

AIの経済的価値 を再定義する

OpenAIの革新的ベンチマーク
「GDPval」が示す、ビジネスと社会の未来



従来のAI評価指標は、現実世界の価値を捉えきれていなかった

これまでのベンチマークは、AIの基礎能力（学術知識やコーディング能力）を測る上で重要でした。

しかし、その多くは実務の複雑な文脈や多様なタスク形式を反映しておらず、AIが「現実の仕事でどれだけ役立つか」を測るには不十分でした。

課題:

- 学術試験（MMLUなど）：実務的な文脈が欠落している。
- コーディング競技（CodeEvalなど）：人為的な問題設定が多く、実務の一部しか反映しない。
- 特定領域の課題（SWE-Benchなど）：評価範囲が狭く、汎用的な業務能力を測定できない。



現実の経済的価値？

新たな羅針盤「GDPval」：AIの”現実世界における経済的価値”を測定する

GDPval (Gross Domestic Product valuable tasks) は、AIモデルが現実の経済活動にどれだけ貢献できるかを定量化するために開発されました。

名称は、GDPへの寄与度が高い主要産業の職務（タスク）を評価対象としたことに由来します。

目的: AIの進歩を、学術的な能力だけでなく、実証的な経済貢献度で追跡する。

目的: AIの進歩を、学術的な能力だけでなく、実証的な経済貢献度で追跡する。



GDPvalは、現実の業務を忠実に再現した 厳格な評価手法で設計されている



収集 (Collect)

米国GDPに大きく寄与する
9業種、44職種の専門的タ
スク1,320件を収集。



設計 (Design)

分野経験平均14年以上の専
門家チームが、法律意見書
や設計図など実際の業務成
果物に基づきタスクを作
成。



実行 (Execute)

単純なテキストだけでなく、
関連資料や背景コンテキスト
を提供。成果物も文書、ス
ライド、図表など多岐にわ
たるマルチモーダル形式。



評価 (Evaluate)

業界専門家が、AIと人間の
成果物をブラインド方式で
品質比較。「AIの成果物が
人間以上（勝ち）または同
等（引き分け）」と判断さ
れた割合をスコア化。

GDPvalは、現実性・多様性・網羅性において既存の評価基準を凌駕する

GDPvalは、実在する業務成果物そのものの、またはそれを模したタスクを重視し、AIの実用的な応用力を試します。

評価ベンチマーク	対象領域・タスク形式	評価の焦点	現実世界との乖離
MMLU	数十分野の四択試験問題	学術知識のカバー範囲	高
CodeEval等	プログラミング問題	コーディング能力	中
SWE-Lancer	実契約に基づく開発タスク	市場環境下での実装力	低
GDPval	44職種の実務タスク	知識労働における成果物の品質	極めて低

注釈: GDPvalは一部マルチメディアも含むマルチモーダル評価であり、現実の仕事により近い形式です。

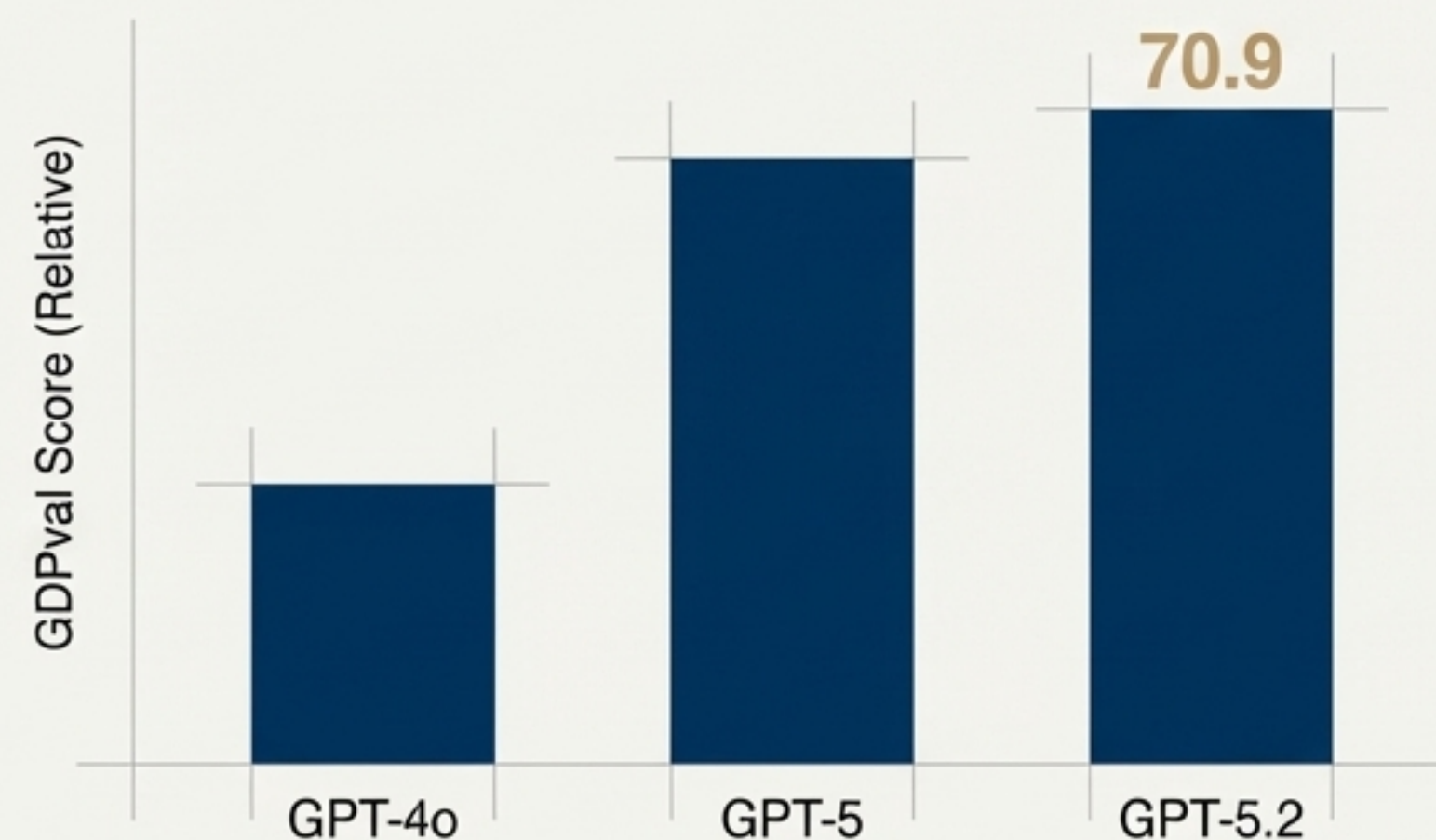
70.9%

OpenAIの最新モデルGPT-5.2が、専門家レベルのタスクにおいて人間と同等以上（勝ち/引き分け）と評価された割合。

最新AIは、専門的タスクの70%以上で人間と同等以上の成果を達成した

この結果は、AIが多くの知識労働タスクにおいて、実用的なレベルに到達しつつあることを示唆しています。

モデル間の性能も飛躍的に向上しており、GPT-4o（2024年春）からGPT-5（2025年夏）にかけて、GDPvalスコアは2倍以上に向上しました。



AIはタスク遂行において、人間を圧倒する 速度とコスト効率を理論上実現する

理論上の最大値

100倍 高速

人間の専門家と比較したタスク遂行速度。

1/100 のコスト

API利用料金に基づく試算。

注意点: この試算には、人間による監督や修正といった現実の業務で不可欠なプロセスは含まれていません。

より現実的なワークフローでの試算

AIによるドラフト作成を専門家がレビュー・改善する協業モデルでは、作業全体が...

約1.4倍 高速

...コストも...

約1.6倍 低減

...したケースが報告されています。

GDPvalは、AIが参入しうる巨大な経済市場の「青写真」である

GDPvalの対象となる44職種は、各業界で賃金総額への貢献が最大級の職種群です。

GDPvalのタスクは、人間専門家が平均7～9時間かけ約400ドルの価値を生み出す仕事に相当します。

戦略的インプリケーション:
GDPvalで高得点を出す職種は、AIによる代替・補完が進みやすい巨大市場であり、企業は業務プロセスの再設計や人材育成方針の見直しを迫られます。

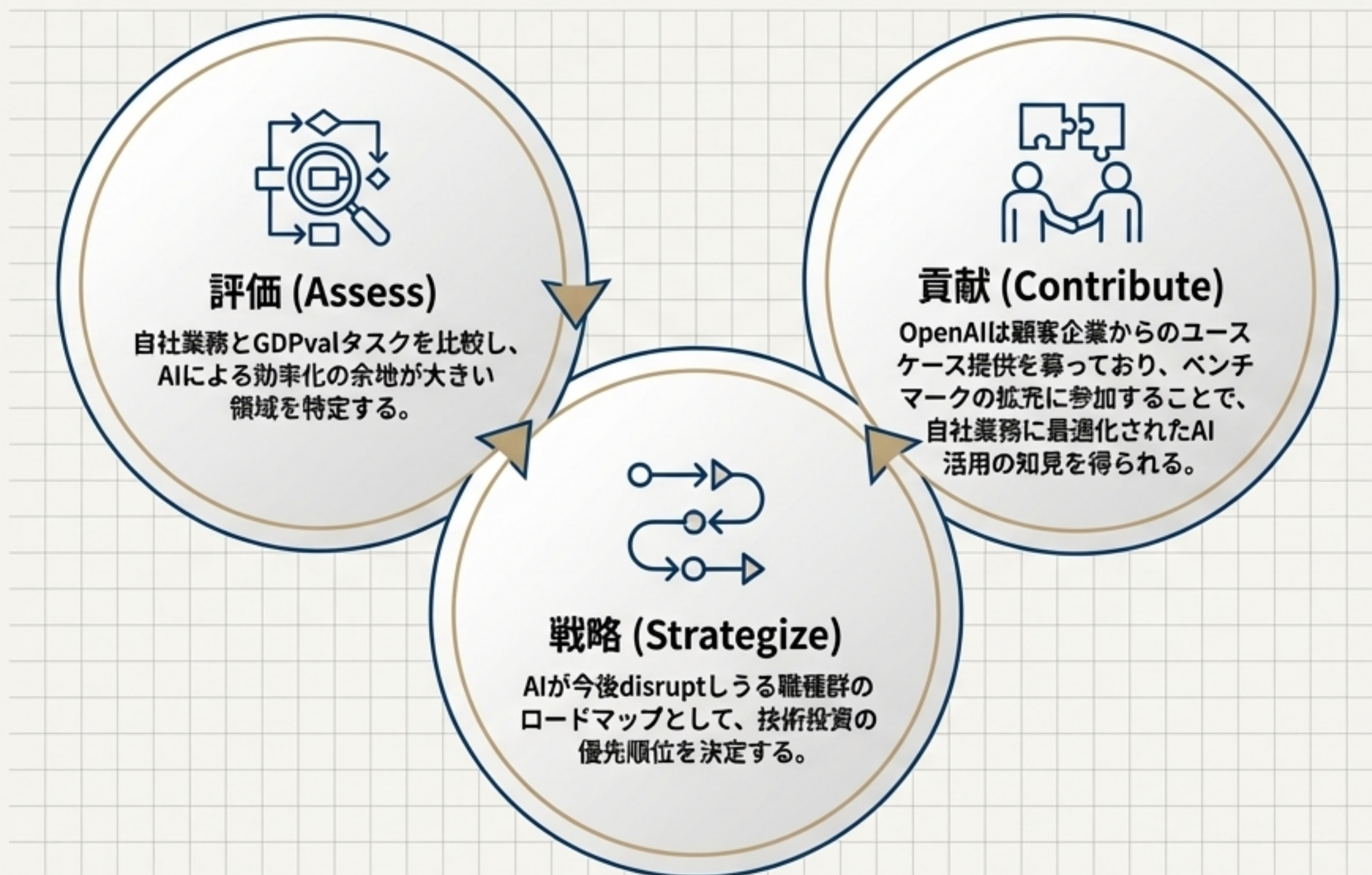


事例:
プロジェクト管理
スペシャリスト

この1職種だけで、
米国経済に年間
約1,087億ドル
の給与貢献があると
されています。

企業はGDPvalを、自社のAI導入戦略を測るベンチマークとして活用できる

企業における活用法



仕事は「AIへの代替」ではなく「役割の再構成」へ向かう

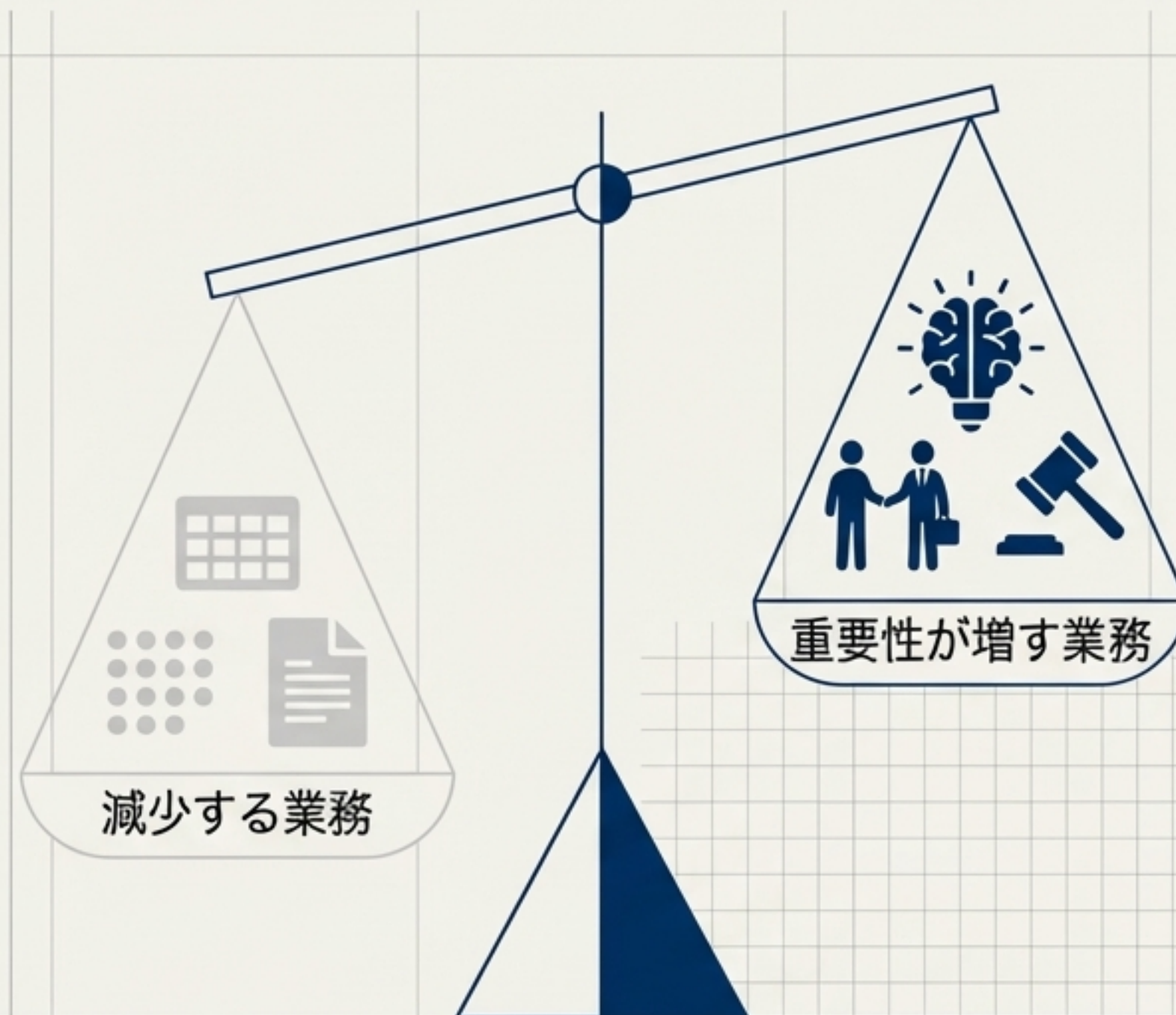
GDPvalの結果は、AIが定型業務を代替する可能性を示していますが、「職業そのものがAIに置き換わる」ことを意味するわけではありません。

人間の役割の変化:

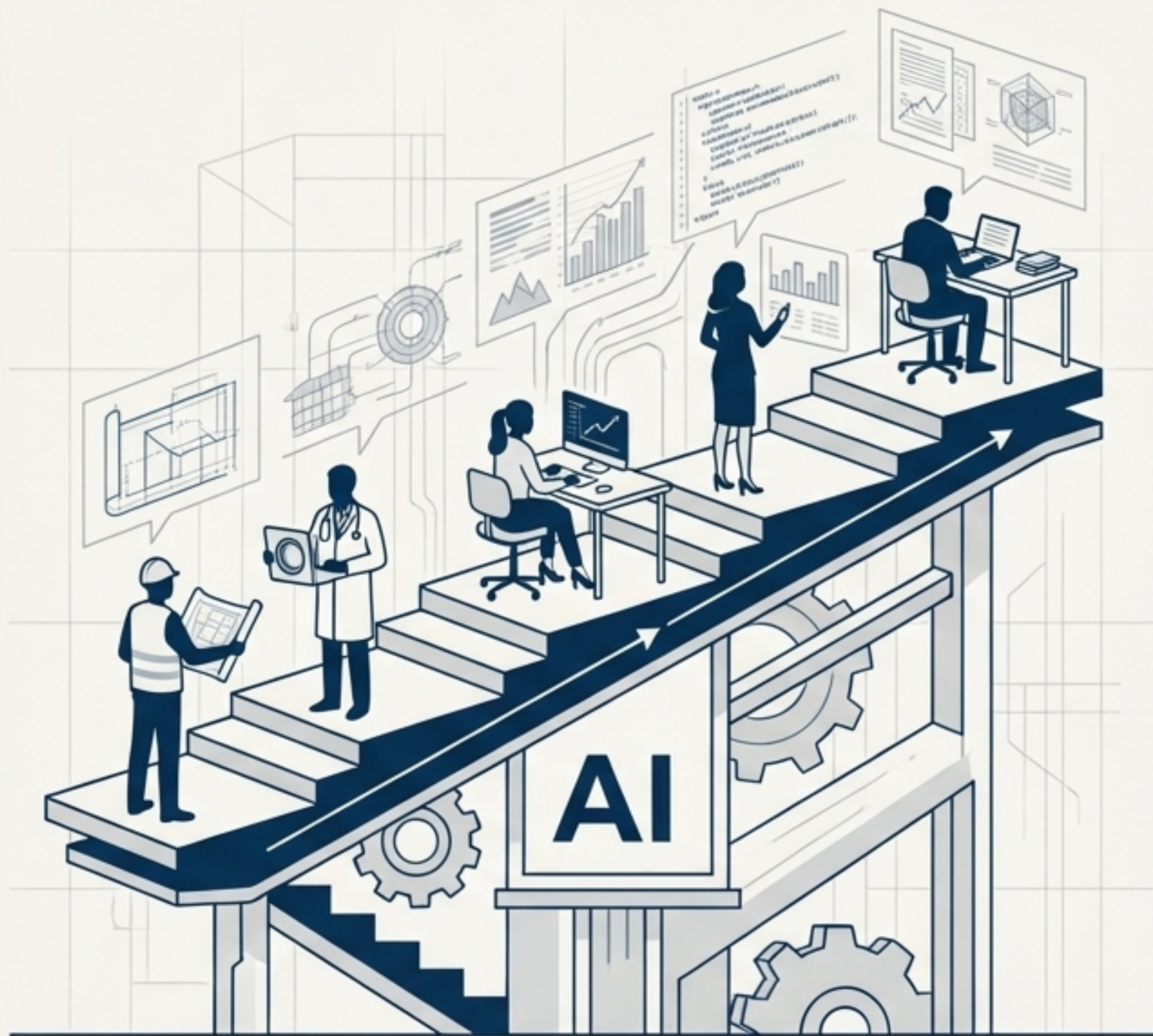
- **減少する業務:** AIが得意な反復的・明確に定義されたタスク。
- **重要性が増す業務:** 創造性、倫理的判断、対人コミュニケーション、AI成果物の高次な監督。

政策的示唆:

- **教育:** 創造的思考力やAIとの協働スキルを育むカリキュラムへの転換が必要。
- **労働政策:** 移行期における労働者への再研修プログラムやセーフティネットの整備が重要になる。



目指すは、AIを人間の能力を拡張する”上りのエレベーター”とすること



OpenAIのビジョン:

- 「我々の目標は全ての人をAIという“上りのエレベーター”に乗せることだ」
- AIを人間の代替ではなく拡張器具として位置づけ、その恩恵を全員に行き渡らせる。

29% 「失敗」

人間の監督は依然として不可欠:

- GDPval評価において、AIの出力の約**29%**は「欠陥が大きい失敗」と判断されました。
- これは、AIの成果物を実用に供するには、人間のレビューと改善が不可欠であることを示唆しています。AIの出力に対する最終的な責任は人間が負う必要があります。

万能ではないGDPval：その解釈には限界とバイアスの認識が不可欠



1. タスク設定の現実との乖離

タスクが「綺麗に定義され過ぎている」。現実の業務にある曖昧な要求や試行錯誤のプロセスは評価範囲外（one-shot型テスト）。



2. 限定的なカバー範囲

44職種に限定されており、肉体労働や対人サービス職は対象外。米国経済がベースであり、他国への適用には注意が必要。



3. 価値観の偏り

経済的価値（GDP寄与）に焦点が当てられており、医療や芸術など、GDPに直接現れない社会的価値は評価の射程外。



4. 信頼性と安全性

約3割のタスクは失敗と判定されており、AIの出力を無批判に信用するリスクは依然として存在する。

技術の進歩は、社会のルール作りとガバナンスの進化を要請する

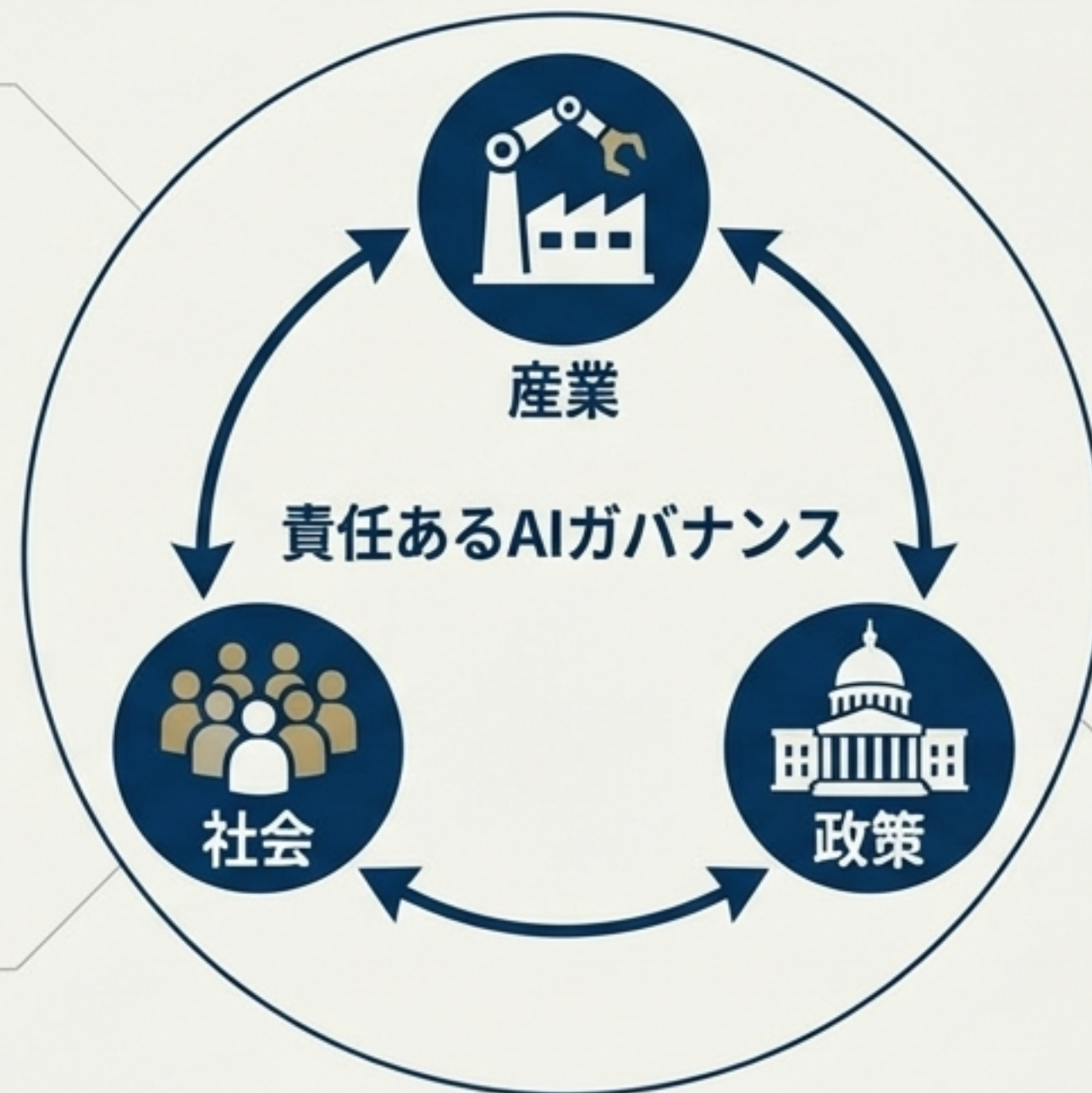
社会が向き合うべき課題

- 雇用の安定と格差

AIによる自動化で生じる失業や所得格差拡大に対し、富の再分配やセーフティネットの議論が不可欠。

- 倫理と人間の尊厳

経済効率を追求するあまり、人間的なケアや安心感が損なわれないか。定量的指標と定性的価値のバランスが重要。



- エビデンスに基づく政策決定

政府はGDPvalのような客観的指標を「年次AI白書」などに活用し、教育投資や雇用対策の優先順位をデータに基づいて決定できる。

GDPvalは進化を続け、次世代AI開発のロードマップとなる

今後の方向性

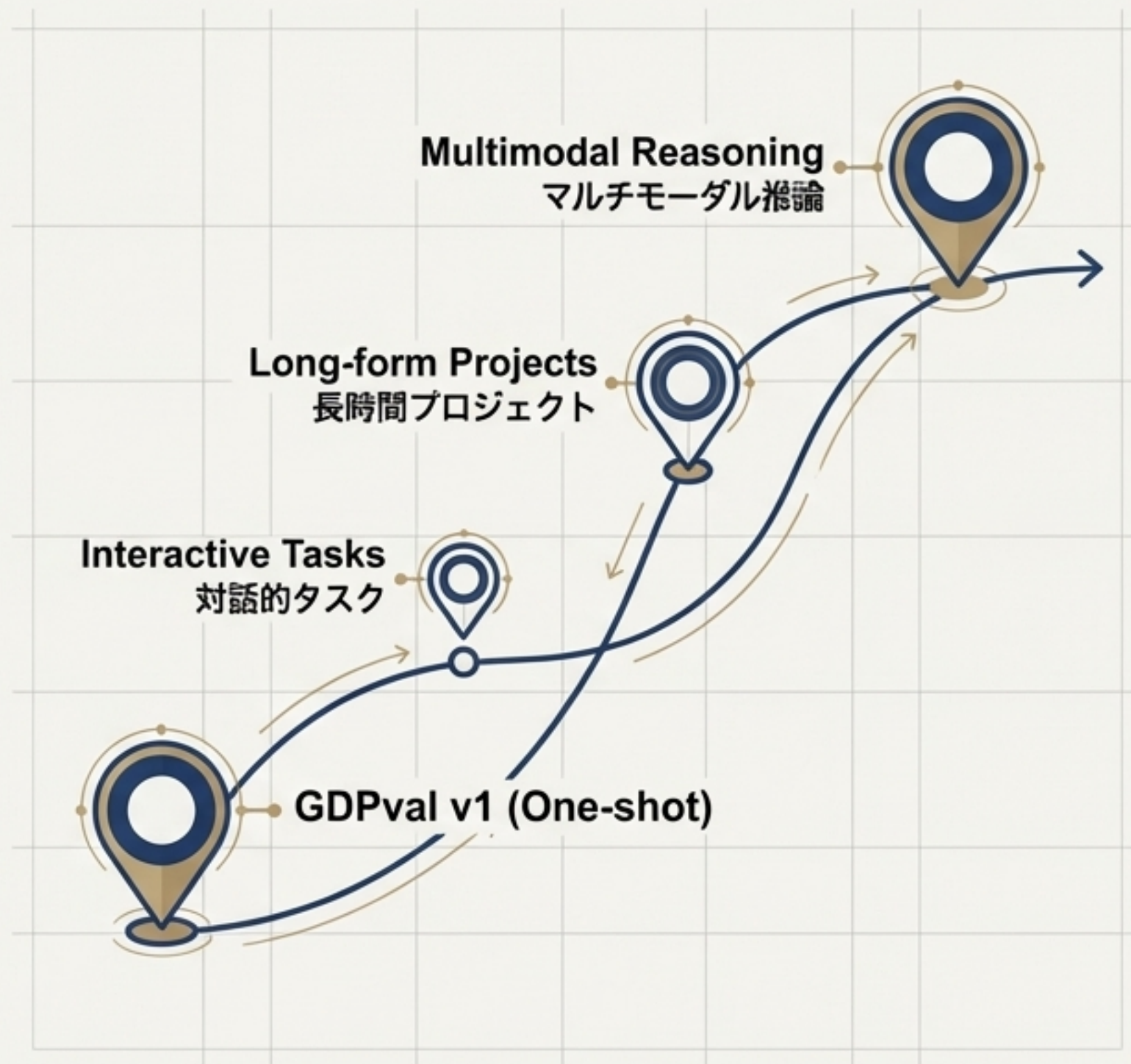
ベンチマークの拡張：OpenAIは、フィードバックに基づく修正や長時間タスクなど、より対話的・反復的なワークフローを評価できるよう、今後のバージョンでの拡張を計画。

オープンなエコシステム：タスクデータセットの一部は公開されており、コミュニティ全体でAIの実務能力向上を競い合う流れが生まれています。

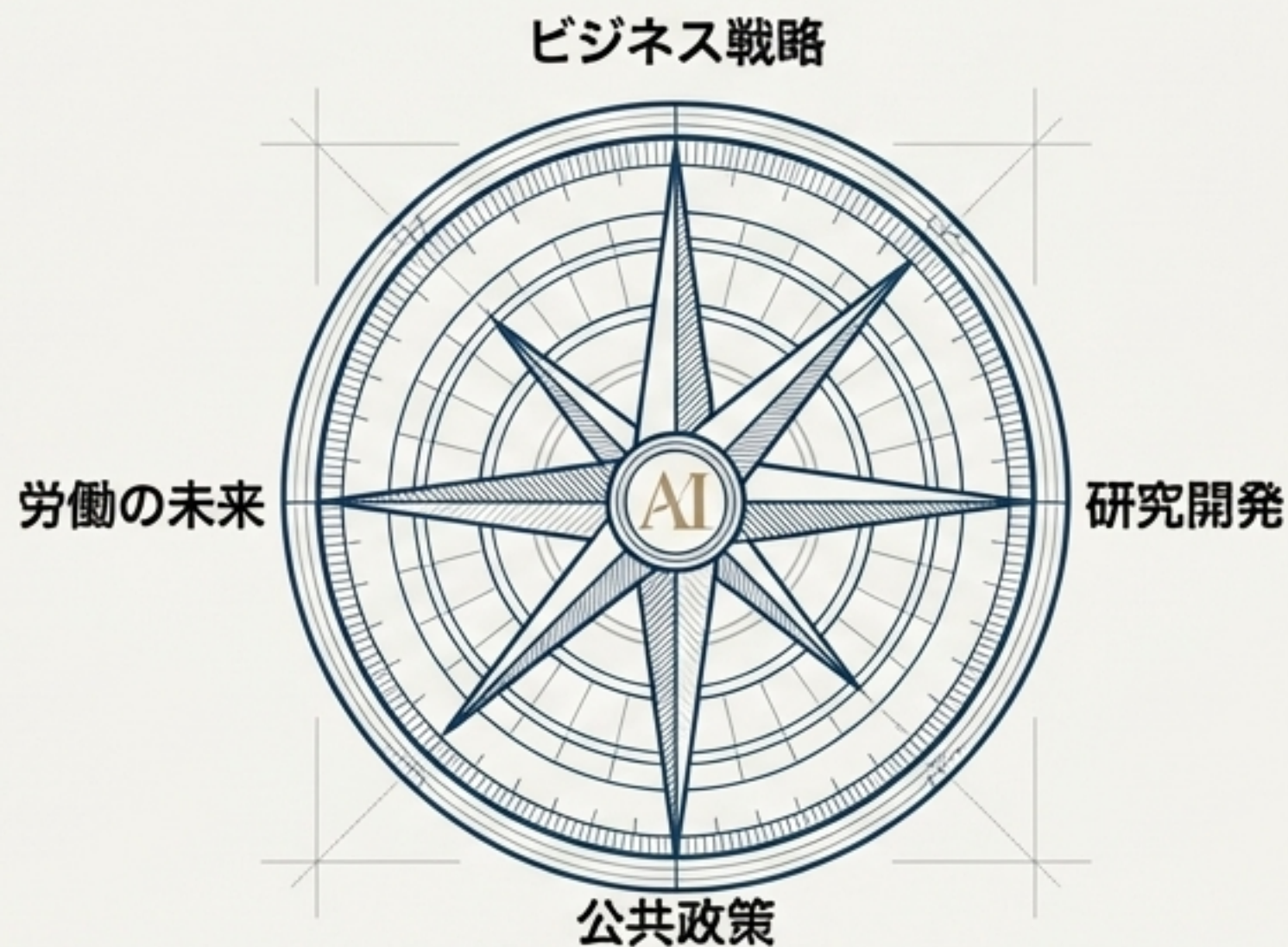
AI研究開発への指針

GDPvalは、AIコミュニティに対し「次に克服すべき課題」を具体的に示します：

- ➊ マルチモーダルな長時間タスクへの対応
- ➋ 指示遵守や計算正確性の向上
- ➌ 書式の整ったプレゼン生成能力の強化



GDPvalは単なる”成績表”ではない。AI経済の未来を形作る”羅針盤”である



GDPvalは、AIに関する議論を「何ができるか」という技術的な問いから、
「どのような経済的・社会的価値を創造できるか」という本質的な問いへとシフトさせました。

この新しい羅針盤を手しに、企業、研究者、政策決定者、そして社会全体が、
AIの能力を最大限に引き出し、その恩恵を公共の利益へと結びつけていくことが求められています。