ChatGPT

ARC-AGI-2基準でみる高性能LLMと低性能LLMの知的財産業務への影響差

エグゼクティブサマリ

ARC-AGI-2は、**「未知の抽象タスクへの一般化」**と**「効率（コスト）」**を同時に強く要求するベンチマークとして設計され、公開当初は「純粋なLLMは0%」という水準で、従来型のスケール則だけでは到達できない能力ギャップを示すものとして位置づけられた。[1] 一方で2025年末〜2026年初頭にかけ、OpenAI[2]のGPT-5.2（Thinking）がARC-AGI-2（Verified）で52.9%（Proで54.2%）を公表し、Google[3]のGemini 3 Deep ThinkはARC-AGI-2で84.6%（ARC Prize Foundation検証）を公表した。[4] これにより、知財実務でボトルネックになりがちな「複数ルールの同時適用」「文脈に応じたルール切替」「記号（用語）に意味を付与する解釈」といった“流動性知能”寄りの能力が、商用フロンティアモデルで急速に押し上げられた可能性が高い（ただしベンチマーク高得点＝法的に安全、ではない）。[5]

知財業務（特許調査、先行技術検索、明細書・クレーム作成、侵害・FTO、契約レビュー、eDiscovery等）における差分は、単純に「正解率が上がる」だけではなく、**(a) 抽象化・一般化に支えられた“判断の安定性”**と、**(b) 長文・RAG・ツール統合・ログ/監査といった“運用上の再現性”**の二層で現れる。ARC-AGI-2が示す難所（象徴解釈・合成推論・文脈規則適用）は、請求項解釈、要件充足の要素分解、引用例対応付け、契約条項の条件分岐、証拠の真正性・同一性の維持など、知財の中核プロセスに近い。[6]

ただし、**高性能LLMでも幻覚（hallucination）は残存**し、複数ターンや長文条件下では誤りが累積しやすい。HalluHard（法律・研究・医療・コーディングを含む950問の多ターン幻覚ベンチマーク）では、Web検索を用いた強構成でも幻覚が大きく残る旨が報告されており、知財のような高リスク領域では「出力を成果物と見なす」運用は危険である。[7] さらに日本弁理士会[8]は、生成AI出力の正確性は保証されず弁理士が確認責任を負うこと、ハルシネーションの存在、無資格事業者によるサービス提供と弁理士法第75条の関係を明示的に論じており、**人間の最終責任と監査可能性**を前提に設計すべきことが公的に補強される。[9]

結論として、実務導入は「高性能モデル一択」ではなく、**高性能モデル中心＋低性能モデル限定用途（多段のフェイルセーフ付き）**が合理的である。ARC-AGI-2が重視する“効率”の観点からも、全工程を高コスト推論に寄せるのではなく、タスクの危険度に応じた**動的ルーティング（マルチホーミング）**、RAGの適法・適切な実装、監査ログ・証拠保全、二重化レビューを組み合わせる設計が、品質とコストの両立に直結する。[10]

前提・定義

本レポートでは「高性能LLM／低性能LLM」を、**層A（ARC-AGI-2で測られる推論・一般化能力）**と、**層B（実務性能：長文コンテキスト、RAG適合性、ファインチューニング可否、監査・セキュリティ等）**に分けて定義する。

層A：ARC-AGI-2（推論・一般化能力）
ARC-AGI-2は、公開・準非公開・非公開の評価集合をもち、各タスクは少数例から規則を推定する形式で、`pass@2`（2回試行で正解なら正）を採用する。[11] また、ベンチマーク設計上の主要難所として、象徴解釈（Symbolic Interpretation）、合成推論（Compositional Reasoning）、文脈依存の規則適用（Contextual Rule Application）などが明示されている。[12] 人間については、評価集合の難易度校正のために400人超の

一般参加者による統制テストを行い、平均スコア60%が報告されている（各タスクは少なくとも2人が2回以内に解けるよう選別）。 12

層B：実務性能（運用上の可用性）
知財実務での性能は、推論力に加え、(1)長文処理と情報統合、(2)RAG/ツール統合、(3)データ管理と監査、(4)セキュリティ・プライバシー、(5)コストと速度、(6)知識更新性（カットオフ／検索）で決まる。 13

本レポートにおける「高性能／低性能」の便宜的レンジ（未指定の扱い含む）

下表は、要求仕様に沿って技術指標を列挙し、**公式に確認できる範囲は“出典付きで実数”**、確認できない範囲は**「未指定」**として整理する。モデル固有値は代表例であり、運用・設定・ツール有無で変動する。

指標	層A/層B	高性能LLM（想定レンジ）	低性能LLM（想定レンジ）	根拠・注記
精度（ARC-AGI-2 `pass@2`）	層A	**50–85%程度**（例：GPT-5.2 Thinking 52.9%、Gemini 3 Deep Think 84.6%） 4	**0–5%程度**（公開当初の公表例：純粋LLM 0%、o3-preview-low推定4%など） 14	ARC-AGI-2は`pass@2`採用で、当初「純粋LLM 0%」が明示。 14
人間基準（平均）	層A	上限目標（85%が目標として提示） 15	—	人間平均60%、2人以上で100%（パネル）などが提示。 14
推論速度（レイテンシ）	層B	**中〜遅**（例：GPT-5.2は“Medium”、Proは“Slowest”、Proは数分かかる場合） 16	**速〜中**（小型モデルで低レイテンシになりやすい）※数値は未指定	速度は実装・推論設定・同時実行で変動。公式に“some requests may take several minutes”等の記載あり。 17
コンテキスト長	層B	**128k〜400k**（例：GPT-5.2 400k、GPT-5.2 Chat 128k） 18	**未指定**（小型でも128kを持つ例あり）	層Aが低くても長文窓が長いモデルはあり、相関は弱い。参考としてLlama 3.1は128Kへ拡張が公式発表。 19
知識更新性（カットオフ／検索）	層B	**新しいカットオフ＋検索**（例：GPT-5.2の知識カットオフ2025-08-31、Web検索ツール対応） 20	**古いカットオフ／検索なし**（未指定）	知財は最新の法令・審決・審査基準・先行例が重要で、検索・引用設計が必須。 21
ファインチューニング適性（可否/手法）	層B	**モデルにより不可の場合あり**（例：GPT-5.2はFine-tuning “Not supported”） 22	**可能な場合が多い（オープンウェイト等）**（例：BedrockでLlama 3.1のFTが提供/告知） 23	日本の著作権整理ではLoRA等の追加学習やRAGのベクトルDB作成が論点化される。 24

指標	層A/層B	高性能LLM（想定レンジ）	低性能LLM（想定レンジ）	根拠・注記
RAG適合性（ツール/引用/整合）	層B	**高**（例：GPT-5.2はFile search等ツールをサポート） [20]	**未指定**（実装次第）	RAGは著作権・契約・秘密管理と不可分。RAG実装時の著作物利用は国内で論点整理がある。 [25]
説明可能性（説明の一貫性・根拠提示）	層A/層B	**高められるが未保証**（例：GPT-5.2で“concise reasoning summaries”等が提供） [26]	**低〜中**（もっともらしい説明で誤りを隠すリスク）	説明そのものが“正しい根拠”である保証はない（評価には根拠照合が必要）。 [27]
hallucination率	層A/層B	**低減傾向**（例：GPT-5.2は誤りが相対38%減と説明） [28]	**高まりやすい**（未指定）	多ターン条件下では最良構成でも幻覚が大きく残る報告（約30%）。 [29]
セキュリティ・プライバシー	層B	**対策の幅が広いが、攻撃面も増える**（ツール統合・外部接続・ログ） [30]	**閉域実行で守れる場合も**（ただし品質が不足し得る）	NISTは来歴（provenance）追跡やエラー記録など、リスク管理行為を体系化。 [31]

（補足）3つの添付レポートは、ARC-AGI-2スコアを知財タスクに接続する観点（象徴解釈＝クレーム解釈等、マルチホーミングによるルーティング等）を含む二次整理として参照した。

主要IPタスク分類と成果差の比較表

知財領域は、技術理解と言語生成（明細書・クレーム）に加えて、法的推論（要件充足・侵害類否・証拠能力）と情報探索（先行技術・引用例・公知例）を統合する「複合ドメイン」であるため、一般のNLPタスクより“誤りのコスト”が高い。 [32] IPBenchは、知財タスクの体系（taxon）化と、理解・生成を含む多数タスクでの評価を提示し、最良モデルでも完全ではないことを示している。 [33] MoZIPも多言語・IP特化の評価で、強力な汎用モデルでも合格水準に届かない課題を指摘する。 [34] また、特許実務に直結する評価枠として、クレーム生成を“法的・構造的”に測るPatentScore（先行詞基準等を含む）が提案されている。 [35]

タスク別：高性能LLM vs 低性能LLMの期待成果差（具体例付き）

下表は、各タスクで**正確性・網羅性・速度・コスト・人的レビュー負荷**がどう変わるかを、実務に寄せて比較した（H=高性能、L=低性能）。

タスク	主要アウトプット	正確性（H/L）	網羅性（H/L）	速度（H/L）	コスト（H/L）	人的レビュー負荷（H/L）	具体例（差が出る場面）
特許調査（技術動向/俯瞰）	分類・出願人・トレンド要約	H: 主要論点の抽象化が安定 / L: 表層要約で誤解混入	H: “抜けの理由”を説明しやすい / L: 抜けが自覚されにくい	H: 中 / L: 速	H: 中 / L: 低	H: 中 / L: 中〜高	JPO統計等の外部根拠を添えると品質が上がる。日本のAI関連出願は2023年約11,400件など、一次統計の参照が重要。 36
先行技術検索（検索式/概念展開）	検索クエリ、同義語・上位概念、候補リスト	H: 概念対応が強い / L: キーワード一致に偏る	H: 類義概念の展開が広い / L: 漏れが増える	H: 中 / L: 速	H: 中 / L: 低	H: 中 / L: 高	「電磁パルスによる雑草制御」を「エネルギー場を用いた非化学的管理」へ抽象化して探索できるかで差。
特許明細書作成（ドラフト）	課題/手段/作用効果、実施形態	H: 論理一貫性が上がる / L: “もっともらしい虚構”が混入	H: 実施形態のバリエーション提示 / L: テンプレ生成に偏る	H: 中 / L: 速	H: 中 / L: 低	H: 高（技術/法の整合レビュー必須） / L: 非常に高	実施可能要件に関わる“架空データ/動作不可能”を混ぜる失敗。
クレームドラフティング	独立/従属請求項、用語定義	H: 要素分解と依存関係が保ちやすい / L: 先行詞基準・整合が崩れやすい	H: 回避設計（フォールバック）提案が可能 / L: 形だけ整う	H: 中 / L: 速	H: 中 / L: 低	H: 高 / L: 非常に高	先行詞基準（例：「第1通信部」→従属で突然「送受信部」）の破綻が低性能で多発しやすい。
侵害分析（クレームチャート）	要素対比表、充足/非充足理由	H: “文脈規則適用”が強く安定 / L: 表層一致で誤判定	H: 反論/非充足論点も提示 / L: 片側ストーリーに偏る	H: 中〜遅 / L: 速	H: 中 / L: 低	H: 高 / L: 非常に高	要素対応は「複数ルール×文脈」で、ARC-AGI-2が苦手とした領域に近い。 12

タスク	主要アウトプット	正確性（H/L）	網羅性（H/L）	速度（H/L）	コスト（H/L）	人的レビュー負荷（H/L）	具体例（差が出る場面）
FTO（自由実施）調査	侵害リスク仮説、検索・例外整理	H: 仮説生成→検証ループを回せる / L: 探索設計が脆い	H: 例外・周辺権利も含めた整理 / L: 漏れが顕著	H: 中〜遅 / L: 速	H: 中〜高 / L: 低	H: 非常に高 / L: 非常に高	「漏れ」のコストが最大級。人間の二重化と検索ログが必須（後述）。[37]
意匠・商標調査	類否理由、類似範囲の説明	H: 視覚・概念の統合が改善し得る / L: 表層類似に偏り	H: 反例の提示 / L: 抜けが多い	H: 中 / L: 速	H: 中 / L: 低	H: 中〜高 / L: 高	“類似の理由付け”の説明はできても、根拠ソースと審査基準に紐づけないと危険。[29]
契約レビュー（知財条項）	リスク抽出、修正文案	H: 条件分岐・例外の扱いが安定 / L: 条項間整合が崩れる	H: 関連条項を横断 / L: 部分最適化	H: 中 / L: 速	H: 中 / L: 低	H: 高 / L: 高	リーガルテック提供と弁護士法72条の関係はケース判断であり、表示・機能・事件性等が論点（サービス設計に直結）。[38]
ライセンス交渉支援	論点メモ、条件案、BATNA仮説	H: 多視点の案出し / L: 典型条項の焼き直し	H: 条件トレードオフ提示 / L: 交渉構造を誤解	H: 中 / L: 速	H: 中 / L: 低	H: 中〜高 / L: 高	交渉は“文脈規則適用”が中心。根拠のない相場提示は危険（幻覚×権限逸脱）。[39]
要約・翻訳	抄録、要旨、対訳	H: 長文でも整合しやすい / L: 重要要素の脱落	H: 用語統一しやすい / L: 用語揺れ	H: 中 / L: 速	H: 中 / L: 低	H: 低〜中 / L: 中	低リスク工程に見えるが、クレーム用語の翻訳揺れは後工程で重大化（ゲート設計が必要）。[35]

タスク	主要アウトプット	正確性（H/L）	網羅性（H/L）	速度（H/L）	コスト（H/L）	人的レビュー負荷（H/L）	具体例（差が出る場面）
証拠保全・eDiscovery	収集方針、ログ、同一性証明	H: 手順の抜け検出 / L: “それっぽい手順”で不備	H: 監査要件と連動 / L: ログ欠落	H: 中 / L: 速	H: 中 / L: 低	H: 高 / L: 高	デジタル・フォレンジックはAs-isでの収集・取得・保全が基盤で、クラウド/ログ範囲が拡大。 [40]

（表の解釈）知財タスクは、**「生成」よりも「根拠に基づく対応付け（claim↔evidence）」**の比重が大きい。新規性評価データセット研究では、生成モデルが一定精度で“関係理解の説明”を生成できる可能性が示されるが、説明の正しさは別途検証が必要である。 [41]

リスク評価

性能差は「便利さ」だけでなく、リスクの**種類**と**増幅の仕方**を変える。低性能は“露骨に間違う”傾向があり検出しやすい一方、高性能は“もっともらしく整合した間違い”になりやすく、レビューが形式チェックに堕ちると見逃しやすい（高リスク領域で危険）。 [21]

性能別リスクの整理（失敗モード・影響度・監査性）

観点	低性能LLM：典型失敗モード	高性能LLM：典型失敗モード	影響度（知財）	検出難度	主要コントロール（要点）
誤情報（hallucination）	架空の先行例・条文・判例・審決を生成／引用を捏造	実在情報を“誤って”結合し、整合的な虚偽ストーリーを生成（多ターンで誤りが累積） [29]	極大（FTO/侵害で致命傷）	低→中（露骨）	**根拠必須**（RAG＋原文引用範囲＋出典ID）、自己一致テスト、別系統モデルで反証探索
機密漏洩（営業秘密/未公開出願）	そのまま外部に投げる／プロンプトに機微を含める	ツール統合（検索/コネクタ）で攻撃面が増える、プロンプト注入の踏み台	極大（秘密喪失、先使用権/新規性問題）	中	データ分類・マスキング、閉域RAG、許可ツールリスト、監査ログ、ゼロトラスト
バイアス（判断の偏り）	典型条項・特許実務の“平均”に寄り過ぎる	高度に説得的だが、特定法域/業界慣行へ過適合	大（交渉・拒絶対応で不利）	中	法域・顧客ポリシーを構造化入力、反対仮説生成、評価指標の分解（属性別） [42]

観点	低性能LLM：典型失敗モード	高性能LLM：典型失敗モード	影響度（知財）	検出難度	主要コントロール（要点）
法的責任（最終判断）	“AIが言った”で根拠なき処理	“高精度だから大丈夫”という自動化バイアス	極大（職責・善管注意義務）	中	弁理士が正確性確認し責任を負うことを明示、AI出力の位置づけを規程化 [43]
証拠能力（eDiscovery/紛争）	手順・ログ欠落、再現不能	出力改変・来歴不明で真正性が争点化	大〜極大	高	As-is収集・ハッシュ・作業ログ、クラウド特有手順、保存期限の設計 [44]
監査可能性（誰が/何を根拠に）	生成過程がブラックボックスになりがち	ツール連携/多段推論でも、設計すれば来歴追跡が可能	大	中	来歴（provenance）追跡、エラー/ニアミス記録、メタデータ管理による追跡 [31]

日本の公的・準公的ガイドラインが示す「リスクの境界」

- 著作権面では、RAG等で既存著作物を用いる場合、**ベクトルDB作成（複製）**や**出力時の“軽微利用”**などが論点となり、一定の場合には許諾が必要という整理が示されている。 [45]
- クリエイター保護の観点では、robots.txt等でクローラ収集を一定程度抑止し得る、ID/パスワード領域での保護等が例示されている。 [46]
- 知財実務（弁理士）では、生成AI出力が正確である保証はなく、弁理士が確認し責任をもって提供すべきことが示される。 [43]
- 無資格事業者がAIで出願書類作成等を提供する場合の弁理士法75条該当性は、実質判断で違反になり得る旨が整理される。 [47]
- 契約関連サービスについては、法務省 [48] が弁護士法72条との関係で、要件（報酬目的、事件性等）に基づく枠組みでの判断を示し、生成AIサービスにも原則同様の枠組みを適用すると整理している。 [38]

運用・導入指針

運用設計は「モデル選定」より先に、**評価指標（KPI）→検証プロトコル→HITL→ログ/データ→セキュリティ→コスト**の順で設計すると破綻しにくい。これはARC-AGI-2が“能力×効率”で測るのと同型であり、知財では特に**監査と再現性**が致命的に重要である。 [49]

KPI設計（例）

KPIは、**成果（アウトカム）KPI**と、**安全・再現性（ガバナンス）KPI**を分離する。

アウトカムKPI例（タスク別に設定）
- 先行技術検索：Recall@K（K=20/50/100）、検索式の概念展開率、重要引用例の再現率（審査引用の再捕捉） [50]
- クレーム：要素分解の妥当性、先行詞基準違反率、従属関係整合率（PatentScore型の構造指標） [35]
- 契約レビュー：リスク抽出の再現率（人手基準）、修正文案採用率、見落とし率（重大条項） [38]

- FTO/侵害：重大誤り率（False negative/positive）、争点の網羅率、反証提示率（counter-argument rate） [21]

ガバナンスKPI例
- Unsupported claim rate（根拠なし主張比率）／Citation accuracy（引用が根拠を支持する率） [7]
- ログ完全性（入力・モデルID・プロンプト・検索クエリ・取得文書ID・出力・人手修正履歴） [51]
- 来歴（provenance）追跡率（文書・出力・編集の追跡可能性） [52]
- 機密自動検知の検出率・誤検知率（DLP）
- インシデント（誤情報/漏洩/誤出願等）の件数・影響度・是正リードタイム [53]

検証プロトコル（公開/準非公開/非公開セット）

ARC-AGI-2が採用する**Public / Semi-Private / Private**の三層構造は、知財AI評価にも転用しやすい。ARC側は、公開120問・準非公開120問・非公開120問、さらにPublic trainingを分け、難易度校正（IDD）を人間テストで担保する設計思想を明示している。 [11]

知財向けの推奨構成（例）
- 公開セット：公開特許・公開判例・公開契約雛形のみ（社外開示可能）。回帰テスト用。
- 準非公開セット：過去案件の匿名化・要約化（機密除去）＋第三者に露出し得る資料（例：API経由で学習済みの可能性がある公開情報）
- 非公開セット：社内案件（未公開出願、交渉中契約、鑑定・警告等）を、アクセス制御下で評価（モデル更新時の最終ゲート）

HITL（Human-in-the-Loop）設計：不確実性ゲート＋二重化

日本弁理士会は、生成AIの生成物は正確性が保証されず、弁理士が確認し責任をもって提供すべきことを明言しているため、HITLは“推奨”ではなく“前提条件”に近い。 [43]

実装上の要点
- 不確実性ゲート：
- 重要主張に根拠未提示／引用不整合／自己矛盾／要素分解の未確定がある場合は自動で“要レビュー”へ
- 二重化：
- 高リスク（FTO/侵害/契約）では、(a)別モデルで反証探索、(b)別担当者レビュー、(c)根拠原文突合を必須化
- 役割分担：
- 低性能モデル＝整形・要約・翻訳・抽出（低リスク）
- 高性能モデル＝要素分解・仮説生成・反証生成（高リスク）
- 人間＝法的評価・最終判断・顧客説明・署名責任

データ管理・ログ保持・証拠能力

- eDiscovery/証拠保全では、デジタル・フォレンジックが「As-is」での収集・取得・保全を基盤とし、クラウド特有の証跡・ログ範囲拡大に対応する必要があることが整理されている。 [44]
- 生成AIリスク管理として、NISTの生成AIプロファイルは、デジタルコンテンツの来歴追跡（provenance metadata、watermark、metadata recording等）や、エラー/ニアミスの記録・追跡、構造化フィードバックの記録などを具体アクションとして提示する。 [54]

セキュリティ対策（知財向けの最小セット）

総務省 [55] と経済産業省 [56] によるAI事業者ガイドラインは、AIガバナンスを“why/what/how”で整理し、チェックリストや仮想事例も含める構成を示す。 [57]

知財用途の最小セットは以下（要点のみ）
- データ分類（公開/社外秘/極秘）と投入可否ルール
- RAGデータの権限管理（文書単位ACL、案件単位隔離）
- 外部ツール／コネクタの許可リスト（デフォルト拒否）
- プロンプト注入対策（取得文書の命令文無効化、システムプロンプト固定、検知）
- 監査ログ（改ざん耐性、保持期間、アクセス証跡）

コスト見積もりの考え方（トークン×単価＋検証コスト＋人件費）

コストは、(1)モデル推論、(2)検索/RAG、(3)検証（自動評価・別モデル・原文突合）、(4)人件費、(5)インシデントコストで見積もる。ARC-AGI-2も“コスト/タスク”を主要指標に組み込み、能力だけでなく効率を重視する。 12

代表例として、GPT-5.2のAPI価格は入力$1.75/100万トークン、出力$14/100万トークン（キャッシュ入力は90%割引）、GPT-5.2-proは入力$21/100万、出力$168/100万である。 58
したがって「入力2万・出力5千」規模の一回処理は、**GPT-5.2で約$0.105、Proで約$1.26**がトークン費の目安となり、ここに検証（追加推論）とレビュー工数が上乗せされる（実務では“工程総量”で逆算する方が安定）。 16

推奨アーキテクチャとワークフロー例

基本方針は「高性能モデル中心」だが、**全工程を高性能に寄せない**。ARC-AGI-2が示す通り、効率は知能の一部であり、知財では“正確性を担保するための二重化”も必要なため、工程ごとに最適化する方が総コストは下がりやすい。 59

添付資料でも、推論コストと要求される論理深度に応じて複数モデルを使い分ける「マルチホーミング」と動的ルーティングが提案されている。　これを実務向けに、**フェイルセーフ（停止・差戻し・人間介入）**を明示して設計すると、法的・監査的に説明しやすい。 60

```
flowchart TD
  A[案件投入: 文書/図面/契約/製品仕様] --> B{データ分類\n公開/社外秘/極秘}
  B -->|極秘| C[閉域RAG: 案件DB\nACL/暗号化/監査ログ]
  B -->|公開/社外秘| D[標準RAG: 公開DB + 許諾DB]

  C --> E{タスク判定\n(先行技術/クレーム/侵害/FTO/契約/eDiscovery等)}
  D --> E

  E --> F{リスクゲート\n高リスク?}
  F -->|低リスク| G[低性能LLM\n抽出/要約/翻訳/整形]
  F -->|高リスク| H[高性能LLM\n要素分解/仮説生成/反証生成\n長文統合]

  G --> I[構造化出力\n(表/箇条/差分)]
  H --> J[根拠付き出力\n(引用ID/原文抜粋/推論要約)]

  I --> K{自動検証\n・引用整合\n・用語統一\n・形式要件}
  J --> K

  K -->|NG| L[差戻し\n追加検索/追加根拠/別モデル反証]
  K -->|OK| M[HITLレビュー\n二重化(必要時)]
  M --> N{承認?}
```

```
  N -->|否| L
  N -->|是| O[成果物化\n(ドラフト/チャート/メモ)]
  O --> P[ログ・来歴保全\n入力/出典/モデルID/編集履歴]
```

設計の鍵は、(1)RAGを“検索”で終わらせず**引用整合性**まで検証すること、(2)監査ログを“残す”だけでなく**来歴追跡可能（provenance）**な形にすること、(3)高リスクでは**反証探索の二重化**を必須工程に組み込むことである。 61

実務導入チェックリストとロードマップ

導入チェックリスト（要点）

ガイドラインを踏まえた「最低限」の観点は、(A)職責・適法性、(B)著作権・データ、(C)品質評価、(D)セキュリティ、(E)監査・証拠の5群に整理できる。 62

- 職責・適法性：弁理士がAI生成物の正確性確認と責任を負う運用になっているか。無資格事業者スキームが弁理士法75条に抵触しない設計か（表示・機能・対価の実質）。 63
- 著作権・データ：RAGのベクトルDB作成・出力が「軽微利用」等の整理に沿うか、許諾が必要なデータは契約で確保できているか。 64
- 品質評価：公開/準非公開/非公開セットでの回帰テスト、Recall@Kや先行詞基準違反率等のKPIが定義されているか。 65
- セキュリティ：データ分類、閉域RAG、許可ツールリスト、DLP、プロンプト注入対策、アクセス監査が実装されているか。 30
- 監査・証拠：入力・出典・出力・編集履歴の真正性が担保でき、eDiscovery/紛争時にAs-is保全とログ提示が可能か。 66

導入ロードマップ（短期・中期・長期）

```
timeline
  title 知財AI導入ロードマップ
  短期 : 対象業務の棚卸しとリスク分類(高/中/低)
       : 公開データで評価基盤(KPI/回帰テスト)を整備
       : 低リスク工程(要約/翻訳/整形)を低性能LLMで先行導入
       : 監査ログ設計(入力/出典/モデルID/編集履歴)
  中期 : 閉域RAGと権限管理(案件分離/ACL/暗号化)を実装
       : 高リスク工程(クレーム要素分解/引用例対応/契約条項抽出)を高性能LLM中心に段階導入
       : 不確実性ゲートと二重化(反証探索+二者レビュー)を組込み
       : 著作権・許諾データ整備(契約/表示/軽微利用の運用ルール)
  長期 : FTO/侵害/交渉支援で多段エージェント化(ツール統合)
       : 来歴(provenance)追跡と証拠保全の標準化(組織横断)
       : 継続監査(インシデント対応/モデル更新/第三者評価)を運用に内蔵
       : 人材育成(新人向けシミュレーション/レビュー基準の形式知化)
```

（補足）知財実務は、AI導入により“作業の圧縮”が進む一方、レビュー基準の形式知化と教育設計が追いつかないと、スキル空洞化・責任所在の曖昧化が起きやすいという論点が添付資料でも論じられている。

参考文献・参照ソース

一次（公式/公的）を優先し、補助的に主要学術論文と、ユーザー提供資料（添付PDF）を用いた。

ARC-AGI-2・ARC Prize（公式）
- ARC-AGI-2の設計、データ構造、難所（象徴解釈等）、人間平均60%、効率指標の導入。 12
- 公開当初のベースライン（純粋LLM 0%、人間パネル、コスト/タスク等）と `pass@2` の説明。 67
- ARC Prize 2025技術報告（人間可解性、2025結果：私有セット24%・$0.20/タスク等）。 68

公式モデル仕様
- GPT-5.2のARC-AGI-2（Verified）スコア、事実性改善の説明、価格。 28
- GPT-5.2 API仕様（コンテキスト400k、出力上限、知識カットオフ、ツール、FT不可）。 69
- Gemini 3 Deep ThinkのARC-AGI-2 84.6%（ARC Prize Foundation検証）。 70

主要学術論文（知財×LLM評価）
- IPBench（知財タスク分類とベンチマーク）。 33
- MoZIP（多言語IPベンチ）。 71
- 新規性評価（請求項↔引用例対応に基づく評価）。 72
- PatentScore（クレーム生成の構造・法的観点評価、先行詞基準等）。 35
- HalluHard（多ターン幻覚、法領域を含む、約30%残存など）。 7

日本の公的・準公的ガイドライン
- 文化庁（AIと著作権、robots.txt等の技術的措置例、RAG/LoRA等の論点）。 73
- 内閣府（AI時代の知的財産権の論点整理：RAGの軽微利用等）。 74
- 総務省・経産省（AI事業者ガイドライン：ガバナンス構成）。 75
- 法務省（契約書関連業務支援サービスと弁護士法72条）。 38
- 日本弁理士会（弁理士業務AI利活用ガイドライン、弁理士法75条整理）。 9
- デジタル・フォレンジック研究会（証拠保全ガイドライン10版）。 44
- NIST（生成AIリスク管理プロファイル：来歴追跡、記録、インシデント等）。 54
- 特許庁（AI関連発明の国内出願動向：2023年約11,400件等）。 76

ユーザー提供資料（添付PDF）
- ARC-AGI-2と知財業務影響の整理（例示・論点）。
- ARC-AGI-2能力要素と知財タスク対応の整理。
- 知財タスクのマルチホーミング/ルーティング、クレーム構造欠陥などの失敗モード整理。

（図表・一次情報への導線：URL）
https://arcprize.org/arc-agi/2/
https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025
https://openai.com/ja-JP/index/introducing-gpt-5-2/
https://developers.openai.com/api/docs/models/gpt-5.2
https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-deep-think/

---

1 2 5 6 11 12 15 49 55 65 https://arcprize.org/arc-agi/2/
https://arcprize.org/arc-agi/2/

3 7 21 27 29 39 https://arxiv.org/abs/2602.01031
https://arxiv.org/abs/2602.01031

4 28 56 58 https://openai.com/ja-JP/index/introducing-gpt-5-2/
https://openai.com/ja-JP/index/introducing-gpt-5-2/

8 38 https://www.moj.go.jp/content/001400675.pdf
https://www.moj.go.jp/content/001400675.pdf

9 37 43 63 https://www.jpaa.or.jp/cms/wp-content/uploads/2025/04/AIservices-guideline.pdf
https://www.jpaa.or.jp/cms/wp-content/uploads/2025/04/AIservices-guideline.pdf

10 14 59 67 Announcing ARC-AGI-2 and ARC Prize 2025
https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025

13 26 Using GPT-5.2 | OpenAI API
https://developers.openai.com/api/docs/guides/latest-model/

16 18 20 22 69 https://developers.openai.com/api/docs/models/gpt-5.2
https://developers.openai.com/api/docs/models/gpt-5.2

17 https://developers.openai.com/api/docs/models/gpt-5.2-pro
https://developers.openai.com/api/docs/models/gpt-5.2-pro

19 https://ai.meta.com/blog/meta-llama-3-1/
https://ai.meta.com/blog/meta-llama-3-1/

23 https://aws.amazon.com/about-aws/whats-new/2024/10/metas-llama-31-8b-70b-models-fine-tuning-amazon-bedrock/
https://aws.amazon.com/about-aws/whats-new/2024/10/metas-llama-31-8b-70b-models-fine-tuning-amazon-bedrock/

24 46 73 https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/seisaku/r06_02/pdf/94089701_05.pdf
https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/seisaku/r06_02/pdf/94089701_05.pdf

25 45 48 64 74 https://www.kantei.go.jp/jp/singi/titeki2/chitekizaisan2024/0528_ai.pdf
https://www.kantei.go.jp/jp/singi/titeki2/chitekizaisan2024/0528_ai.pdf

30 57 75 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20250328_2.pdf
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20250328_2.pdf

31 42 51 52 53 54 60 61 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

32 33 https://arxiv.org/abs/2504.15524
https://arxiv.org/abs/2504.15524

34 71 https://arxiv.org/abs/2402.16389
https://arxiv.org/abs/2402.16389

35 https://arxiv.org/abs/2505.19345
https://arxiv.org/abs/2505.19345

36 76 https://www.jpo.go.jp/system/patent/gaiyo/sesaku/ai/document/ai_shutsugan_chosa/gaiyo.pdf
https://www.jpo.go.jp/system/patent/gaiyo/sesaku/ai/document/ai_shutsugan_chosa/gaiyo.pdf

40 44 66 https://digitalforensic.jp/wp-content/uploads/2025/04/shokohozenGL10.pdf
https://digitalforensic.jp/wp-content/uploads/2025/04/shokohozenGL10.pdf

41 50 72 https://arxiv.org/abs/2502.06316
https://arxiv.org/abs/2502.06316

47 62 https://www.jpaa.or.jp/cms/wp-content/uploads/2025/04/AIservices-article75.pdf
https://www.jpaa.or.jp/cms/wp-content/uploads/2025/04/AIservices-article75.pdf

68 ARC Prize 2025: Technical Report
https://arxiv.org/pdf/2601.10904

70 https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-deep-think/
https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-deep-think/