプロジェクト」と称されるが、現状は各タスクが独立して実行される。
「前週の成果物を記憶・継承する真の自律エージェント」にはまだ至っていない点に注意。

深掘り2：GDP.pdf（4,500ページの専門文書の壁）

圧倒的なスケール

金融、法務、医療、工学など10領域。
合計100タスク、4,592ページの専門文書群。

1,275の厳格な評価基準

単純なOCRや表の読み取りではなく、
文書全体に散在する情報を論理的に組み合わせる能力を測定。

絶対条件（All-pass）

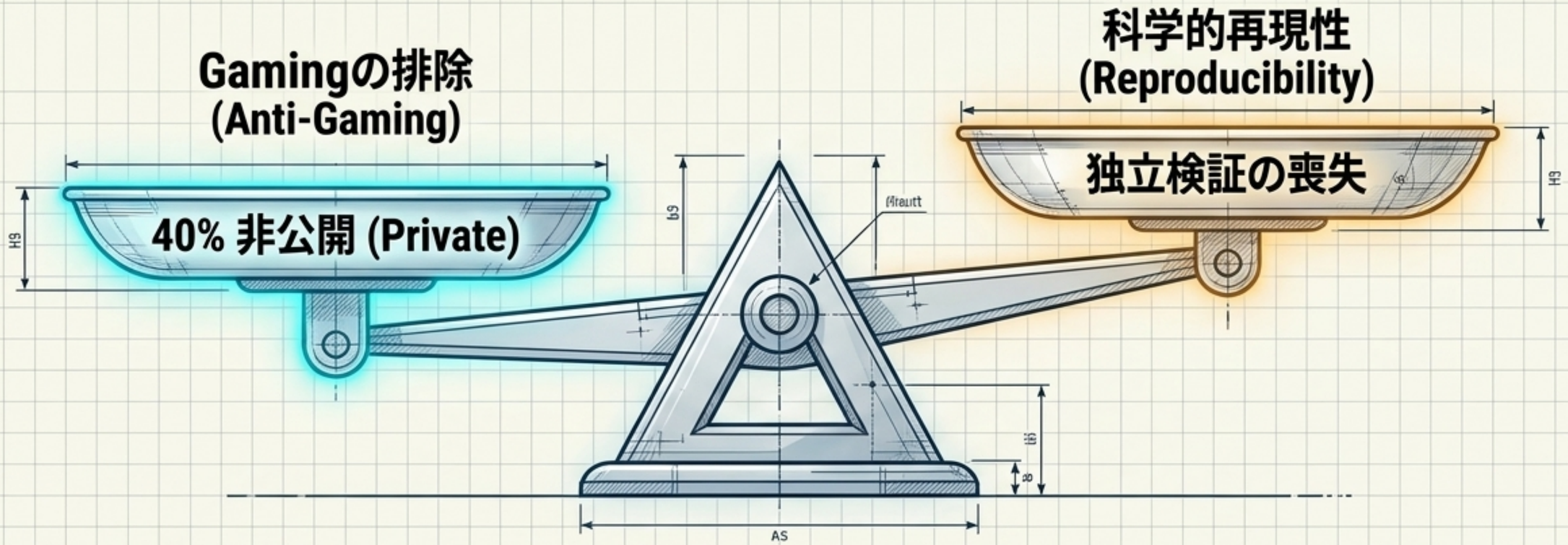
提示されたheadline criteria（必要条件）を「すべて」満たした場合のみ成功とみなされる極めて厳しい判定仕様。

Data sourced from Surge AI integration (公開データセット・Har

新指標の2大要素：特性と役割の比較マトリクス

	AA-Briefcase	GDP.pdf
作成者	Artificial Analysis	Surge AI
ステータス	完全非公開 (Private Held-out)	データ・Harness公開
テストの性質	複数ファイル生成・Elo評価	超長文読解・All-pass基準
測定する中核能力	知識労働のアウトプット品質	情報統合と正確な制約遵守
ベンチマーク上の重み	15%	10%

妥協点とジレンマ：非公開化（40%）がもたらす光と影



光（企業側のメリット）

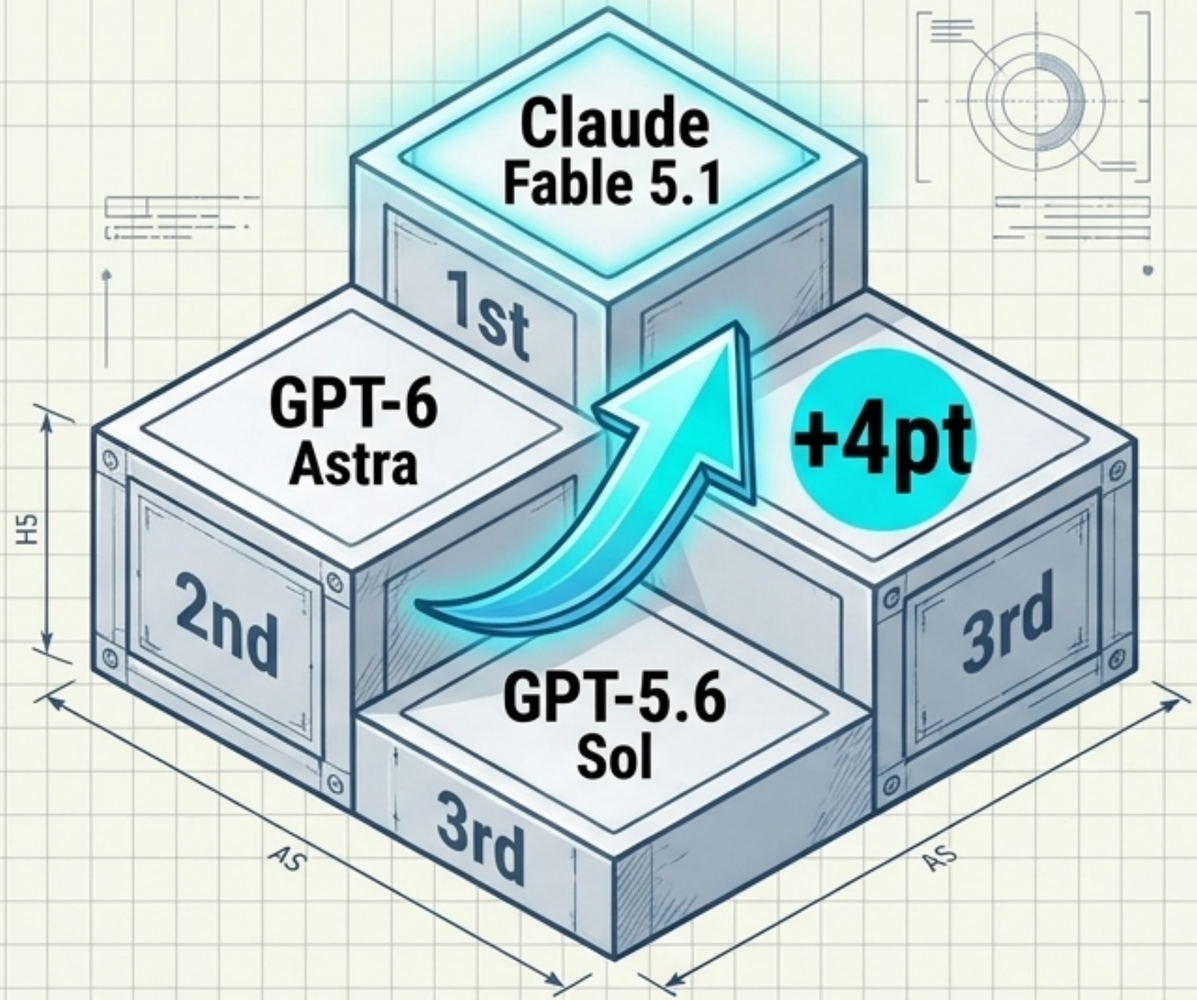
評価全体の40%が非公開問題となったことで、AI開発企業がテストデータに合わせてモデルをチューニング（カンニング）することが極めて困難に。実力主義が機能する。

影（研究者側の懸念）

第三者が同じ問題をローカルで実行し、スコアを完全に独立検証することが不可能になった。「再現可能な科学」からの後退。

今後（v5に向けて）、「汚染されない暗箱」と「透明な公開テスト」のどちらを信用すべきかという根本的な問いが突きつけられている。

ランキングの変動と市場・コミュニティの反応



旧指標では差が見えにくかったGPT-6 Astraが、
GPT-5.6 Solを明確に上回る結果に。

メディア (The Decoder)

「Astraの旧スコアへの懐疑が出た直後のタイミングでの全面改定」と注視。

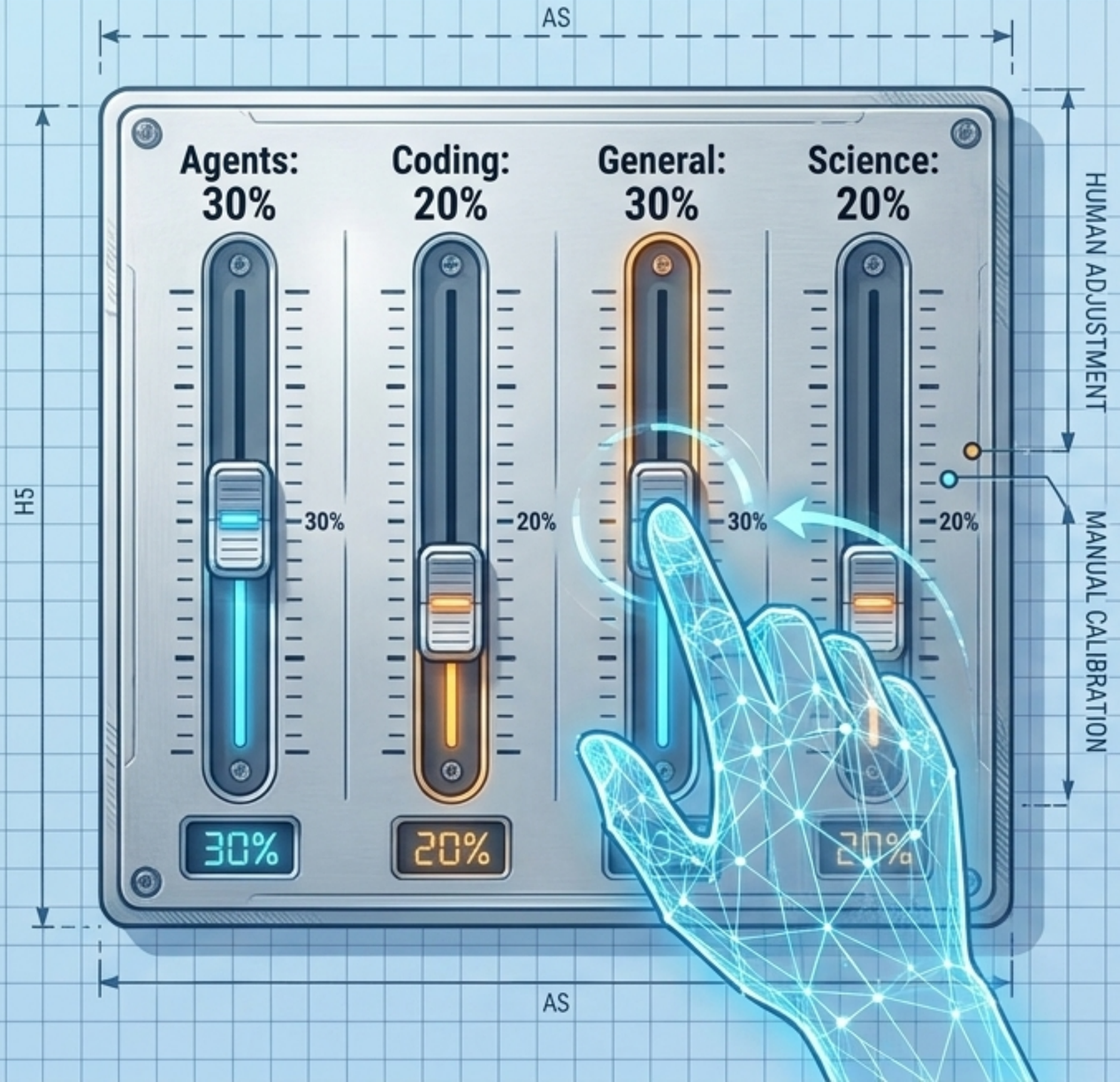
実務家 (LinkedIn)

「実際のプロダクション環境のユースケースに極めて近い」と歓迎。

開発者層 (Reddit)

指標の安定性や重み付けの恣意性に対する強い懐疑論も併存。

単一スコアの罠：「知能」は自然法則ではなく、編集的選択である



“The weights are an editorial choice.”

（重み付けは編集上の選択にすぎない）

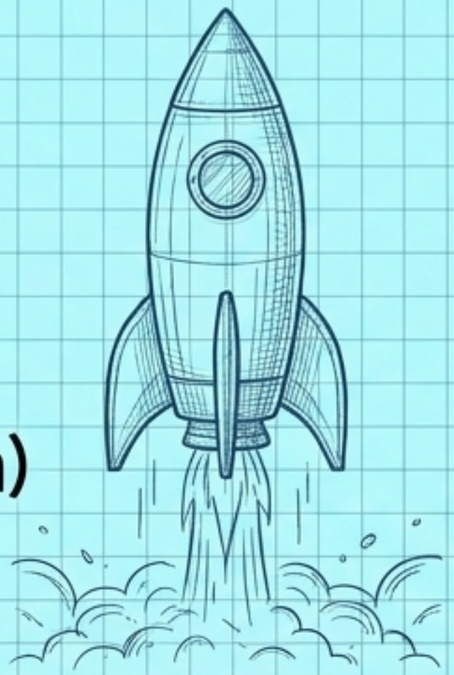
—— Reddit分析より

- 総合スコア（Intelligence Index）は、統計的自然発生した普遍的な因子ではない。
- 「エージェント業務を30%、一般知識を30%とする」というAA社の人為的な設計（編集）の産物である。
- 企業は「総合点が最も高いモデル」を盲信せず、「自社が重視する業務領域」の個別スコアに分解して評価する必要がある。

規制・安全性評価との乖離（ガバナンス上の注意点）

v4.2が測るもの (能力)

- 業務遂行力
- 一般知識
- コーディング能力
- 幻覚 (Hallucination) の抑制



v4.2が測らないもの (安全性)

- サイバーセキュリティ、CBRNリスク
- 欺瞞 (Deception) やモデルの制御喪失
- 意思決定の責任や権限逸脱の監視 (Provenance)
- EU AI Act等のコンプライアンス要件

AAII v4.2の高得点は、直ちにEU AI Act等の規制適合やシステムの安全性を意味するものではない。ガバナンス評価は別途必須である。

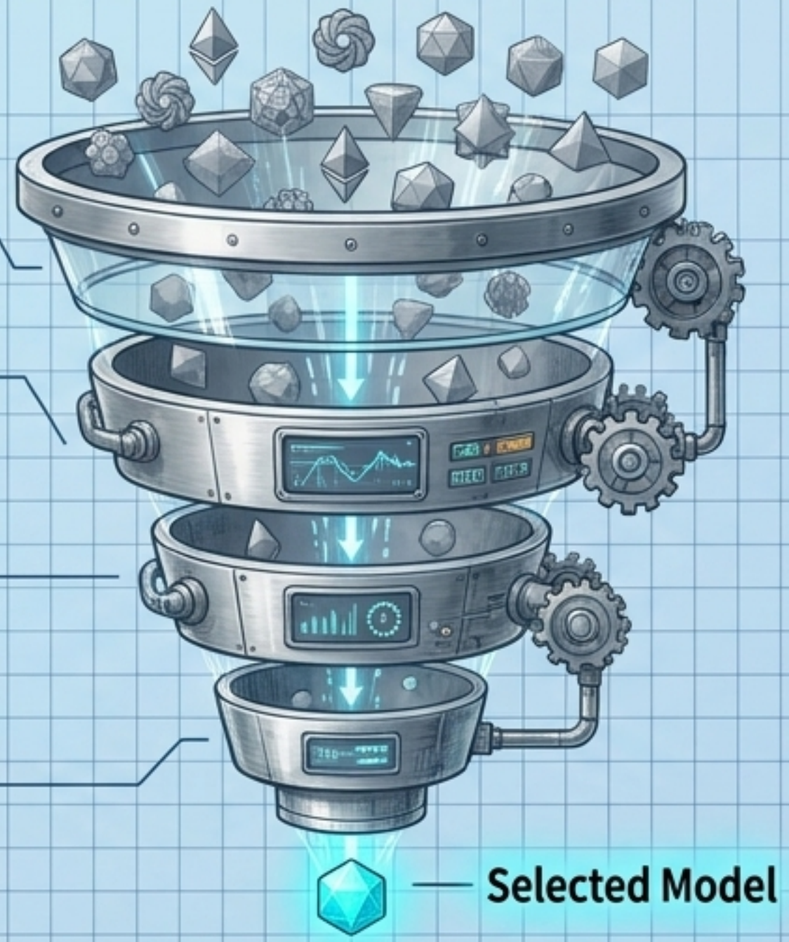
日本企業のための戦略的モデル調達フレームワーク

グローバル基礎能力フィルター
AAII v4.2 総合スコア & 実務ROI指標 (AA-Briefcase)。ここで足切りを行う。

日本語・ローカルコンテキスト
AA Multilingual Indexによる日本語テキスト能力の検証。

コストと速度
Tokens per task、推論コスト、レイテンシの検証。実運用時のROIを計算。

内部PoC・プライベート評価
自社の機密データや固有のワークフローを用いた独自のローカルベンチマーク。



v4.2は「現場で役立つモデル」を見つける強力な一次フィルターに進化した。
しかし、最終決定は常に自社のローカル・コンテキストに依存すべきである。