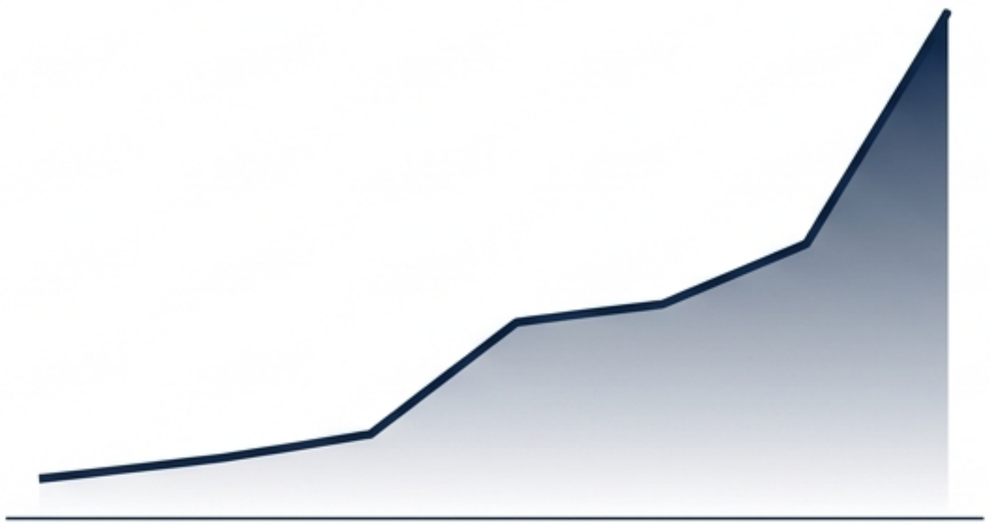
PatRe AI特許審査のパラダイムを「静的分類」から「動的攻防」へ

LLMによるOffice Action（拒絶理由通知）とRebuttal（反論）のフルステージ生成ベンチマーク

深圳先進技術研究院（SIAT）, 大連理工大学, UNSW Sydney による最新研究（2026-05）

限界を迎える人間の審査官と、現実と乖離した既存のAI評価

人間の限界：出願の爆発的増加



USPTOの未審査案件は837,928件（待ち時間20.5ヶ月）。LLMによる「特許自動生成」技術の台頭が、審査官の負荷をさらに悪化させている。

既存AIの欠陥：ベンチマークの構造的限界

これまでの研究（HUPD等）は、特許審査を「静的な二値分類」としてしか扱っていない。実際の審査で繰り返される「理由付けと反論の動的な対話プロセス」が完全に無視されている。

特許AIベンチマークの進化：PatReの立ち位置

Dataset	Task	Statute (法的根拠)	Evolution (クレーム進化)	Adversarial (マルチターン)	Full-stage (フルステージ)
HUPD (2023)	Discriminative	✗	✗	✗	✗
PANORAMA (2025)	Discriminative	✓	✗	✗	✗
Patent-CR (2025)	Generative	✗	✓	✗	Partial
PatRe (Ours)	**Generative**	✓	✓	✓	✓

PatReは、特許審査を「結果の予測」ではなく、「正当性の主張（Justification）の動的プロセス」として再定義した初のベンチマークである。

PatRe フレームワーク：法的主張と反論のマルチターン・シミュレーション

Step 1: 審査官タスク (OA Generation)

1. OA-DP (Zero-shot):
外部情報なしでの直接生成

2. OA-RO (Oracle):
正解の先行技術を付与した上限評価

3. OA-RS (Simulate):
BM25検索ノイズを含む現実的シミュレーション

Step 2: 出願人タスク (Rebuttal Generation)

OAの拒絶理由に対し、技術的特徴と法的枠組みに基づき、説得力のある反論を生成する。

表面的な一致を超えて：法的・技術的ニュアンスを捉える多次元評価

生成された法的文書
(OA / Rebuttal)

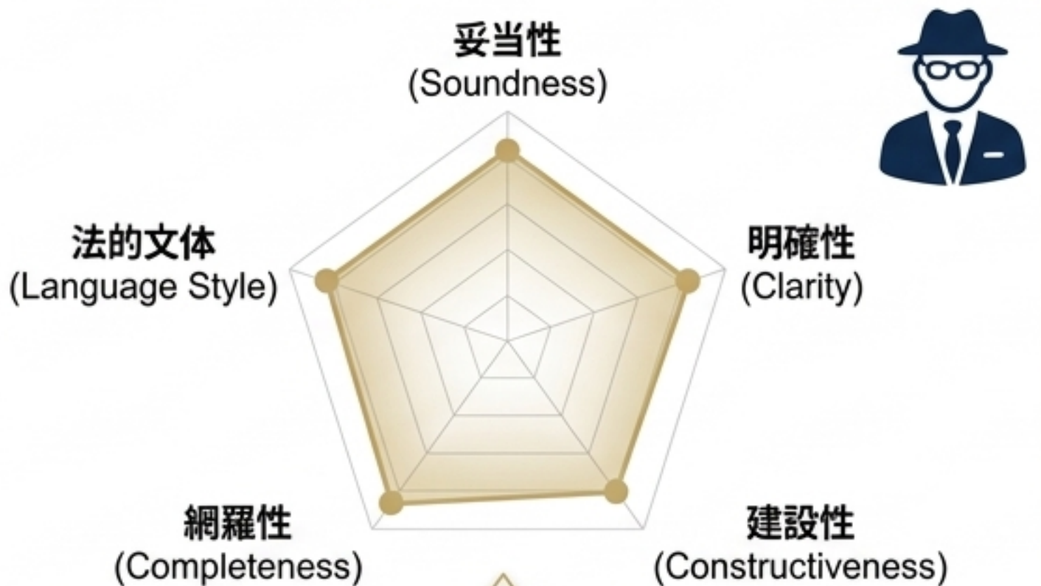
Level 1: 客観的検証 (Objective Deterministic)

Decision Accuracy:
最終的な審査判断（許可/拒絶）の精度

Statute Precision:
35 U.S.C.（米国特許法）の引用正確性

Rouge-L:
テキストベースの表面的な一致度

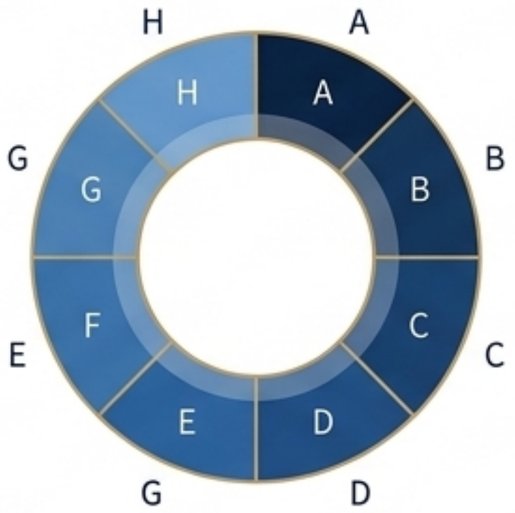
Level 2: LLM-as-a-Judge による意味論的監査



Rebuttalタスク専用指標:
Point-wise Coverage（反論網羅率）

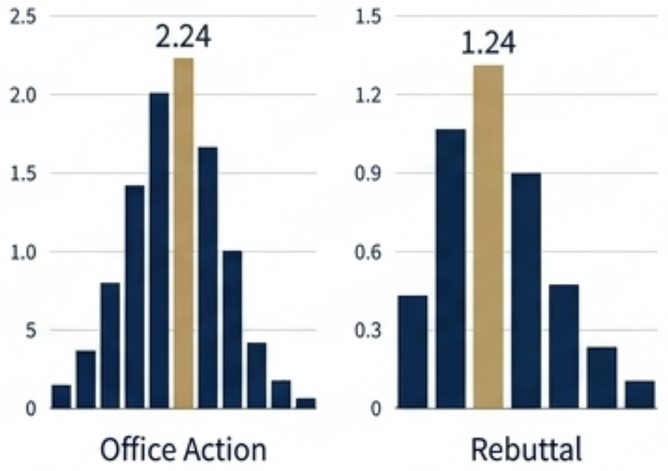
PatRe: 480件のリアルワールド・データセットの解剖学

IPC Diversity



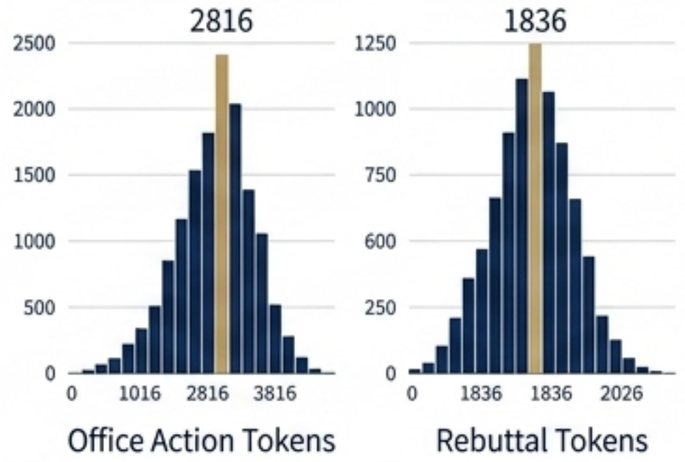
多様な技術分野: 全8つのIPCセクション（A~H）を網羅。特定の技術分野に偏らない現実的な難易度。

Multi-turn Reality



マルチターンの現実: 1件あたり平均2.24回のOAと1.24回の反論。特許審査が単発の分類ではないことを裏付ける。

Context Length



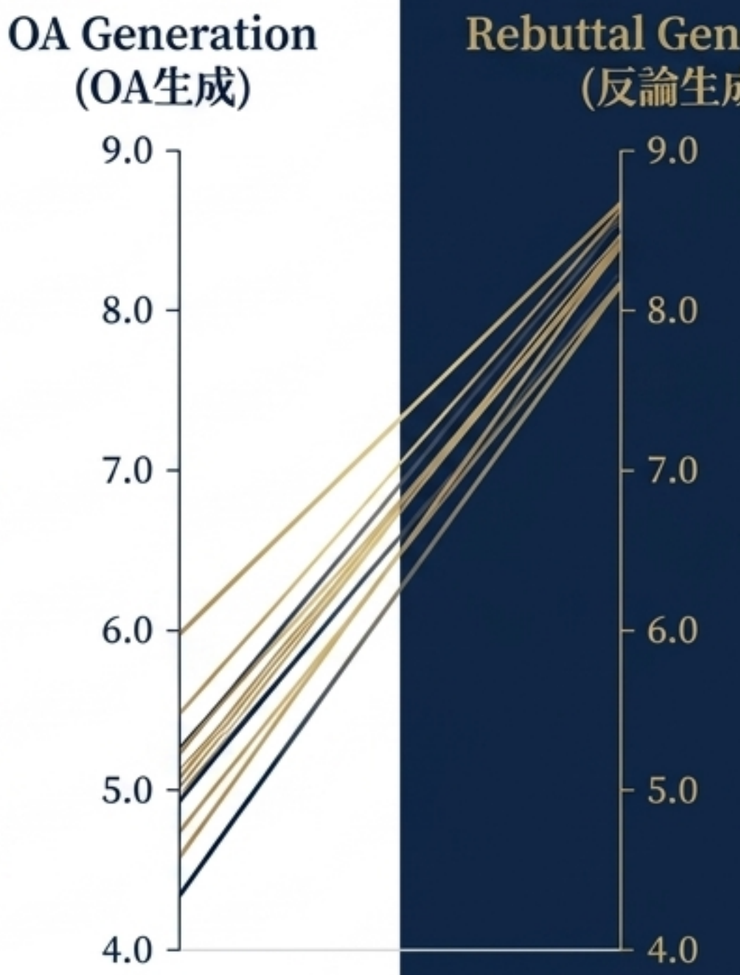
長大なコンテキスト: OAは平均2,816トークン、Rebuttalは1,836トークン。法的手続きにおける長文での高度な推論能力が要求される。

洞察①：LLMは「能動的な問題発見」よりも「受動的な防御」に長けている

Proactive (能動的審査)

OA生成時の平均スコア: ~4.5

新規クレームと先行技術から
自律的に法的矛盾を見つけ出す
「審査官としての問題発見能力」
には大きな壁がある。

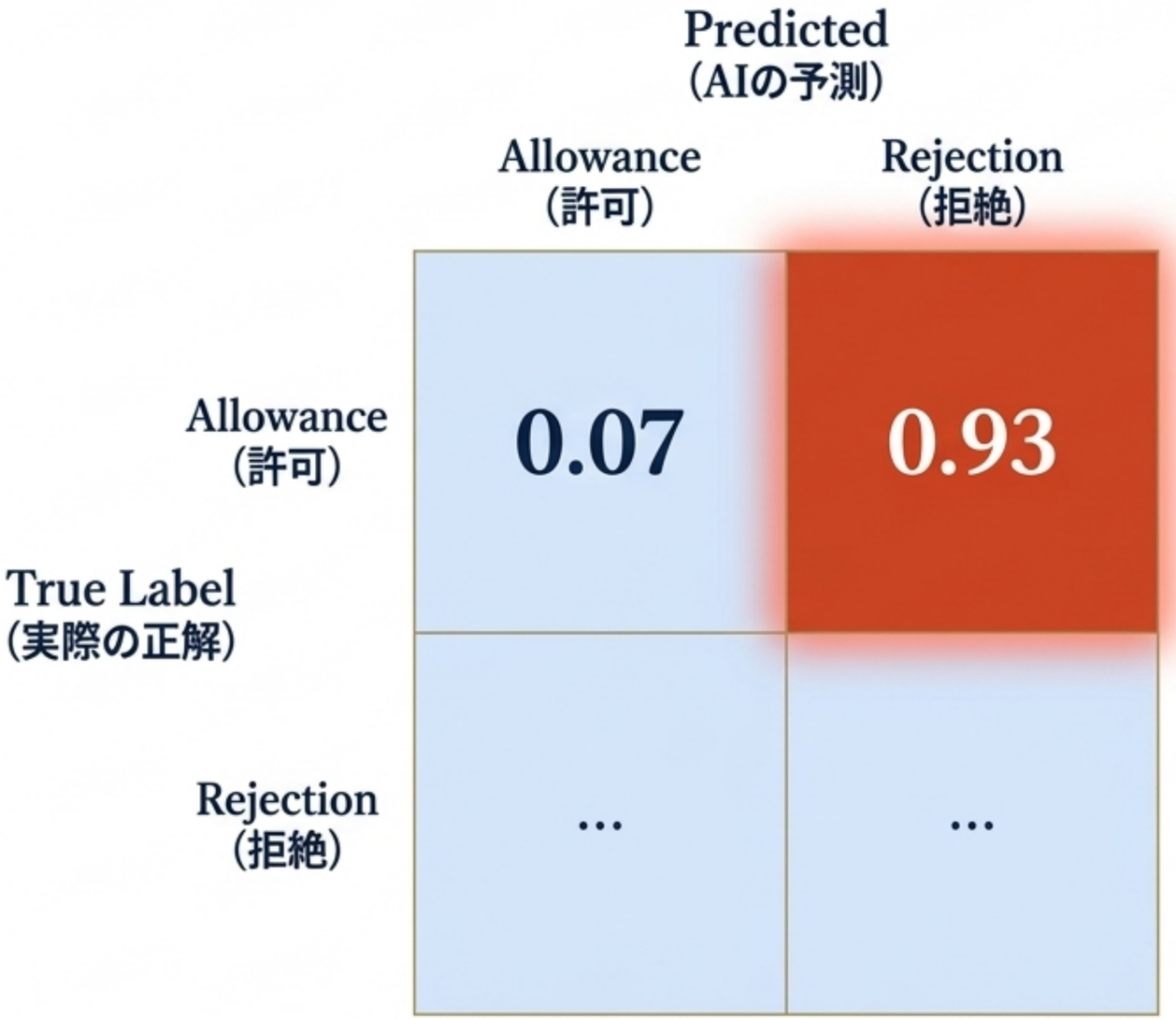


Reactive (受動的防御)

反論生成時の平均スコア: ~8.5

明示された指摘事項に対する
「応答・反駁」の論理構築には
優れている。

洞察 ②：「ハイパークリティカル・バイアス」～存在しない欠陥を捏造するLLM～



Data Insight

許可されるべき特許に対する過剰な拒絶。例えば LLaMA-3.3-70B は、本来許可される案件の精度がわずか 9.7% に留まり、残りの大部分を不当に拒絶している。

Conclusion

モデルは「厳格な審査＝拒絶すること」という強いバイアスを持っており、特許性のある発明を正当に評価できず、存在しない先行技術との矛盾（幻覚）を捏造して粗探しに走る傾向がある。

洞察 ③：条項によって異なるLLMの法的エラーモード

Dual Failure (不安定な幻覚): §101 / §102

存在しない違反を捏造する（FP: 72.8%）一方で、実際の欠陥を頻繁に見落とす（FN: 48.8%）。

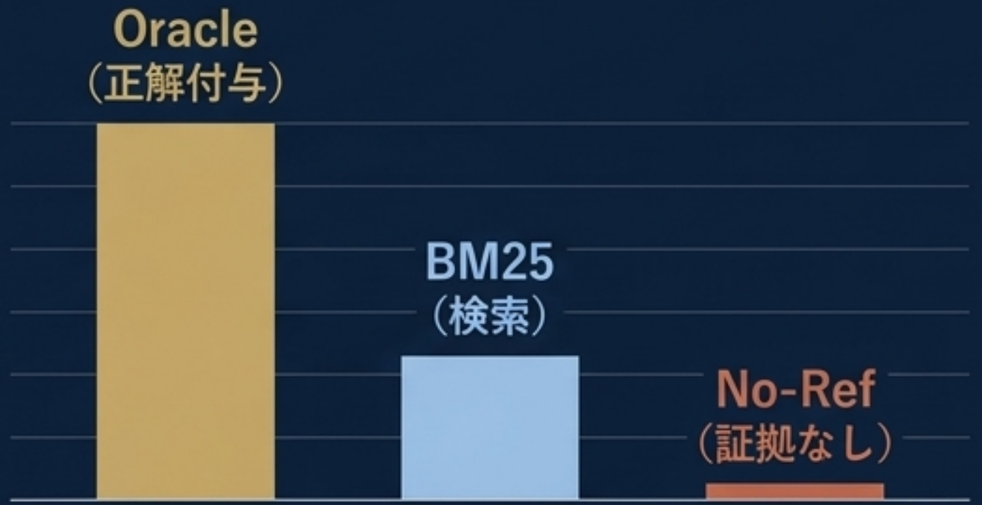
Over-Enforcement (過剰な法執行): §103 / §112

見落としは少ないが、ルールの境界線を拡大解釈しすぎる傾向が強い（FP: 60.0%）。

[考察] LLMは「法定要件」と「非法定要件」の境界線を正確に引く能力が著しく不足している。

洞察 ④&⑤：外部証拠への依存と、テキスト一致度指標の欺瞞

証拠の必須性
（外部証拠への依存）



引用の正確性は、モデルの推論力ではなく「外部証拠の質」に完全に依存する。

証拠がない場合、モデルは平気で引用を幻覚（Hallucinate）する。

指標の乖離
（Lexical vs Reality）

表面的なテキスト一致度は、人間の専門家の評価とほとんど相関しない（$\tau = 0.0258$）。

もっともらしい「法律用語風の文体」を生成できることは、内容の法的正当性を一切保証しない。

ベンチマーク結果：Proprietary vs. Open Source

🥇 Top Proprietary

GPT-5-mini

- 両タスクで首位を獲得。
- 反論時 (Rebuttal) の Point-wise Coverage (90.5%) と論理的妥当性 (Soundness: 8.71) で他を圧倒。

🥈 Top Open Source

Qwen3.5-27B

- オープンソース勢の中で突出した成績。
- 決定精度（Decision Accuracy）等の表層的な指標では商用モデルに肉薄する健闘を見せる。

【The Reasoning Gap】

表面的なフォーマット（Language Style）の模倣はOSSでも成熟しているが、高度な敵対的推論においては、商用トップモデルとの間に明確な「論理構築力」のギャップが存在する。

結論：AI特許審査の未来へ向けて

【The Facade of Professionalism（専門性の仮面）】
LLMは「特許庁風の文書（文体・フォーマット・明確さ）」を模倣することには既に長けている。

【The Reasoning Gap（推論の壁）】
しかし、厳密な証拠に基づく推論、ハイパークリティカル・バイアスの克服、
および法定要件の正確な適用には、依然として重大な課題が残されている。

PatRe は、AIが単なる「法的文書生成器」から脱却し、真に論理的で対話可能な
「特許審査パートナー」へと進化するためのマイルストーンである。