

OpenAIのベンチマーク指標「GDPval」の調査レポート

GDPval指標の概要と目的

GDPval (Gross Domestic Product valuable tasks) は、OpenAIが2025年9月に発表した新しいAI評価手法であり、AIモデルの「現実世界での経済的価値」を測定することを目的としています¹。具体的には、米国GDPに大きく寄与する上位9業種から選ばれた44職種について、それぞれの職業で実際に行われている専門的タスク1320件を集め、AIモデルがそれらをどの程度うまくこなせるか評価します²。GDPvalという名称は主要経済指標である「国内総生産 (GDP)」の概念に由来し、GDPへの寄与度が高い産業の主要職種からタスクを抽出したことにちなみます³。この評価により、AIモデルが現時点で現実の仕事にどれだけ役立つか（すなわち経済にどれだけ貢献できるか）を定量化し、AIの進歩を実証的に追跡するのが狙いです⁴。従来の評価（学術試験やコーディング競技など）はAIの推論能力を押し広げる上で重要でしたが、日常業務の実態を十分に反映していない場合が多く、このギャップを埋めるためにGDPvalが開発されました⁵。例えば、「人間の専門家ならどう処理するか」を想定してタスクを設計し、弁護士や映画監督など現実の職業としての性能を測定できるようにしています⁶⁷。

タスク設計と評価手法: GDPvalのタスクは、各職種の分野経験平均14年以上の専門家チームが作成・精査しており、法律意見書・設計図・カスタマーサポート対話・看護ケア計画など実際の業務成果物に基づいています⁸。全タスクのうち220件は「ゴールド標準」としてオープンソース公開されており、他の研究者も利用可能です⁹。評価はブラインド方式で実施され、業界の専門家がAIモデルと人間専門家の成果物を見分けがつかない状態で品質比較を行いました¹⁰。この際、AIモデルの出力が人間の成果物より優れている（勝ち）か同等（引き分け）と判断された割合が記録され、モデルごとの性能指標となっています¹¹。また、タスクは単なるテキストプロンプトではなく、関連資料ファイルや背景コンテキストが与えられ、期待される成果物も文書・スライド・図表・スプレッドシート・マルチメディアなど多岐にわたります¹²。このように実務そのものに近い形式でモデルをテストすることで、AIが理論上できることだけでなく日常の仕事への支援としてどこまで有用かを評価できるよう工夫されています¹³¹²。

技術的意義（他のベンチマークとの比較・優位性・AI研究への影響）

GDPvalの登場は、AIモデル評価における新たな技術的マイルストーンとされています。その技術的意義を、従来のベンチマークとの比較を通じて整理します。

- ・**現実性・多様性の向上:** GDPvalは幅広い職種・業種横断のタスクを含む点で他の評価と一線を画しています。他の指標が特定分野の経済価値に特化していたり（例：ソフトウェア開発者向けのSWE-Lancerなど）¹⁴、あるいは人為的に作られた試験問題（MMLUやHumanity's Last Examなど）でモデルの学術性能を測るのに留まっていたのに対し、GDPvalは実在する業務成果物そのもの、またはそれを模したタスクを重視しています¹⁴。タスク形式もマルチモーダルで、モデルには関連資料を参照しつつ文書や図表を作成させるなど、仕事の実務プロセスに近い設定となっています¹²。この点は単純なテキストQA形式のベンチマークとは大きく異なり、AIの実用的な応用力を試す上で優位性があります¹⁵¹⁶。
- ・**タスク範囲と網羅性:** GDPvalは米国経済で重要な9業種から知識労働職を抽出しており、業界横断的な評価を可能にしています¹⁷。例えば対象職種には、ソフトウェアエンジニアや弁護士といったAI活用が進んでいるような職種だけでなく、私立探偵や薬剤師、ソーシャルワーカーなど従来あまりAI浸透

していない職種も含まれています^{18 19}。これは評価範囲を偏らせず、**AIの得意不得意を幅広く測定する意図**と考えられます。さらに選定プロセスでは、ONETデータから業務タスクの60%以上が肉体労働を伴わない職種を抽出することで、AIが影響を与えやすい知識労働に焦点を当てています²⁰。このような包括的かつ合理的なタスク設定により、GDPvalはAIモデルの総合的な能力評価*を目指しています。

- モデル性能向上の指標:** GDPvalで得られた結果から、**最新のAIモデルが専門家レベルに近づきつつある**ことが示唆されています。初回評価では、OpenAI GPT-5やAnthropic Claude Opus 4.1など最先端モデルの出力が**ほぼ半数のタスクで人間専門家の成果物と同等以上**と評価されました^{21 22}。特にClaude Opus 4.1は**文書フォーマットやスライド体裁などの美的品質**で強みを示し、GPT-5は**領域知識の正確性**で優れていたと報告されています^{23 24}。さらにモデル間の比較では、2024年春リリースのGPT-4oから2025年夏リリースのGPT-5にかけて、**GDPvalスコアが2倍以上に向上**しており、性能の線形的な伸びが確認されました²⁵。実際、OpenAI最新モデルのGPT-5.2では**70.9%のタスクで人間専門家と同等以上の結果**を達成したとの報告があり（専門家に対する勝敗・引分率）、モデルの飛躍的進歩を裏付けています²⁶。
- AI研究開発への影響と応用:** GDPvalは**AIモデルの改良目標を現実的なタスクで定める**ことで、研究開発の指針となります。例えば、OpenAIはGDPvalを用いてGPT-5モデルの実験的バージョンを追加学習させる実験を行い、**性能の更なる向上が可能**なことを示しました²⁷。この結果は「モデルサイズの拡大」「推論ステップ数の増加」「豊富なタスクコンテキスト提供」といった施策が**GDPvalスコアを押し上げる**ことを示唆しており、今後のモデル開発に具体的な方向性を与えます²⁸。また、GDPvalの一部タスクや評価スキームはオープンに公開されているため、**外部の研究者や企業も自社モデルをGDPval基準で評価**したり、タスクデータセット（Hugging Face上の公開データなど）を活用して**モデルのfine-tuning**を行うことができます^{9 29}。これによりAIコミュニティ全体で**共通ベンチマークとしてGDPvalを発展**させ、AIの実務能力向上を競い合う流れが生まれています。実際、GDPvalで測定したモデル間の差異（例：GPT-5は指示遵守や計算正確性に秀で、Claudeは書式の整ったプレゼン生成に秀でる³⁰）は、モデルアーキテクチャや訓練手法の比較分析にも役立ちます。総じて、GDPvalは**AIモデルを現実業務に近い土俵で評価する新基準**を打ち立て、AI研究者にとっては**次に克服すべき課題**（例えばマルチモーダルな長時間タスクへの対応）を浮き彫りにするという意義があります。

以下の表に、従来の代表的なAIベンチマークとGDPvalの特徴を比較します。

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 （学術試験ベンチマーク）	数十分野の四択試験問題 ³¹	学術知識の力 バー範囲と推論力	高：試験問題形式 （実務文脈なし）
コードコンテスト課題 （CodeEval等）	プログラミング問題、アルゴリズム競技	コーディング能力・論理思考	中：人工的な問題 設定が多い
SWE-Bench （バグ修正課題）	ソフトウェアエンジニアリング（バグ修正） ³¹	専門領域内の実装力	中：狭い領域、実務の一部のみ
SWE-Lancer （フリーランス開発課題）	実契約に基づく開発タスク ³²	市場環境下での実装力	低：現実の依頼だが分野限定
GDPval	44職種の実務タスク（文書作成・分析・設計等） ^{17 12}	知識労働における成果物の品質	極めて低：実在の成果物ベース

(注: 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
12。)

経済的意義（経済分析・企業活用・産業界での評価）

GDPvalはAIの現実的な経済インパクトを定量化する初の試みであり、その結果は**経済分析や企業戦略においても重要な示唆**を与えています。

・**生産性・コストへの影響:** GDPvalの評価によれば、最先端のAIモデルは人間の専門家より**タスク遂行が約100倍速く、コストは1/100程度**で済むことが明らかになりました³³。これは純粋なモデル推論時間とAPI利用料金に基づく試算ですが、もしAIを適切に活用できれば**極めて大きな生産性向上とコスト削減**の可能性があることを示しています³⁴。もっとも、OpenAI自身も指摘するように、この比較には現実の職場で必要となる**人間による監督や結果の統合といった手間は含まれていません**³⁴。実務でAIを使う際には確認・修正のプロセスが不可欠ですが、それでも「モデルが得意なタスクは人間が手を付ける前にAIに任せることで時間と費用を節約できる」だろうと述べられています³⁵。実際の現場での試算では、AIによるドラフト作成を専門家がレビュー・改善するワークフローにより、**作業全体が約1.4倍速く、コストも1.6倍程度低減**したケースが報告されています³⁶。これは100倍という理論値には及ばないものの、**AIと人間の協業でもすでに顕著な効率化が測定可能**であることを意味します。企業にとっては、このような効率向上を積み重ねることで**大幅なコスト削減や業務スピードアップ**が期待でき、競争力向上につながります。

・**タスクの経済価値と市場インパクト:** GDPvalが対象とする44職種は、米労働統計局のデータに基づき**各業界で賃金総額への貢献が最大級の職種群**です²⁰。例えば、この中には「プロジェクト管理スペシャリスト」も含まれ、その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³⁷³⁸。GDPvalで扱われるタスクは平均で**人間専門家が7〜9時間かけ約400ドルの価値を生み出す仕事**であり³⁹⁴⁰、それらをAIが担えるとなれば**巨大な経済的付加価値の創出**につながります。事実、AIモデルの性能がさらに向上し**専門家並みのアウトプットが常時得られる**になれば、一人の人間がAIを駆使してこれまで数日かかった仕事を数時間で完了する、といった生産性革命も現実味を帯びます⁴¹⁴²。企業や投資家にとって、GDPvalは「**AIが参入し得る市場（労働コスト）の青写真**」とも言えます。AIラボやベンチャーキャピタルは、人件費総額（言い換えれば**各職種の給与と市場規模**）に着目してAI導入の有望分野を探りますが、GDPvalはまさに**給与総額の大きい知的業務領域**に沿った評価であり、**AIが今後 disrupt（創造的破壊）し得る職種群を示すロードマップ**になっているとも指摘されています⁴³⁴⁴。言い換えれば、GDPvalで高得点を出せる職種ほど**AIによる代替・補充が進みやすい領域**であり、企業はその領域の業務プロセスを再設計したり人材育成方針を変える必要が出てくるでしょう。

・**企業での活用と産業界の評価:** GDPvalの指標は、**企業が自社のAI活用戦略を検討する上でのベンチマーク**になり得ます。例えば、ある企業の業務がGDPvalのタスクと類似しており、最新モデルがそれらを高い水準でこなせると分かれば、**その企業はAI導入による効率化余地が大きい**と判断できます。既にOpenAIは顧客企業向けに「GDPvalの今後のラウンドに貢献しませんか」と呼びかけており⁴⁵、ユーザ企業から具体的なユースケースやデータを募ってベンチマークを拡充しようとしています。これは**産業界の実課題を評価指標に取り込む動き**であり、企業自身もGDPvalの枠組みに参加することで**自社業務にAIをどう活かすかを模索**できます。また、産業ごとのAI適用度合いを比較する材料としてもGDPvalは有用です。44職種には**金融、製造、医療、公共サービス、IT**など多彩な分野が含まれているため、各業界団体や規制当局も**どの業界でAIの経済効果が大きく現れそうか**を把握できます。経済分析面では、GDPval結果をもとに「AIが何%のタスクをこなせるようになったか」を定点観測すれば、それを元に**潜在的なGDP押し上げ効果や労働市場への影響**を定量モデルに組み込むことも可能でしょう⁴⁶⁴⁷。実際、スタンフォード大学やNBERのエコノミストらは、AIが**経済成長・所得分配・人々の幸福に与える影響**について**包括的な研究アジェンダ**を提言し始めています⁴⁶⁴⁸。GDPvalのデータはそうしたマクロ経済研究の実証的基盤にもなり得ます。総じて、GDPvalは**AIの経**

済価値を測る共通指標として、企業の技術投資判断から政府の経済政策まで幅広い領域で参考にされる意義を持っています。

社会的・倫理的意義（政策・社会実装・公共利益への示唆）

GDPvalによって浮かび上がったAIの能力と限界は、社会や政策の文脈でも重要な示唆を与えます。モデルが高度化し実務タスクを担える割合が増えてくる中、我々はどのようにAIと共存し、その利益を公共のものとするかという課題に取り組む必要があります。

・**雇用と労働の未来：** GDPvalの初期結果は、AIが反復的で明確に定義されたタスクであれば既に専門家より高速・低コストでこなせることを示しました⁴⁹。これは、今後の仕事の在り方に大きな影響を与えます。ただし同時に、「ほとんどの仕事は明確に記述できるタスクの単なる集合ではない」ことも指摘されています⁴⁹。現実の職場では、複数のタスクが繋がったワークフローや、明文化されていない暗黙知、判断力が必要となる局面が数多く存在します⁵⁰⁵¹。そのため、GDPvalが高得点だからといって即座に「職業そのものがAIに置き換わる」わけではなく、職務内容や役割の再構成という形で変化が現れると考えられます⁵¹⁵²。政策決定者にとって、この結果は職業訓練や教育の方向性を見直す契機となります。AIに任せられる定型業務は今後ますます増える一方で、人間には創造性・倫理的判断・対人コミュニケーションなどAIが苦手な領域での価値発揮が求められるでしょう⁴⁹⁵³。教育政策ではSTEM教育だけでなく、より創造的思考力や批判的思考、AIとの協働スキルを育むカリキュラムへの転換が示唆されます。また労働政策面では、AIによる自動化で移行期に職を失う人々への再研修プログラムや、職種転換を支援するセーフティネットの整備が公共利益の観点から重要になるでしょう。

・**人間とAIの協働・補完：** OpenAIは「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」と述べています⁵³。これは、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という社会的ビジョンです。GDPvalの結果からも、単純作業はAIが肩代わりし、人間はより創造的で主観的判断を要する業務に時間を割ける可能性が示されています⁴⁹⁵³。このような人間とAIの協働が進めば、生産性向上による経済成長だけでなく、働き方の質的向上（人間はより付加価値の高い仕事に専念できる）という公共的利益も期待できます。社会実装の面では、企業や公共機関がAIを導入する際に人間の役割をどのように再定義するかが問われます。例えば、プロジェクトマネージャーの仕事は従来の計画書作成などから、AIが生成した成果物の検証・調整といった「人間による高次の監督」へとシフトしつつあると分析されています⁵⁴⁵⁵。実際GDPvalでも、モデル出力の約29%は「欠陥が大きい失敗」であった一方、残りの多くは「一部利用可能だが完璧ではない」程度の失敗であったと報告されました⁵⁶⁵⁷。このことは人間のレビューと改善を組み合わせれば多くのタスクで実用に耐えること、そして人間の熟練が依然不可欠であることを示唆します。政策的には、こうした人間-AI協働スキームが各業界で浸透するよう促進策を講じること（例：ガイドライン策定、ベストプラクティス共有）が有用でしょう。また、AI時代の労働倫理として、AIが出力した成果物に対する最終的な責任は誰が負うのか、AIを使うことで生じる判断の偏りをどう是正するか、といった新たな課題にも向き合う必要があります。GDPvalはモデルの性能のみを評価しますが、社会実装にあたっては説明責任・公平性・透明性といった倫理基準も並行して確立しなければなりません。

・**政策決定と公共ガバナンス：** GDPvalのフレームワークは、政府や規制当局にとってもAIガバナンスの参考指標となります。AIが経済に与える恩恵とリスクを把握するため、公共政策として定期的なAIインパクト評価を行うことが考えられます。例えば「年次AI白書」にGDPvalのような指標を盛り込み、各年でAIの能力がどこまで進展しどの産業に影響を与えたかを報告すれば、政策の優先順位（教育投資や雇用対策すべき領域等）をエビデンスに基づいて決めやすくなるでしょう。さらに、GDPvalは各タスクに経済価値（人間換算の所要時間や賃金）を紐づけているため³⁹⁴⁰、AIが創出する付加価値を測定し課税や再分配を検討する材料にもなり得ます。例えば、AIシステムが年間〇ドル相当の業務を処理したというデータがあれば、その恩恵を社会全体に還元するための仕組み（「AI産業への課

税とリカレント教育への充当」など）を議論する根拠となるでしょう。倫理的観点では、GDPvalが示すようにAIの能力が急速に高まる中で、人々の**心理的な不安や社会的受容**にも目配りが必要です。**AIへの過信**によるリスク（例えばAIが作成した看護計画が誤っていた場合の責任問題）と、逆に**AI忌避**による非効率（AIが使えるのに使わないことで損失が出る場合）をどうバランスするか、これは公共政策における難題です。GDPvalは客観指標として「AIは既にここまでできる」と示してくれるため、**冷静かつ具体的な社会対話**を促す材料となります。実際、この評価の公開により「AIはまだ大したことがない」という見方から「**自分の仕事にもAIが十分使える部分があるかもしれない**」という実感へと、世論も変化しつつあるとの指摘があります⁵⁸。こうした認識の変化自体が、社会全体でAIを正しく理解し適応していく上で重要な一歩と言えるでしょう。

GDPval指標に対する批判や懸念

GDPvalは画期的な評価基準ですが、その手法や結果についていくつかの批判や懸念も提起されています。以下に主なポイントを整理します。

・**評価手法の限界（タスク設定と現実の乖離）**：最大の指摘は、GDPvalのタスクが「あまりに綺麗に定義され過ぎている」点です。実世界では、多くの業務が曖昧な要求や不確実な情報から始まり、試行錯誤や対話を経てタスクが固まります。対してGDPvalでは、タスク内容や目的が最初からはっきり提示され、必要な資料もすべて参照として与えられています¹²。これはAIモデルの性能を客観に測るため意図的にそうしているのですが、「**ユーザー自身がタスクのやり方を知らない場合、AIに完璧な指示を書くことなどできない**」という指摘がある通り、実際の利用場面では人間がそこまで詳細なプロンプトを用意できないケースも多々あります。このため、一部の論者は「完全に仕様が決まった一回きりのタスク」での評価は**リアルな仕事ぶりを測るには不十分**であり、**対話的・反復的なフローを含む評価**が今後必要だとしています⁵⁹⁵⁰。OpenAI自身も、現行GDPvalは**one-shot型（一回プロンプト回答のみ）のテストに限定**されており、**フィードバックを受けた修正作業や長時間かけてコンテキストを構築するタスク**は範囲外だと認めています⁶⁰⁶¹。今後のバージョンではよりインタラクティブなワークフローや曖昧さへの対処能力も評価できるよう拡張する計画が述べられています⁶²。したがって現時点のGDPval結果を「**AIが仕事を完全に肩代わりできる証拠**」と過大解釈するのは禁物であり、むしろ「**AIが今日できる具体的タスク**」を示すベンチマークと捉えるべきだという意見があります⁵⁰⁶³。

・**カバー範囲とバイアスの問題**：GDPvalは当初44職種に限られており、**全ての職業領域を網羅していない点も指摘**されています⁵⁹。特に、肉体労働や対人サービスの比重が高い職種（例：介護士、建設労働者など）は評価対象外であり、**AIがそうした領域に与える影響**はこの指標からは読み取れません。また地理的にも米国の産業構造・職種に基づいているため、新興国や日本など他国経済への適用には注意が必要です。GDPvalという名前自体、「経済価値（＝GDP寄与）のあるタスク」にフォーカスしており、これは裏を返せば**GDPに直接は現れない価値（例：ボランティア活動、芸術的創造性、家庭内労働など）**は評価の射程外だということです。AIの社会的意義は経済以外の文脈（例えば医療や福祉での人命やQOL向上など）にもあるはずですが、GDPvalはあくまで**市場経済上の価値尺度**に則っているため、そこに**価値観の偏り**があるとの批判も考えられます。例えば「AIがGDPに貢献しないような領域で良い影響を与えてもGDPvalでは評価されない」という点は、政策判断でこの指標だけに頼ることへの警鐘と言えます。また、タスク作成者や評価者が業界の熟練専門家であることから、**専門家バイアス**の可能性も議論されています。すなわち、成果物の評価基準が伝統的な専門家の観点に縛られていれば、斬新なAIソリューションが提示する**新しい価値を過小評価**する恐れがあります（例えばAIが独自のアプローチで課題を解決しても「人間らしくない」という理由で減点される等）。ただし今回のGDPvalでは、評価者はAI出力か人間出力か区別せず**出来栄そのもので採点**しているため¹⁰、そうした露骨なバイアスは避けられています。さらにOpenAIは人間評価の限界を補うため**自動評価AI**の試行も報告しましたが、人間ほど信頼できず当面代替は難しいとも述べています⁶⁴⁶⁵。評価手法自体の公平性・一貫性の担保も今後の課題と言えるでしょう。

・**社会的影響への懸念:** GDPvalが示すAIの高性能ぶりに対し、**労働市場や社会への影響を懸念する声も**上がっています。例えば「2026年にはAIが丸一日自律的に働けるようになり、年末までに複数の産業で人間専門家と同等のモデルが登場、2027年には多くのタスクで人間を凌駕する」という予測さえ専門家から出ており⁴⁶⁴²、急速な技術進歩に社会が追いつけるか不安視されています。このような予測に基づき、経済学者たちは**変革的AI（transformative AI）が雇用・所得分配・福祉に与える影響**を研究し、政策対応を検討すべきだと警鐘を鳴らしています⁴⁷⁴⁸。具体的な懸念としては、AIによって**一部の知的職業が大幅に自動化された場合の失業や格差拡大**の可能性があります。GDPvalの対象には高所得専門職も多く含まれますが、もしそれらの仕事がAIで効率化され企業利益が増えても、その果実が労働者に還元されなければ所得格差が拡大しかねません。さらに、AIの導入により**労働者のスキル陳腐化**（AIにできない創造的・高度スキル以外は価値が下がる）が起これば、中間層の雇用が圧迫され社会不安につながるリスクも指摘されています。「誰もがAIの上りエレベーターに乗れるようにする」というOpenAIの理念⁵³はこの懸念への回答ですが、実現には**教育制度の改革や生涯学習支援**、さらには**富の再分配**など政策レベルでの対応が不可欠でしょう。また倫理面では、AIが経済価値を生み出すことに過度に注目すると「**経済効率 vs 人間の尊厳**」という問題も浮上します。例えば医療や福祉の現場でAIが事務作業を代行しコスト削減できても、患者や利用者が感じる安心感や人間的なケアは経済指標に表れにくい部分です。AI導入でサービスの効率は上がっても人間味が失われ公共の福祉が損なわれるようなことが無いよう、**定量的指標と定性的価値のバランス**を取ることが大切だという指摘もなされています。

・**モデル依存と安全性:** GDPvalの結果が注目される一方で、「モデルが出力した成果物をそのまま信用してよいか」という**AIの信頼性・安全性**の問題も依然残ります。GDPval評価でも**約3割のケースでAIのアウトプットが専門家から見て“悪い/壊滅的”と判定されており**⁵⁶⁵⁷、例えば法律文書に致命的な誤りが混入したり、設計図に安全上の欠陥がある可能性を示唆します。こうした失敗を人間が検知できずに実務に使ってしまえば、法的紛争や事故等の重大な結果を招きかねません。従って、**AIを仕事に使う際の責任の所在や誤出力の検出体制**といった安全面の枠組み作りが急務です。GDPvalは**AIの能力を示すもの**であって、そのまま実運用での無謬性を保証するものではありません。評価上は人間と互角でも、AIは**確率的推論により不確実な回答をする**性質があるため、重要な場面では常に**人間の監査とフェイルセーフ**が必要だという点は強調しておくべきでしょう。また、AIモデルがトレーニングデータに由来する**バイアスや不公平さ**を持っている可能性もあり、GDPvalではそうした側面は評価対象外です。例えばカスタマーサポートの対話文書をAIが作成する場合、特定の人種や性別に不適切な応対をするリスクも考えられますが、GDPvalは成果物の品質評価が中心で倫理的チェックは含んでいません。そのため、**性能評価と並行して倫理基準・規制による補完**が必要となります。「責任あるAI」や「AIガバナンス」に関する議論では、まさにこうした**技術指標と社会価値の統合**が求められており、GDPvalのデータもその議論の土台として活かされるべきでしょう。

以上のように、GDPvalはAIの**現実世界での経済的貢献度**を測る革新的なベンチマークですが、その解釈や活用にあたっては**評価範囲の限界や現実とのギャップ**を正しく認識することが重要です。同指標を起点に、技術コミュニティと社会全体が協調しつつ、**AIの能力を最大限公共の利益へ結びつけ、リスクを低減する方策**を講じていくことが求められています。

参考文献・情報源: GDPval公式発表¹¹⁷、OpenAIブログ日本語版¹²⁴⁹、OpenAI論文（arXiv）¹⁵⁶⁶、各種メディア報道および専門家コメント³³⁵⁰⁵⁶など。

- 1 2 3 4 5 8 9 10 11 12 13 14 17 20 23 24 27 28 31 32 49 53 60 61 62 実際のタスクに
対するモデルのパフォーマンスを測定する | OpenAI
<https://openai.com/ja-JP/index/gdpval/>
- 6 7 25 33 34 35 AIの現実世界での能力を測定するベンチマーク「GDPval」をOpenAIが開発、弁護士や
映画監督など現実の職業としての性能を測定可能 - GIGAZINE
<https://gigazine.net/news/20250926-openai-benchmark-gdpval/>
- 15 16 21 22 42 43 44 46 47 48 58 66 OpenAI's New Benchmark Shows AI Does Knowledge Work
100X Faster and Cheaper Than Experts
<https://www.marketingaiinstitute.com/blog/openai-gdpval>
- 18 19 64 65 AIモデルの実務能力を測定--OpenAIの新たな評価指標「GDPval」とは - ZDNET Japan
<https://japan.zdnet.com/article/35238606/>
- 26 OpenAI、最新モデル「GPT-5.2」発表 専門家レベルの推論能力 - ケータイ Watch
<https://k-tai.watch.impress.co.jp/docs/news/2070782.html>
- 29 openai/gdpval · Datasets at Hugging Face
<https://huggingface.co/datasets/openai/gdpval>
- 30 36 37 38 39 40 54 55 56 57 OpenAI's GDPval Explained | IPM
<https://instituteprojectmanagement.com/blog/openais-gdpval-explained/>
- 41 50 51 52 59 63 OpenAI's GDPval evaluation: a benchmark for AI's economic value | Jared Spataro
posted on the topic | LinkedIn
https://www.linkedin.com/posts/jaredspa_openais-new-gdpval-evaluation-is-sparking-activity-7380996649827028992-qen2
- 45 OpenAI の GDPval に貢献する
<https://openai.com/ja-JP/form/gdpval-customer-contribution/>