ARC-AGI-2 の性能差が知的財産業務に与える影響分析

エグゼクティブサマリー

ARC-AGI-2（Abstraction and Reasoning Corpus for AGI – Version 2）は、AI の**流動的知性**（蓄積された知識ではなく、新しい状況で推論し問題を解決する能力）を測定するベンチマークである。純粋な LLM は ARC-AGI-2 でスコア 0%を記録し、高度な推論システムでも一桁台にとどまる一方、人間は平均 60%を達成する。2026 年 2 月時点では、GPT-5.2 ベースのシステムが 72.9%、Claude Opus 4.6 が 68.8%に到達するなど急速な進歩が見られる。[1][2][3][4]

この性能差は、知的財産（IP）業務における「抽象的推論」「新規性判断」「クレーム設計」など、まさに流動的知性が求められるタスクに直結する。本レポートでは、ARC-AGI-2 で高性能なモデルと低性能なモデルが知財業務でどのような違いを生むかを体系的に分析する。

ARC-AGI-2 ベンチマークの概要

ベンチマークの設計思想

ARC-AGI-2 は、François Chollet が 2019 年に提唱した抽象的推論テストの進化版であり、以下の 3 つの原則に基づいて設計されている：[5][6]

- 暗記や過学習では対応できない、完全にユニークなタスクで構成
- 必要な前提知識を最小限に限定し、純粋な推論能力を測定
- すべてのタスクが少なくとも 2 人以上の人間によって 2 回以内の試行で解決されている[2]

ARC-AGI-2 は ARC-AGI-1 よりも大幅に難易度が高く、ARC-AGI-1 で 87.5%を記録した OpenAI o3 でさえ、ARC-AGI-2 では初期テストで 4%程度に低下した。ブルートフォース解法への耐性を持ち、合成的推論（compositional reasoning）やグローバルなルール導出（global rule induction）が要求される。[7][1]

2026 年 2 月時点のリーダーボード

モデル / システム	ARC-AGI-2 スコア	備考
Johan Land（GPT-5.2 改造）	72.9%	非商用カスタムシステム[4]
Claude Opus 4.6	68.8%	Anthropic 商用モデル[4]
GPT-5.2 Pro (High)	54.2%	OpenAI 商用モデル[3][4]
Gemini 3 Pro	約 45%	Google 商用モデル[3]
GPT-5.1	17.6%	前世代モデル[3]
Grok-4	15.9%	xAI（2025 年 7 月時点）[8]
純粋な LLM（非推論モデル）	0%	パターン認識のみに依存[1][2]
人間平均	**約 60%**	**全タスクが人間に解決可能**[2]

この表から明らかなように、モデル間の性能差は極めて大きい。以下では、この差異が知財業務の各領域にどう影響するかを具体的に検討する。

知財業務への影響：高性能モデル vs 低性能モデル

発明の本質理解と抽出

知財業務の出発点は、発明者が開示した技術情報から**発明の本質的特徴**を抽出し、抽象化することにある。これはまさに ARC-AGI-2 が測定する「新しい状況での抽象的推論」と同質の能力である。[9]

高性能モデル（ARC-AGI-2 スコア 50%以上） は、発明者の説明から技術的原理を抽出し、個別の実施形態を超えた上位概念レベルで発明を把握できる可能性が高い。複数の実施例を観察して共通する抽象的なルールを導出する ARC-AGI-2 のタスク構造は、発明の複数の実施形態から共通する技術的思想を抽出する作業と構造的に類似している。

低性能モデル（ARC-AGI-2 スコア 5%未満） は、発明者の説明をそのまま繰り返すパラフレーズに留まる傾向がある。ARC-AGI-2 で示されたように、パターン認識のみに依存するモデルは新規の問題に対する一般化能力に欠ける。発明の本質がどこにあるかを自律的に判断する能力は期待できない。[10][7]

特許クレーム起案

特許クレームの起案は、発明を適切な抽象度で法的に表現する高度な作業である。研究によれば、LLM は詳細な明細書記述に基づく独立クレームの生成では一定の品質を達成するが、従属クレーム間の論理的連鎖を維持する能力は依然として課題である。[11][12]

業務要素	高性能モデルの期待挙動	低性能モデルの限界
独立クレームの抽象度設定	発明の本質に基づき適切な上位概念化が可能	実施例に引きずられた狭いクレームになりがち
従属クレームの階層設計	技術的特徴の論理的階層を推論し構築	特徴間の依存関係を正確に把握できない[11]
構成要件間のリンケージ	発明全体の技術的一貫性を維持	個別構成要件の記述は可能だが全体整合性に欠ける[13]
記載要件への適合	発明と技術的効果の因果関係を論理的に説明	形式的には正しいが技術的深みが不足

GPT-4 が包括的な人間評価で最良の性能を示した研究は、推論能力の高いモデルほど特徴の網羅性、概念的明確性、技術的一貫性において優れていることを裏付けている。[11]

先行技術調査と新規性・進歩性判断

新規性・進歩性の判断は、特許クレームと先行技術文献の間の**対応関係を推論**し、技術的差異を評価する知的作業である。LLM の新規性審査能力を評価した研究では、生成モデルが妥当なレベルの精度で予測を行い、ターゲット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ている。[14]

高性能モデルは以下の点で優位性を発揮する：

- クレームと先行技術の構成要件を多次元的に分析し、技術的特徴・応用領域・権利範囲の各レベルで類似性を評価（PatentMind フレームワークが示すような構造的アプローチ）[15]
- 単なるキーワードマッチングを超えた意味的理解に基づく先行技術の発見[16]
- 複数の先行技術を組み合わせた進歩性論理の構築[17]

低性能モデルは、分類タスクにおいて新規性を効果的に評価できず、特にクレーム要素と先行技術開示の間の微妙な対応関係の把握に苦慮する。ARC-AGI-2 で測定される「複合的推論」と「マルチステップ変換」の能力不足が、先行技術の組み合わせによる進歩性判断という知財特有のタスクに直接的に反映される。[7][14]

侵害分析・非侵害設計

侵害分析は、特許クレームの各構成要件が対象製品・方法に充足されているかを判断する作業であり、均等論の適用を含めて高度な抽象的推論を必要とする。

高性能モデルは、クレームの文言と被疑侵害品の技術的構成の間の「機能・方法・結果」の対応関係を推論する能力を持つ。これは ARC-AGI-2 のタスクにおける「入力パターンから変換ルールを導出し、新しい入力に適用する」能力と本質的に同じ構造を持つ。[5][10]

低性能モデルは、クレーム文言の字義通りの解釈に留まり、均等論的な判断や設計変更による回避可能性の評価など、抽象的な推論を要する分析では信頼性が低い。AI 技術関連発明の権利行使では、クラウド利用発明における構成要件充足の立証が困難な場合があり、こうした複雑なシナリオでの推論能力の差が結果を大きく左右する。[18]

IP 戦略策定とポートフォリオ分析

IP 戦略の策定には、技術動向の把握、競合分析、ポートフォリオ最適化など、多角的な推論が求められる。

高性能モデルの活用可能性：

- 特許出願トレンドから技術開発の方向性を推論
- 競合のポートフォリオ分析に基づく空白領域の特定
- 出願戦略の費用対効果分析

低性能モデルの限界：

- データの集約・分類は可能だが、戦略的示唆の導出に弱い
- 複数のデータソースからの統合的判断が困難
- 「なぜ」「だからどうする」という推論の深度が不足

ARC-AGI-2 が測定する能力と知財業務の対応関係

ARC-AGI-2 が要求する認知能力と知財業務における具体的タスクの対応関係を以下に整理する。

ARC-AGI-2 の認知能力	知財業務における対応タスク	性能差の影響度
パターン認識と一般化	発明の本質抽出、クレーム起案	極めて高い
複合的推論（compositional reasoning）	進歩性判断（複数先行技術の組合せ）	極めて高い
グローバルなルール導出	発明の上位概念化、権利範囲の設定	高い
マルチステップ変換	侵害分析の論理構築	高い
空間的推論	図面解析、構造的特許の理解	中程度
最小限の事前知識での問題解決	未知技術分野の迅速な理解	高い

ARC-AGI-2 で測定される流動的知性は、蓄積された知識（結晶性知性）とは異なり、未知の問題に対する適応力を表す。知財業務では、発明はそれぞれ固有であり、個々の案件に対して新たな推論を行う必要があるため、この区別は極めて重要である。[9]

2026 年の知財業務における AI 活用の実態

エージェンティックワークフローの台頭

2026 年は、AI エージェントが知財業務のワークフローを自律的に実行する「エージェンティックワークフロー」の元年とされている。これは単なる補助ツールから、人間の監督下で自動化されたプロセスを実行する AI への転換を意味する。高い推論能力を持つモデルほど、このエージェンティック環境で信頼できるタスク遂行が可能となる。[19]

生成 AI は 2 つの波で知財オペレーションを変革しつつある：弁護士の作業成果を向上させる生成 AI（第 1 波）と、コアワークフローを自動化するエージェンティック AI（第 2 波）である。ARC-AGI-2 で高いスコアを持つモデルは、第 2 波のエージェンティック AI としての信頼性がより高い。[20]

効率性メトリクスの重要性

ARC-AGI-2 は、正確性だけでなく**コスト・パー・タスク**という効率性指標も組み込んでいる。知財業務においても、AI モデルの推論能力と計算コストのバランスは実務上の重要な判断基準である。高性能モデルが必ずしもすべてのタスクに最適とは限らず、タスクの複雑度に応じたモデル選択が戦略的に重要となる。[2][7]

実務的な示唆と推奨事項

ARC-AGI-2 の性能差が知財業務に与える影響を踏まえ、IP 専門家には以下のアプローチが推奨される。

- **タスク別モデル選択**：定型的な文書作成や翻訳には標準的なモデルを使用し、発明の本質抽出・クレーム設計・新規性判断などの抽象的推論を要するタスクには ARC-AGI-2 高性能モデルを割り当てる
- **推論能力のベンチマーク活用**：ARC-AGI-2 スコアを、知財業務に使用する AI モデルの選定基準の一つとして採用する。ただし、ARC-AGI-2 は視覚的パズルベースであるため、言語的推論能力を測る他のベンチマーク（HLE、SWE など）との複合的評価が望ましい[21]
- **人間と AI の協調設計**：ARC-AGI-2 で最高性能のモデルでも人間平均の 60%に到達したのは比較的最近であり、知財業務の最終判断には依然として人間の専門知識が不可欠である。AI を「自律的な労働力」ではなく「プロフェッショナルパートナー」として位置づけるアプローチが現実的である[4][19]

- **AI 出力の検証体制**：特にクレーム起案や侵害分析など法的リスクの高いタスクでは、AI の推論プロセスの透明性と人間による検証を制度化する。シンボリックかつグラフ中心のアプローチは、検証可能な中間推論の段階的説明を可能にする[10]

結論

ARC-AGI-2 は、AI モデルの「流動的知性」—すなわち未知の問題に対する抽象的推論能力—を測定するベンチマークとして、知財業務における AI 活用の質的差異を予測する有力な指標となり得る。高性能モデルと低性能モデルの間で最も顕著な差が生じるのは、発明の本質抽出、クレームの上位概念化、複数先行技術に基づく進歩性判断、均等論を含む侵害分析といった、まさに知財専門家の「知的判断力」が問われる領域である。

2026 年現在、トップモデルの ARC-AGI-2 スコアが 60〜70%台に到達し人間平均に近づく中で、知財業務への AI 活用は「補助ツール」から「推論パートナー」への転換期を迎えている。IP 専門家には、ARC-AGI-2 のような推論能力ベンチマークを理解したうえで、タスクの性質に応じた最適なモデル選択と、人間の専門知識と AI の推論能力を組み合わせたハイブリッドワークフローの設計が求められる。[3][4]

References

1. LLMs Hit 0% on ARC-AGI-2 benchmark - Introduction to ARC-AGI-2 benchmark The ARC-AGI-2 benchmark—an evolution within the Abstraction and ...
2. ARC-AGI-2 - ARC-AGI-2 - the next iteration of the benchmark - is designed to stress test the efficiency and capa...
3. GPT-5.2 & ARC-AGI-2: A Benchmark Analysis of AI Reasoningintuitionlabs.ai › articles › gpt-5-2-arc-agi-2-benchmark - An in-depth analysis of OpenAI's GPT-5.2 achieving a 54% score on the ARC-AGI-2 benchmark for abstra...
4. Claude Opus 4.6 が ARC-AGI-2 で 68.8% - 2026 年 2 月 5 日に発表された Anthropic 社の最新 AI モデル Claude Opus 4.6 は、100 万トークンという膨大な情報を一度に処理できる記憶容量を持ち、複数の AI が協調するエージェントチ...
5. ARC-AGI-2: A New Challenge for Frontier AI Reasoning ... - The Abstraction and Reasoning Corpus for Artificial General Intelligence (ARC-AGI), introduced in 20...
6. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems
7. ARC-AGI In 2026: Why Frontier Models Still Don't Generalize - ARC-AGI-2 exposes the real gap: generalization efficiency under budget constraints, where refinement...
8. ARC-AGI v2 Benchmark: Complete Leaderboard & Performance Analysis (2025) - Comprehensive ARC-AGI v2 benchmark results comparing 3+ AI models from 3 organizations. Top performe...
9. Do AI Reasoning Models Abstract and Reason Like Humans? - Going beyond simple accuracy for evaluating abstraction abilities

10. ARC-AGI: Benchmarking Abstraction in AGI - ARC-AGI is a comprehensive benchmark suite designed to evaluate human-level abstraction and composit...

11. Can Large Language Models Generate High-quality Patent ...

12. [PDF] Can Large Language Models Generate High-quality Patent Claims?

13. Multi-dimensional Evaluation of LLM-Generated Patent Claims - arXiv

14. Can AI Examine Novelty of Patents? - This paper introduces a novel challenge by evaluating the ability of large language models (LLMs) to...

15. PatentMind: A Multi-Aspect Reasoning Graph for Patent ... - Y Yoo 著 · 2025 · 被引用数: 5 — We introduce PatentMind, a novel framework for patent similarity assessme...

16. Uncover Weak Patent Claims with AI Invalidator LLM - Discover how AI transforms patent invalidation. Explore Invalidator LLM's claim analysis, semantic s...

17. Patent Claim Analysis for Novelty and Inventive Step - Understand key legal principles and analyses for evaluating novelty and inventive step in patents. G...

18. AI 技術関連発明の特許出願及び権利行使

19. 2026 AI Predictions: Agentic Workflows Will Define IP Management - Artificial intelligence in IP has long sparked fears about job replacement, but 2026 tells a differe...

20. The Future of IP Operations: Integrating AI, Agents ... - AI is reshaping IP operations in two waves: generative AI that enhances attorney work product, and a...

21. The new Gemini Deep Think incredible numbers on ARC- ... - Opus 4.6 scored 30% higher on ARC-AGI2 than Opus 4.5 but actually regressed by scoring 1% less on SW...