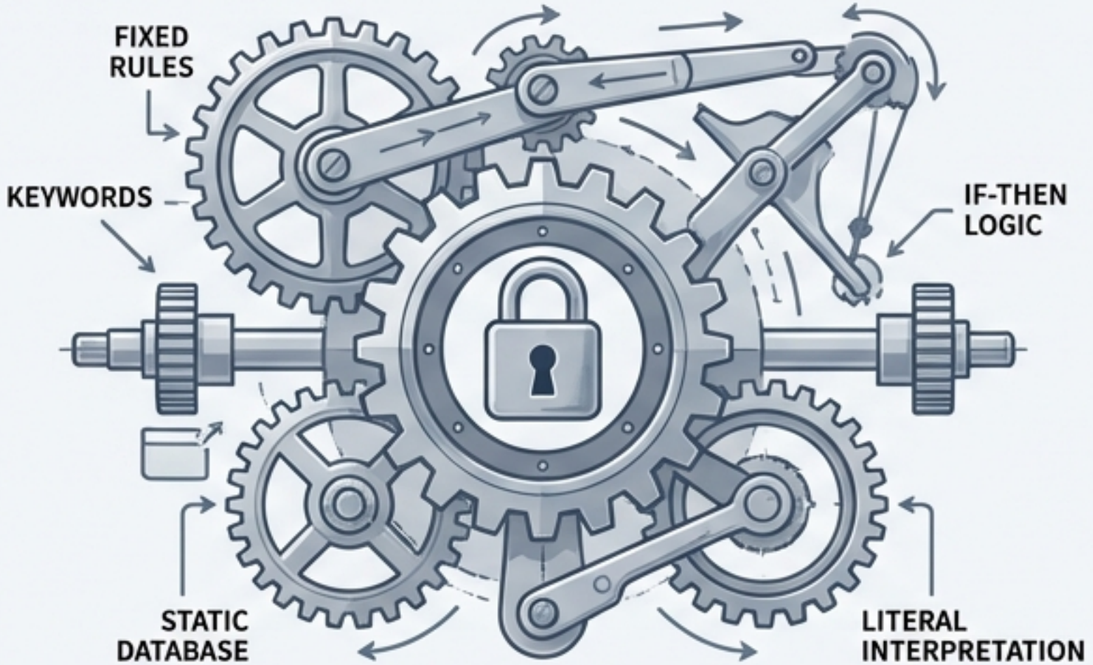
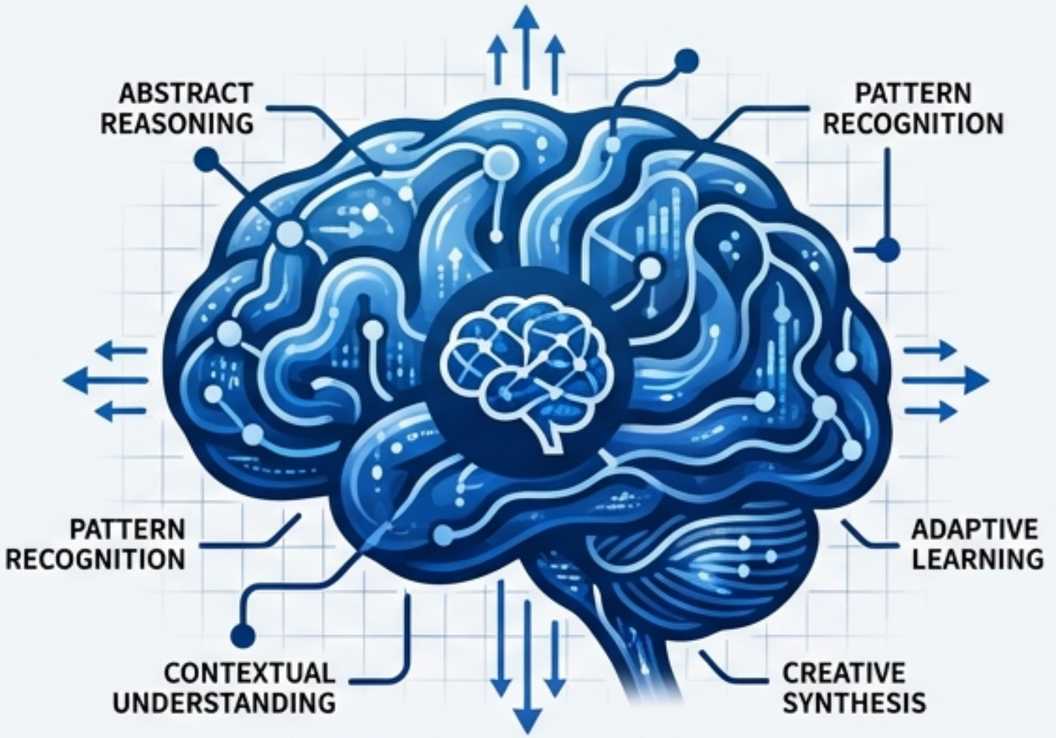
ARC-AGI-2基準が変える知財実務の「流動性知能」

ARC-AGI-2 Standards & The Future of IP Operations



従来型ルール (Rigid Rules)



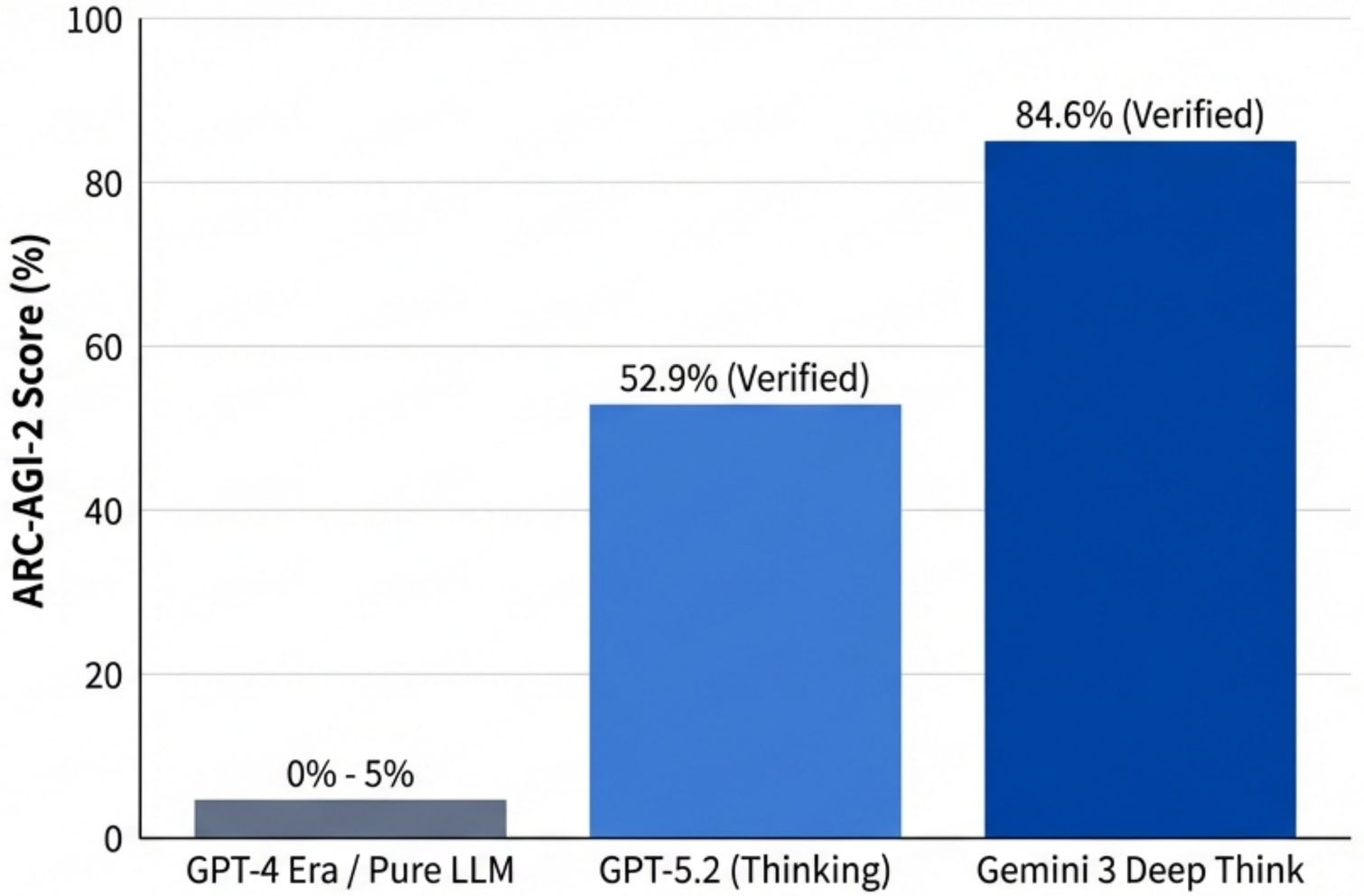
流動性知能 (Fluid Intelligence)

従来のLLMは「純粋な推論能力（ARC-AGI-2）」が0%に近く、知財業務はキーワードマッチングの延長にあった。

2025-26年、Gemini 3 Deep Think（84.6%）や GPT-5.2（52.9%）の登場により、AIは「未知の抽象タスク」を処理可能になった。

本資料では、高性能LLMと低性能LLMの決定的な能力差と、それを前提とした「ハイブリッドな知財実務アーキテクチャ」を提示する。

0%から85%への跳躍：ARC-AGI-2が示す性能ギャップ



ARC-AGI-2とは

「未知の抽象タスクへの一般化」と「効率（コスト）」を同時に測定するベンチマーク。

•

意味合い

従来型のスケール則だけでは到達できなかった「複数ルールの同時適用」や「記号への意味付与」が可能になったことを示す。

•

実務への影響

このスコア差は、単なる正解率の違いではなく、複雑な法的推論における「判断の安定性」の差として現れる。

「高性能」と「低性能」を定義する2つの層

Layer A: 推論・一般化能力 (The Brain)

定義：ARC-AGI-2（Pass@2）スコア

高性能 (High):

50–85%

- モデル: Gemini 3, GPT-5.2
- 能力: 象徴解釈・合成推論が可能

低性能 (Low):

0–5%

- モデル: 純粋LLM, o3-preview-low

Layer B: 実務性能 (The Utility)

定義：実務性能（長文、RAG、セキュリティ）

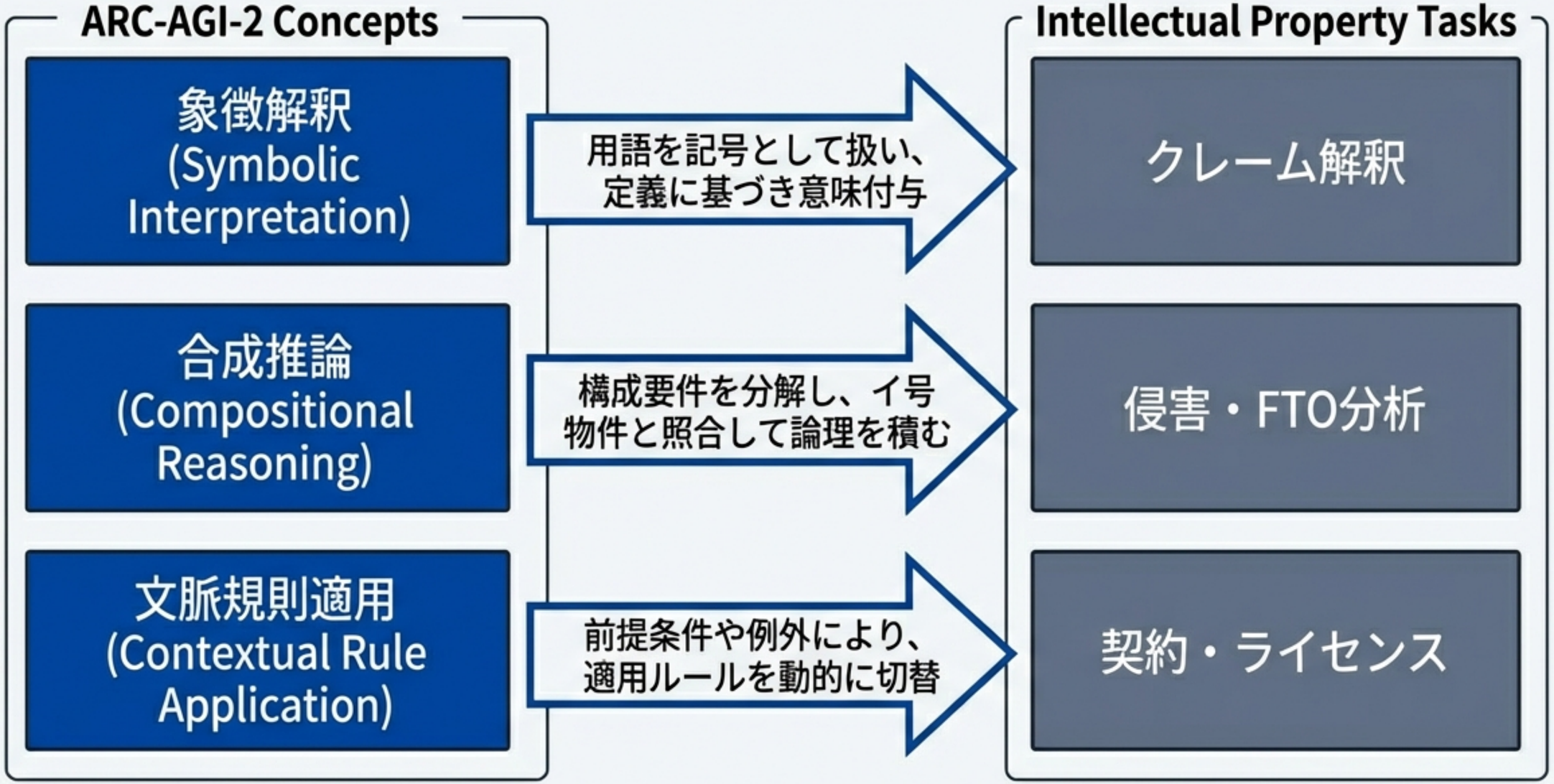
高性能 (High):

- Context: 128k-400k
- Knowledge Cutoff
- Knowledge Cutoff: 2025-08
- Tools: 統合あり (File Search etc)

低性能 (Low):

- RAG適合性不明、幻覚リスク高

抽象能力の知財タスクへの翻訳



タスク別影響差（1）調査・明細書作成

Task	Low Performance (0-5%)	High Performance (50-85%)
特許調査・先行技術検索（Patent Search）	キーワード一致に偏る。「電磁パルス」で検索し「エネルギー場」を見落とす。	概念対応が強い。「雑草制御」を「非化学的管理」へ抽象化して探索可能。Recall@Kが向上。
明細書作成（Drafting）	「もっともらしい虚構」が混入。実施不可能要件や架空データを生成するリスク。	実施形態のバリエーション提示。論理一貫性が高く、クレームの先行詞基準（例：「第1通信部」と「送受信部」の不整合）を防ぐ。

タスク別影響差（2）侵害分析・契約レビュー

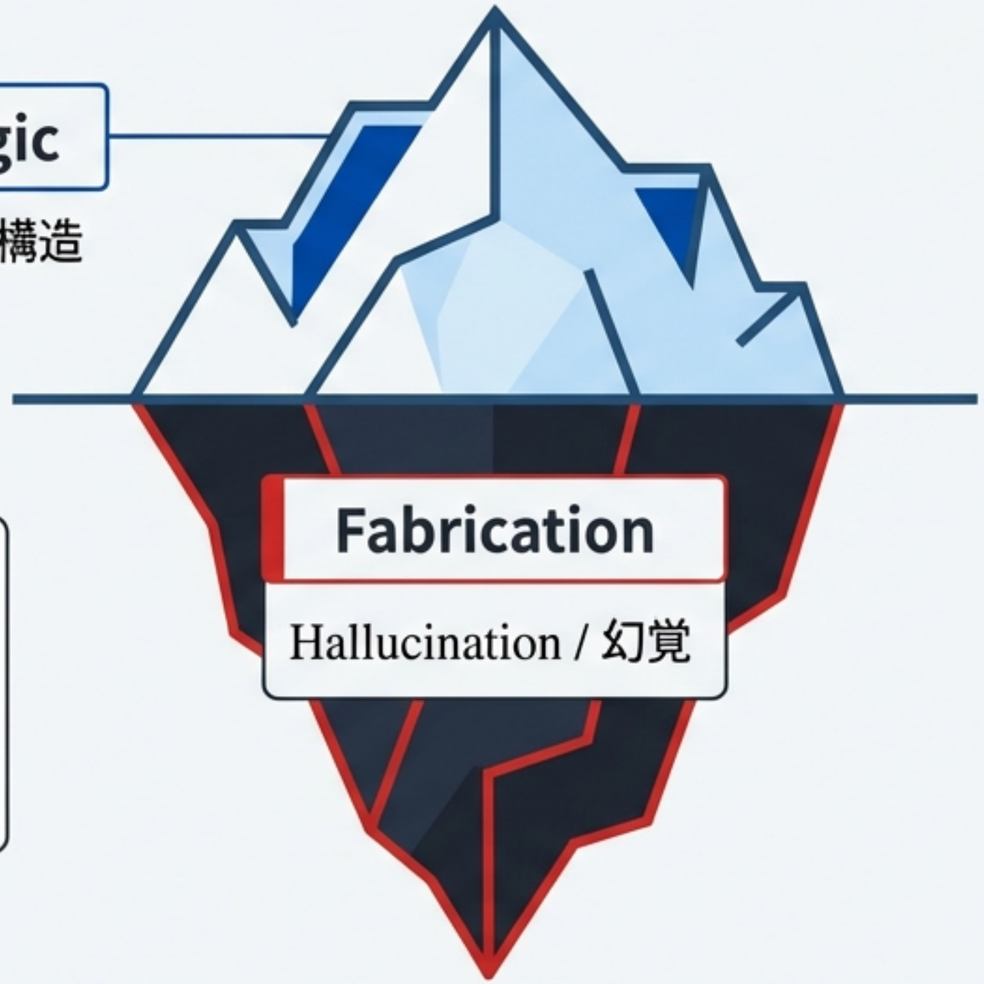
Task	Low Performance (0-5%)	High Performance (50-85%)
侵害分析（Infringement/Claim Chart）	表層的な文言一致で誤判定。片側のストーリーに偏る。	複数ルール×文脈の適用。反論や非充足論点も提示可能。
契約・FTO（Freedom to Operate）	条項間の整合が崩れる。部分最適化により、グローバルなクロスリファレンスを見落とす。	条件分岐・例外の扱いが安定。漏れのコストが最大級であるFTOにおいて、仮説→検証ループを回せる。

High Performance models overcome the ‘Contextual Rule Application’ deficit.

「もっともらしい嘘」という新たなリスク

Grammar & Logic

文法的に正しく、論理構造も完璧に見える。



1. HalluHardベンチマーク:
法律・医療含む難問において、高性能モデルでも約30%の幻覚（エラー）が残存。

2. リスクの質的変化:

Low Performance Risk:
露骨な間違い（架空の判例など）。検出しやすい。

High Performance Risk:
「整合的な虚偽」。実在情報を誤って結合し、形式チェックをする。FTOでは致命傷。

結論: 「出力をそのまま成果物とする」運用は危険。検証プロセスの設計が必須。

日本の法的ランドスケープと責任境界

弁理士法 第75条

信用失墜行為の禁止。
日本弁理士会ガイドライン：AI生成物の正確性は保証されず 、弁理士が確認責任を負う。**「AIが間違えた」は免責事由にならない。**

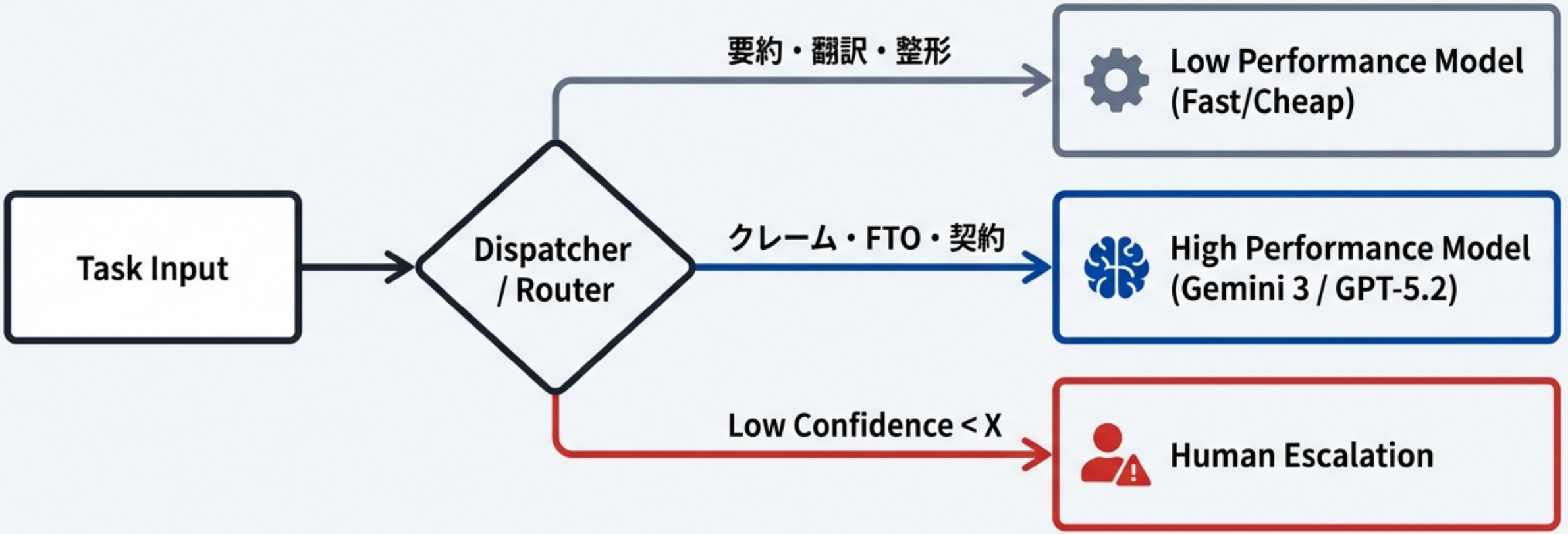
著作権法とRAG

ベクトルDB作成（複製）や出力時の「軽微利用」が論点。
RAGデータの権利処理と、契約による許諾確保が必要。

弁護士法 第72条

非弁活動の禁止。
リーガルテック提供における表示・機能・事件性が論点。**無資格事業者による鑑定的なAIサービスはリスクが高い。**

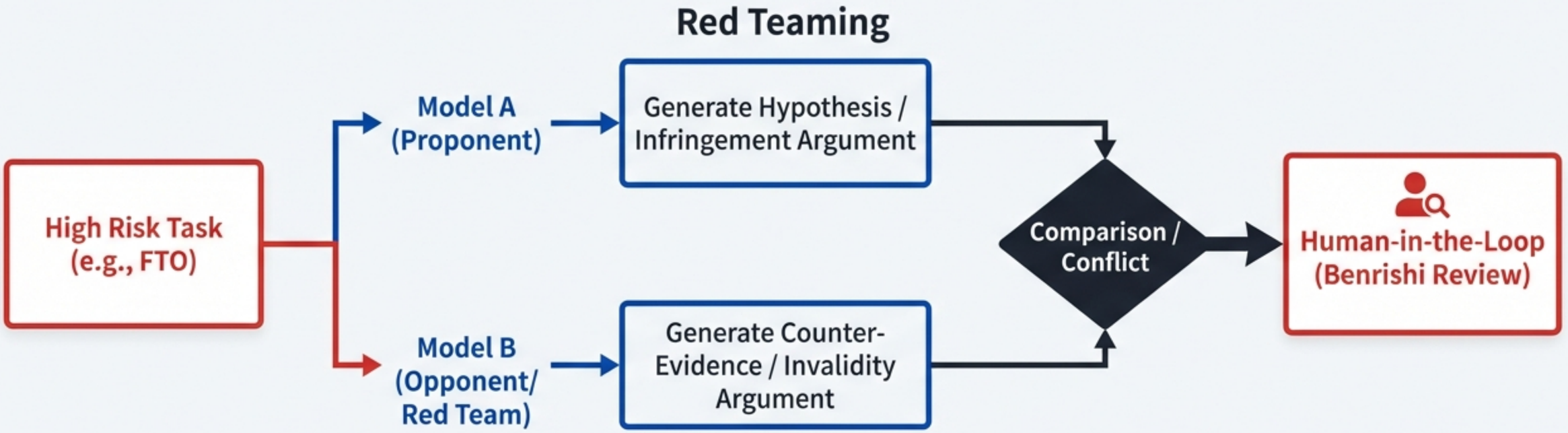
推奨アーキテクチャ：動的ルーティング



基本方針: 全工程を高性能モデルに寄せない「マルチホーミング」でコストと精度を最適化。

Swiss Legal Modern

ワークフロー設計：「二重化」とHITL



AIによるレッドチーミング:
一方向のハルシネーションを防ぐため、対立仮説を生成させる。

不確実性ゲート:
根拠未提示や自己矛盾がある場合、強制的にレビューへ。

データガバナンスとRAGの適法性

Noto Sans JP (Bold)

監査ログ:

「いつ、誰が、どのAIで」を記録し、事後検証に耐えうる状態にする。

コスト試算：トークン単価 vs 検証コスト

Total Cost = [Token Cost] + [Verification Cost] + [Risk Cost]

Token Cost

High Performance (GPT-5.2 Pro):
High ($168/1M tokens)

Low Performance:
Low Cost

Verification Cost (Human Time)

Low Perf Models → High Verification Cost (手直し工数増)

High Perf Models → Lower Verification Cost

Insight

知財実務では**専門家**（弁理士）の時給が高いため、**高精度モデル**で**手直しを減らす**方が**総コストは下がる。**

評価プロトコルとKPI設計

Outcome KPI (成果)

Recall@K (Search)

Structural Integrity
(PatentScore)

Governance KPI (安全)

Unsupported Claim Rate
(根拠なし主張率)

Provenance Tracking Rate
(来歴追跡率)

Testing Sets (Evaluation)

- Public Set: 回帰テスト用 (Regression)
- Semi-Private Set: 匿名化済み (Anonymized)
- Private Set: 実案件・最終ゲート (Real Cases)

実装ロードマップ：短期から長期へ

実務導入のための最終チェックリスト

☐	職責・適法性: 弁理士法75条に基づき、人間が最終責任を負うプロセスか？
☐	著作権・データ: RAG利用におけるデータ許諾と「軽微利用」の範囲整理。
☐	品質評価: PatentScoreやRecall@Kなどの定量的指標で評価しているか？
☐	セキュリティ: データ分類とプロンプト注入対策は実装済みか？
☐	監査・証拠: 「いつ、誰が、どのAIで、何を根拠に」出力したか追跡可能か？

知財AIの本質は「速度」ではなく、「監査可能性」と「再現性」にある。