ARC-AGI-2 が示す LLM 推論力と知的財産業務への実務的影響

Claude Opus 4.6
2026 年 2 月 22 日作成

ARC-AGI-2 ベンチマークにおける LLM モデル間の性能差は極めて大きく、その差は知的財産（IP）業務の質に直結する。2025 年 3 月の公開時点で主要 LLM の正答率は 0〜4%にとどまったが[1][2]、2026 年 2 月現在、Gemini 3 Deep Think が 84.6%に到達し、人間平均の 60%を大きく超えた[3]。ARC-AGI-2 が測定する「抽象推論」「新規パターンへの適応」「一般化能力」は、特許文書の本質的理解、先行技術調査の精度、発明の新規性・進歩性判断において決定的に重要な能力であり、モデル間の推論力の差が特許実務の信頼性と効率に重大な違いを生んでいる。

1. ARC-AGI-2 は「流動性知能」を測定する初の AI ベンチマーク

ARC-AGI-2 は、François Chollet が 2019 年に発表した ARC（Abstraction and Reasoning Corpus）の第 2 版として、2025 年 3 月に ARC Prize 財団から公開された[1][4]。心理学者 Raymond Cattell の**流動性知能**（fluid intelligence）と**結晶性知能**（crystallized intelligence）の概念に基づき、ARC-AGI-2 は前者——すなわち事前知識に依存せず未知の問題を解決する能力——を測定する[4][5]。

タスクの形式は色付きグリッド（1×1〜30×30、10 色）の変換規則推論である。各タスクで 2〜5 組の入出力ペアが示され、システムはそこから抽象的な変換規則を推論し、未見のテスト入力に適用する。ARC-AGI-1 との主要な違いは 3 点ある。第一に、すべてのタスクが完全に新規で暗記が不可能。第二に、グリッドが大きく 1 タスクあたりの情報量が増大。第三に、複数の規則を新しい組み合わせで適用する深い構成的一般化が求められる[4][5][6]。

ARC-AGI-2 が要求する核心的な認知能力

シンボルの意味解釈：視覚的パターンを超え、記号に意味を付与して推論する能力。**多段階の構成的推論**：複数の相互作用する規則を順序立てて適用する能力。**文脈依存の規則適用**：グリッド内の特定文脈に応じて規則を柔軟に変える能力。**少数事例からの一般化**：わずか 2〜5 例から抽象的法則を抽出する能力[5][6]。

日本の AI 分析メディア Innova（イノーバ）は、これを Kahneman のシステム 1（直感的・高速）とシステム 2（論理的・低速）の枚組みで解説し、現在の LLM がシステム 1 に優れる一方、ARC-AGI-2 が要求するシステム 2 的思考で根本的に弱いと指摘している[7]。

2. モデル間の性能差——0%から 84.6%まで

ARC-AGI-2 のスコア推移は、LLM の推論能力の進化と限界を鮮明に映し出している。純粋な

LLM（推論モード非搭載）のスコアが 0%である事実は注目に値する。日本のテクノエッジ誌はこれを「歯が立たない」と表現し、ナゾロジーは「全滅」と報じた[8][9]。しかし 2025 年後半から 2026 年初頭にかけて、推論拡張機能を搭載したモデルが急速にスコアを伸ばした[3][10]。

モデル	ARC-AGI-2 スコア	タスク単価	時期	出典
GPT-4o / GPT-4.5 / o3-mini	0%	—	2025 年 3 月	[1]
Claude 3.7 Sonnet	0.9%	—	2025 年 3 月	[1]
o3（低コスト設定）	4%	約$200	2025 年 3 月	[1]
NVARC（コンペ 1 位、4B）	24%	$0.20	2025 年 11 月	[11]
Claude Opus 4.5（Thinking）	37.6%	$2.20	2025 年 12 月	[12]
GPT-5.2 Pro（X-High）	約 54%	高額	2025 年 12 月	[10]
Claude Opus 4.6	68.8%	—	2026 年 2 月	[12]
Gemini 3.1 Pro	77.1%	—	2026 年 2 月	[13]
Gemini 3 Deep Think	84.6%	—	2026 年 2 月	[3]
人間平均	60%	$17	2025 年基準	[1]

表 1: ARC-AGI-2 における主要モデルのスコア推移

ARC Prize 2025 コンペティション（Kaggle）には 1,455 チーム・15,154 エントリーが参加した。優勝した NVARC チーム（NVIDIA の Ivan Sorokin & Jean-Francois Puget）は、わずか 40 億パラメータのモデルを合成データ生成とテスト時学習で微調整し、数千億パラメータの商用モデルを上回る 24%を$0.20/タスクで達成した[11]。85%の大賞（$700K）は未達成のまま終了している[12]。

3. 抽象推論能力が IP 業務の質を左右する 5 つの局面

3.1 先行技術調査における意味的理解

特許調査で「電磁パルスによる雑草制御」の先行技術を探す場合、「エネルギー場を利用した非化学的害虫管理」との関連性を認識するには、表面的なキーワードを超えた抽象的概念の対応付けが必要である。Carnegie Mellon 大学の Ikoma & Mitamura（2025 年）による初の体系的研究では、生成型 LLM が特許新規性判断において合理的な精度で予測を行い、特許と先行技術の関係を理解するのに十分な説明を生成できることが示された[14]。

3.2 発明の本質的特徴の抽出

特許文書は平均で説明部分が 5,451 トークン、請求項が 962 トークンに達する長大な文書である。ScienceDirect（2025 年）の研究では、GPT-3.5-turbo による特許明細書の要約が RoBERTa の分類精度を F1 値で 2.9〜3.0 ポイント向上させた。これは LLM が長い詳細な説明から発明の本質的な課題と解決策を抽出できることを示す[15]。

3.3 特許の類似性評価と分類

PatentMind（2025 年）が開発した多面的推論グラフ（MARG）フレームワークは、GPT-4o-mini を使用して特許類似性評価で専門家評価とのピアソン相関 0.938 を達成した。この「構造化された多次元推論」こそが優れた性能の鍵であり、単純な LLM 呼び出しでは達成できない[16]。

3.4 非自明性（進歩性）の判断

「引用文献 A と B を組み合わせることが当業者にとって自明か」という判断は、本質的に類推推論である。異なる技術分野の解決策間の構造的類似性を認識し、組み合わせの動機付けの有無を評価する必要がある。ARC-AGI-2 が測定する「少数事例からの規則抽出と未見の文脈への適用」は、この判断プロセスと構造的に同型である[5][15]。

3.5 IP ランドスケープ分析における傾向認識

数千件の特許を処理して競合他社の技術戦略を把握する際、LLM のパターン認識能力が決定的になる。EvoPat（2024 年）の多エージェント LLM システムは、RAG を活用し、単一エージェントの GPT-4 を要約・比較分析・技術評価のすべてで上回った[17]。

4. 高性能モデルと低性能モデルで実務はこう変わる

IP 専用ベンチマークの結果は、モデル間の推論力の差が IP 業務の実用性を根本的に左右することを裏付けている。

IPBench（2025 年、最も包括的な IP ベンチマーク）は、8 つの IP 機構にわたる 20 タスクで 17 の LLM を評価した。Webb の知識の深さ（DOK）分類体系に基づき、情報処理→論理推論→判別分析→拡張思考の 4 段階で評価した結果、**高次タスクほど高性能モデルの優位性が拡大**することが明らかになった[18]。つまり、単純な情報抽出では性能差が小さいが、進歩性判断やクレーム解釈のような高次推論が求められるタスクでは、性能差が決定的になる。

MOZIP ベンチマーク（2024 年）の結果はさらに厳しい。ChatGPT が最高性能を示したものの「合格水準」に達せず、7B パラメータのモデル（BELLE-7b）は PatentMatch-en で 14.2%（ランダム推測以下）にとどまった。長文の専門テキスト処理が主要な障壁である[19]。

業界調査（HGF ウェビナー、2025 年）によれば、特許弁護士の 52%が AI を試用した経験があるが、定常的に使用しているのはわずか 15%にとどまる。最大の懸念は「精度とハルシネーション」（42%）であり、推論能力の高いモデルほどハルシネーションが少なく、専門家の信頼を獲得しやすい[20]。欧州特許庁（EPO）の審決 T1193/23 は、LLM によるクレーム解釈を追加的証拠なしに信頼することに明確に警告を発しており、モデルの推論信頼性は IP 業務での採用可否を決定する閾値となっている。

5. 日本の IP 業界における AI 活用と今後の展望

日本の特許実務における LLM 活用は急速に進展している。日本特許庁（JPO）は 2023 年度に AI 関連特許出願が約 11,400 件に達したと報告し、2024 年 5 月には内閣府の「AI 時代の知的財

産権検討会」が中間報告書を公表している[21][22]。

日本発の注目すべき取り組みとして、**Tokkyo.Ai**（リーガルテック社）が ChatTokkyo（対話型特許検索）や Deep Agent（AI の思考過程を可視化するシステム）を展開している[23]。**Emuni**（東京大学松尾研発スタートアップ）は Meta Llama-3-70B を 600 件以上の翻訳ペアで微調整し、特許翻訳で GPT-4o と DeepL の両方を BLEU と RIBES 指標で上回る専用モデルを開発した[24]。

日本の AI 研究コミュニティは、ARC-AGI-2 の意義を鋭く捕らえている。Reinforz Insight は、人間が限られた情報から帰納的に意味を構築するのに対し、AI は膨大なデータから演繹するという根本的な差異を強調した[25]。note.com の MBBS/佐藤源彦は、GPT-5.2 が o3 モデルの 1 年前と比べてコスト効率が 390 倍改善したことを指摘し、推論能力の向上とコスト低下の両面で IP 業務への実用化が加速すると論じている[10]。

ARC Prize 財団は 2026 年初頭に ARC-AGI-3 を発表予定であり、初めて対話型推論（探索、計画、記憶、目標獲得）をテストする形式変更を予告している[12]。これは特許業務におけるエージェント型 AI——段階的な調査戦略の立案、仮説検証の反復、複数データベースの横断探索——への展開と方向性を同じくするものである。

6. 結論：推論力の差は IP 業務の信頼性閾値を決定する

ARC-AGI-2 の結果は、LLM の推論能力が「あれば便利」な付加機能ではなく、IP 業務への適用可否を決定する構造的要件であることを示している。推論モード非搭載の LLM が ARC-AGI-2 で 0%を記録した事実は、単純なパターンマッチングでは IP 業務の核心——抽象的な技術概念の対応付け、複数規則の構成的適用、少数事例からの法則抽出——に到達できないことの直接的な証拠である[1][8]。

2026 年 2 月現在、Gemini 3 Deep Think（84.6%）と Gemini 3.1 Pro（77.1%）が人間平均（60%）を超え、Claude Opus 4.6（68.8%）が続く[3][12][13]。しかし、ARC Prize チームはフロンティアモデルの学習データに ARC 関連データが混入している可能性を指摘しており、スコアの額面通りの解釈には注意が必要である[12]。

IPBench が示すように、高次推論タスクほどモデル間の性能差が拡大する傾向は、特許業務で最も高度な判断——進歩性評価、クレーム範囲の解釈、技術的本質の抽出——において、モデル選択が業務品質を決定的に左右することを意味する[18]。推論能力の急速な向上とコスト低下は、IP 業務における AI 活用の本格化を加速させるが、人間の専門家による検証なしに LLM の出力を信頼することは依然として許容されない。ARC-AGI-2 が測定する「真の推論力」と、IP 業務が要求する「法的・技術的正確性」の交差点に、今後の AI×IP 実務の発展軸がある。

参照文献

[1] ARC Prize Foundation. "ARC-AGI-2." https://arcprize.org/arc-agi/2/ (2025 年 3 月公開)

[2] ARC Prize Foundation. "Announcing ARC-AGI-2 and ARC Prize 2025."

https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025 (2025 年 3 月)

[3] MarkTechPost. "Is This AGI? Google's Gemini 3 Deep Think Shatters Humanity's Last Exam And Hits 84.6% On ARC-AGI-2." https://www.marktechpost.com/ (2026 年 2 月 12 日)

[4] Chollet, F. et al. "ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems." arXiv:2505.11831 (2025 年)

[5] テクノエッジ. "主要 AI モデルはどれも"歯が立たない"、新しい「人間には簡単だが AI には難しい AGI 問題」登場." https://www.techno-edge.net/ (2025 年 3 月 31 日)

[6] イノーバ. "AI の「本当の賢さ」を測る：新テスト「ARC-AGI-2」が暴く推論モデルの弱点." https://innova-jp.com/media/ai-weekly/47

[7] イノーバ. "AI が苦手なパズル：ARC-AGI ベンチマークが示唆する AGI に向けての課題." https://innova-jp.com/media/ai-weekly/13

[8] テクノエッジ. 前掲[5]

[9] ナゾロジー. "主要な AI モデルが AGI テストで全滅：汎用人工知能の高い壁." https://nazology.kusuguru.co.jp/archives/173968

[10] 佐藤源彦@MBBS. "【Gemini3 Deep Think 超え】GPT-5.2 が「抽象的推論」で 50%超え." note.com (2025 年 12 月)

[11] NVIDIA Developer Blog. "NVIDIA Kaggle Grandmasters Win Artificial General Intelligence Competition." https://developer.nvidia.com/blog/ (2025 年)

[12] ARC Prize Foundation. "ARC Prize 2025 Results and Analysis." https://arcprize.org/blog/arc-prize-2025-results-analysis (2026 年 1 月)

[13] Blockchain News. "Gemini 3.1 Pro Launch: Latest Benchmark Breakthrough with 77.1% ARC-AGI-2 Score." https://blockchain.news/ (2026 年 2 月)

[14] Ikoma, S. & Mitamura, T. "Can AI Examine Novelty of Patents?: Novelty Evaluation Based on the Correspondence between Patent Claim and Prior Art." arXiv:2502.06316 (2025 年)

[15] 特許文書の要約と分類に関する LLM 研究 (ScienceDirect, 2025 年; PatentPC 解説記事)

[16] PatentMind MARG フレームワークによる特許類似性評価研究 (2025 年)

[17] EvoPat: 多エージェント LLM システムによる特許分析 (2024 年)

[18] Jiang, Y. et al. "IPBench: Benchmarking the Knowledge of Large Language Models in Intellectual Property." arXiv:2504.15524 (2025 年)

[19] Yang, H. et al. "MoZIP: A Multilingual Benchmark to Evaluate Large Language Models in Intellectual Property." LREC-COLING 2024. arXiv:2402.16389

[20] HGF ウェビナー. 特許業界における AI 活用調査 (2025 年)

[21] 特許庁. "AI 関連発明の出願状況調査." https://www.jpo.go.jp/ (2024 年)

[22] 内閣府. "AI 時代の知的財産権検討会 中間とりまとめ." https://www.kantei.go.jp/ (2024 年 5 月)

[23] Tokkyo.Ai. https://www.tokkyo.ai/ (リーガルテック社)

[24] Emuni. "AI で特許調査のコストを 1000 分の 1 に." https://media.emuniinc.jp/ (2025 年 1 月 10 日)

[25] Reinforz Insight. "OpenAI や Google の AI がわずか数% 新ベンチマーク「ARC-AGI-2」が示す AGI 到達の現実." https://reinforz.co.jp/