PatRe 論文の内容整理・批評的評釈

公開後の学術的・実務的な評価と反響

調査・評釈報告書 ｜ 調査基準日：2026 年 9 月 12 日
対象論文：arXiv:2605.03571v1（2026 年 5 月 5 日）
GPT-6 Astra

要旨

PatRe は、**特許審査を、単発の判定ではなく、審査通知・反論・補正の連続として評価する**という研究の方向性を提示している。この着想は高く評価できる。一方、「生成 AI は審査官としては弱いが、出願人側の反論作成では実務的に十分な能力を持つ」とまで結論づけるには、評価設計と検証が不足している。とくに、本文の高い自動採点に対して、付録の人手評価はかなり慎重な結果となっている点が重要である。［1］（pp.4–5、15、表 12）

公開後の反響として、英語・中国語・日本語での紹介が確認できる。ただし、その中心は論文ダイジェスト、ニュース解説、著者自身の告知である。第三者による再現実験、特許実務家による本格的な実証評価、実務導入の成果まで確認できた状況ではない。著者サイトは調査時点でも PatRe を「Manuscripts」に分類しており、学会採択を示す記録は確認できていない。［2］［3］

> **本報告書の中心的な評価**
> PatRe は「AI が特許審査・中間処理を代替できることを示した論文」ではなく、「特許審査・応答における AI の能力を、どのように測るべきかを一歩進めた論文」と位置づけるのが妥当である。

本書の構成と資料の扱い

本書は、①論文内容の整理、②批評的評釈、③公開後の学術的・実務的な評価・反響、④特許実務への含意、⑤総合評価の順に構成する。論文の主張、外部資料で確認した事実、本報告書の分析・提案を区別して記述する。

文中の［番号］は文末の参考文献一覧に対応する。原論文については必要に応じて頁・図表・節を併記した。ウェブ上の数値は参照時点の表示であり、被引用数とは区別する。「未確認」は「存在しない」という断定ではない。書誌整備時に確認できた日付の相違、再照合できなかった情報は、文末の「編集上の補記」に明記した。

1. 論文内容の整理

1.1 何を提案した論文か

正式題名は、**“PatRe: A Full-Stage Office Action and Rebuttal Generation Benchmark for Patent Examination”**である。Qiyao Wang ら、中国科学院深圳先進技術研究院、大連理工大学、UNSW Sydney 等の研究者による論文で、arXiv への初回投稿は 2026 年 5 月 5 日である。調査時点で、arXiv の投稿履歴に確認できるのは v1 である。［1］（p.1、arXiv 投稿履歴）

従来研究が主として「この出願は許可されるか」「どの拒絶理由に該当するか」という分類や情報抽出を扱っていたのに対し、PatRe は、次の一連の過程を評価対象に据えている。［1］（pp.1–4、図 1・表 1）

> 現在のクレーム → 審査官の Office Action → 出願人の反論 → クレームの変更 → 次の審査段階

論文 2 頁の図 1 は、このデータ構造、生成タスク、評価方法を一体として示す。ここでいう Office Action（OA）は、拒絶理由通知だけでなく、許可通知等も含む広い意味で用いられている。［1］（p.2、図 1；pp.13–14、表 7）

この論文の中心的な貢献は、新しい特許専用モデルを開発したことではなく、**特許審査・応答の生成能力を測定するためのベンチマークを構築したこと**にある。［1］（pp.2–4）

1.2 データセットとタスク

データセットは、USPTO の実際の審査記録に基づく 480 件の特許案件で構成される。各案件について、OA、出願人の応答、クレームの変更、引用文献等を対応づけている。IPC の A〜H の 8 セクションを含み、図 2 では各セクションがほぼ均等な比率となっている。［1］（pp.5–6、§3.3–3.4・図 2）

表 1　PatRe の評価タスク・条件

タスク・条件	モデルに与える情報	測定しようとする能力
OA-DP：Direct Prompting	クレーム等。外部の先行技術文献は与えない。	内部知識だけによる審査通知の生成。
OA-RO：Reference-Oracle	実際の審査で引用・検討された参照文献。	有力な証拠が与えられた場合の審査・理由づけ。
OA-RS：Retrieval-Simulated	BM25 検索文献と、正解側の参照文献を混ぜた候補群。	不適切な候補を除き、必要な文献を選んで審査する能力。
Rebuttal Generation	クレーム、OA、引用文献、審査履歴等。	拒絶理由に対応した技術的・法的反論の生成。

出典：原論文§3.1、図 8・図 9 を基に整理。OA-RO の参照文献は、審査官の引用文献だけでなく、出願人が引用して審査で検討された文献も含む。［1］（p.4、pp.21–22）

1.3 評価方法

評価は二層構造である。第一層は、実際の審査記録との一致を測る指標であり、OA の判断結果の一致、適用条文の適合率、ROUGE-L による文章の重なりを測定する。第二層では、Gemini-3.1-Flash-Lite を採点者として、妥当性、明確性、建設性、完全性、文体を採点する。反論生成には、拒絶理由の各論点にどれだけ対応したかを表す Point-wise Coverage も用いる。［1］（p.5、§3.2）

比較対象は GPT-5-mini、Gemini-2.5-Flash、DeepSeek-V3.2、GPT-4o-mini、Llama、Qwen、Gemma 系の計 10 モデルである。したがって、この結果は**当該比較対象と実験条件についての評価**であり、2026 年 9 月時点のすべての最新モデルの能力を直接示すものではない。［1］（pp.6–8、表 2–4）

1.4　主な実験結果

代表的な 4 モデルの結果を抜き出すと、次のとおりである。

表 2　OA-RO と反論生成の主要結果

モデル	OA-RO の判断一致率	OA-RO の自動採点平均／10	反論生成の自動採点平均／10	反論の Point-wise Coverage
GPT-5-mini	50.0%	4.89	**9.18**	**90.5%**
Gemini-2.5-Flash	46.4%	4.36	8.34	81.5%
DeepSeek-V3.2	49.7%	4.34	8.37	79.1%
Qwen3.5-27B	47.6%	4.37	8.29	79.2%

出典：原論文 7 頁の表 2、8 頁の表 4。自動採点平均は LLM-as-a-Judge による。［1］（pp.7–8、表 2・表 4）

第一に、OA 生成より反論生成の方が高得点である。 著者らは、問題を自ら見つけ出す審査官側の作業より、既に示された問題に対応する出願人側の作業の方が、LLM に適していると解釈している。［1］（pp.8–9、Finding 1）

第二に、専門的な文体と、技術・法律上の妥当性に大きな差がある。 例えば GPT-5-mini の OA-RO では、明確性 7.77、文体 7.84 に対し、妥当性は 3.07、建設性は 2.65 である。「審査通知らしく書けること」と「審査内容が正しいこと」は一致していない。［1］（p.8、表 3）

第三に、参照文献を与えても、判断の正確さが一貫して改善するわけではない。 文献の提供によって条文や引用の整合性は改善しても、クレームと先行技術の実質的な対比が十分になるとは限らない、という結果である。［1］（p.7、Observation 3）

2. 批評的評釈

2.1　評価すべき点――「結論」から「理由と応答」へ評価対象を広げた

本論文の最も重要な意義は、特許実務で必要な能力を、単なる許可・拒絶の予測に還元しなかったことにある。

クレームのどの要素が問題なのか、引用文献の何が根拠になるのか、出願人の反論に対して何を維持し何を撤回するのか――こうした生成物まで評価しようとする設計は、分類精度だけを測るより、実務上の問題を可視化しやすいと評価できる。論文自身も、先行研究との違いを、クレームの変遷、審査官・出願人の相互作用、生成タスクの組合せに置いている。［1］（p.3、§2・表 1）

また、人手評価を実施し、ROUGE-L のような文章一致指標だけでは不十分だと検討した点も評価できる。ただし、その人手評価は、本文の強い解釈をむしろ限定する材料にもなっている。［1］（p.10、表 5；p.15、付録 C）

2.2 最重要の留保——人手評価では「反論生成 9 点台」ではない

原論文 15 頁の表 12 を確認すると、結果の見え方が変わる。

表 3 反論生成における自動採点と人手評価

モデル	自動採点平均	人手評価平均	妥当性：自動採点	妥当性：人手評価
GPT-5-mini	9.18	**6.85**	8.71	**5.94**
Gemini-2.5-Flash	8.34	**6.67**	7.77	**5.75**
DeepSeek-V3.2	8.37	**5.05**	7.66	**4.00**

出典：自動採点は原論文表 4、人手評価は表 12。いずれも 10 点満点。ただし両表の評価対象集合は同一ではなく、数値差を同一サンプルに対する採点差として扱うことはできない。［1］（p.8、表 4；p.15、表 12）

人手評価は、7 モデルにまたがる OA100 件・反論 100 件の抽出評価であり、自動採点表と対象集合が同じではない。したがって、表 3 の差を、そのまま「AI 採点の過大評価幅」として扱うこともできない。評価者は知財を専門とする博士課程学生 3 名で、モデル名を伏せて評価している。米国特許実務の資格・経験等は、この記述だけでは確認できない。［1］（p.15、Human Evaluation Setup）

それでも、次の読み方は妥当である。

> 反論生成が OA 生成より容易である可能性は、人手評価でも支持される。しかし、反論の技術的・法的妥当性が「ほぼ完成している」との解釈は、人手評価によっては支持されない。

なお、人手評価の OA 平均では DeepSeek-V3.2 が 4.33、GPT-5-mini が 3.59 で、自動採点の順位とも一致していない。したがって、「GPT-5-mini があらゆる観点で最良」とする読み方も避けるべきである。［1］（p.15、表 12）

2.3 OA と反論は、同じ物差しで採点されていない

OA と反論の得点差を、純粋な能力差と解釈するうえで、評価プロンプトの非対称性が重要である。

OA の採点では、実際の OA を絶対的な参照とし、Non-Final／Final／Allowance のラベルが一致しなければ、妥当性の点数を 3 以下に抑えるよう指示している。これに対し、反論の採点プロンプトでは、soundness が主として「OA の各論点に具体的に対応しているか」という尺度になっている。また、付録の反論評価項目と、本文表 4 の 5 項目との対応も、完全には明示されていない。［1］（pp.23–24、図 10・図 11）

したがって、OA の 4.89 と反論の 9.18 には、少なくとも、**タスクの難しさ、入力される手掛かりの量、採点基準の厳しさ**という複数の要因が混在している。

本報告書の評価では、著者の「反応的な防御の方が得意」という仮説は興味深いものの、これを確かめるには、共通の証拠検証基準を用いた採点や、採点条件をそろえた追加実験が必要である。

また、**Point-wise Coverage の 90.5%は、反論の正答率でも、拒絶理由の解消率でも、権利化成功率でもない。** 付録では論点対応の程度を 1〜10 の尺度で評価する形式が示されており、この値を「90.5%の案件で有効な反論を作れた」と読み替えることはできない。［1］（p.8、表 4；p.24、図 11）

2.4 「全段階」のデータはあるが、自律的な権利化プロセスを実証したわけではない

ここは、実務家が最も誤解しやすい部分である。

PatRe には複数段階の審査記録が含まれる。しかし、公開コードを確認すると、過去のコンテキストはデータセット内の OA や応答文から構築され、各段階のクレームもデータセットから取り出す構成

である。少なくとも調査時点で公開されている処理は、**実際の履歴を手掛かりに、その段階の文書を生成する評価**が中心である。［4］（_extract_round_texts、_build_previous_round_context 等）

これは、AI が自分で作った補正・反論を次の AI 審査官に渡し、その結果を受けてさらに補正し、最終的に妥当な権利範囲へ到達する「閉じた自律ループ」の検証とは異なる。

したがって、PatRe が直接示したのは、**審査過程の各場面における条件付き文書生成能力**と捉えるのが適切である。自律ループ全体の誤りの累積、補正による権利範囲の毀損、適切な収束、審査回数の削減等は、別途評価すべき課題である。

2.5 「検索付き」も「十分な審査資料付き」も、実務条件とは違う

OA-RS は、通常の先行技術調査の再現率を測る設定ではない。 付録のプロンプトでは、候補群は基本的に「BM25 の 5 件＋Oracle の 5 件」とされ、正解側の文献があらかじめ混ぜられている。これは、重要文献をゼロから探し出せるかというより、正解文献を含む候補群から適切なものを選べるかを測る設定である。［1］（p.21、図 8）

また、生成用プロンプトの入力はクレーム、OA、参照文献のクレーム・要約、履歴等が中心で、出願当初の明細書全文・図面が独立した入力項目として明示されていない。それにもかかわらず、実施可能要件、明細書による裏付け、新規事項の有無まで判断させている。［1］（p.22、図 9）

米国の記載要件や補正の適法性は、出願当初の開示に基づいて検討する必要がある。このため、本設定では、**モデルの法的推論の誤りと、必要な資料を与えていないことの影響を分離しにくい**という問題がある。［5］（§2163.01、§2163.03、§2163.05–.06）

さらに、システムプロンプトには、不足情報があっても仮定を置いて進め、追加資料を求めず、限界を述べないよう指示している。この設計では、「証拠が足りないので判断を保留する」という実務上重要な対応が評価されない。［1］（pp.21–22、図 8・図 9）

したがって、観察された幻覚や過剰拒絶をすべてモデル固有の性質とみなすのではなく、**強制的に結論を出させるプロンプトの影響**も考える必要がある。

2.6 サンプルと指標には、一般化を制限する点がある

付録の OA 内訳は、合計 1,075 件のうち、Notice of Allowance が 479 件、Non-Final Rejection が 435 件、Final Rejection が 152 件、Ex Parte Quayle Action が 9 件である。480 案件に対して許可通知が 479 件あることから、少なくとも、最終的な許可に至る履歴を強く含んだ標本である。放棄・不成立に終わる案件を含む一般的な出願母集団と同じとはいえない。［1］（pp.13–14、付録 A・表 7）

IPC をほぼ均等に配置した設計は、分野横断の比較には有用であるが、そのまま企業の実案件構成における平均性能を示すものではない。日本語・日本の審査基準への適用も、論文自身が今後の拡張課題としている。［1］（p.6、図 2 ; p.11、§5）

表 4　評価指標が測るものと、その限界

指標	実際に測るもの	それだけでは分からないもの
判断一致率	実際の OA の分類との一致。	別の、しかし法的に妥当な審査判断の可能性。
Statute Precision	挙げた条文が正解側の条文集合と一致する割合。	条文の適用理由、要件ごとの充足、見落としの全体像。
Reference Citation Accuracy	引用した文献 ID と参照集合との一致。	その文献が、主張された技術内容を本当に開示するか。
ROUGE-L	実際の文書との語句・系列の重なり。	法的妥当性や、権利範囲を守る戦略の良さ。

出典：原論文§3.2、Finding 4 の定義を基に本報告書で整理。［1］（p.5、pp.9–10）

これらの限界は、各指標の定義から生じる。とくに、**文献番号が合っていることと、その文献に基づくクレーム対比が正しいことは別問題**である。

また、「判断一致率が 50%程度だから、二択の当てずっぽうと同じ」とする批判も適切ではない。公開コードは Non-Final と Final を別分類にしており、評価対象・除外条件を確かめずに二値分類の偶然水準と比較できないからである。［6］［7］

2.7　公開コードの点検で、採点に影響し得る問題も見つかった

公開コードの静的点検では、metrics_library.py の本文から判断を抽出する処理に、**「final rejection」を「non-final rejection」より先に判定する分岐**が確認された。「non-final rejection」には「final rejection」という文字列が含まれるため、この分岐では Non-Final を Final として扱う可能性がある。［7］（extract_disposition）

ただし、調査時点のコードには、本文抽出より優先して機械可読の OA 分類を読む経路もある。したがって、この問題から直ちに「論文の結果は誤計算だった」と断定はできない。**論文実験時のコード版、モデルが返した分類形式、本文抽出への切替頻度を確認する必要がある。**［7］（disposition_from_oa_type）［6］（OA 分類の読取り・採点処理）

これは、論文が Llama の誤りとして「Non-Final を早まって Final とする」現象を論じているだけに、優先的に照合すべき点である。［1］（p.9、Finding 2）

公開コードの静的点検と、論文実験の再現は異なる。本調査では、元データと実験時の全出力を使った再計算までは行っていない。コードの参照先は更新され得る main ブランチであり、論文実験時の版との同一性は未確認である。

2.8　「商用対オープンソース」という一般論にも慎重さが必要

論文では DeepSeek-V3.2 を、計算資源上の理由で API から利用したため「proprietary」側に分類したと説明している。したがって、この比較には、モデルの公開形態だけでなく、モデル規模、提供方法、推論環境等の違いも含まれる。［1］（p.14、付録 B・表 9）

実際、Qwen3.5-27B の反論生成平均 8.29 は、Gemini-2.5-Flash の 8.34 や DeepSeek-V3.2 の 8.37 に近い値である。ここから導くべき結論は、「商用が常に優れる」ではなく、**モデル・タスク・評価項目ごとに適性を検証する必要がある**ということである。［1］（p.8、表 4）

2.9　反論生成の事例自体にも、検証を要する箇所がある

20 頁の表 16 では、引用文献の入力要約に「M153 座位の遺伝的改変」と「サイトカイン等をコードする導入遺伝子」が記載されている。一方、生成された反論は、その引用文献が M153 座位の特定の改変や導入遺伝子との組合せを教示しない、と述べている。［1］（p.20、表 16）

掲載事例は省略を含むため、これだけで実際のクレームと引用文献の相違を法的に確定することはできない。しかし、**提示された説明だけでは「引用例にない」と断定する根拠が十分に示されていない**ことは指摘できる。

この事例は、「反論らしい文章ができていること」と「引用例の開示を正確に踏まえた反論であること」を分けて評価する必要性を、反論生成側についても示している。

3. 公開後の学術的・実務的な評価・反響

3.1 学会採択・査読状況

調査時点では、次の状況である。

表 5 学会採択・査読状況の確認結果

確認対象	確認結果	解釈
arXiv	2026 年 5 月 5 日の v1 を確認。	プレプリントとして公開。
著者ホームページ	PatRe を「Manuscripts」に掲載。	採択済み論文としての表示は確認できず。
ACL Anthology・OpenReview 等	完全題名・番号による公開検索で、PatRe の採択を示す記録を特定できず。	「未採択」との断定ではなく、採択未確認。
著者の他の研究	採択実績の記載あり。	PatRe 自体の採択と混同しないことが必要。

出典：arXiv 書誌情報、著者ホームページ、および公開検索の確認記録。検索対象の全記録を網羅したものではない。［1］［2］

arXiv と著者サイトから確認できる範囲では、**「査読付き国際会議で評価が確立したベンチマーク」と紹介する根拠は、現段階では得られていない。**［1］［2］

3.2 引用状況――総被引用数は確定できなかった

Google Scholar の著者プロフィールには、本論文の掲載を示す検索結果の抜粋がある。ただし、最新の被引用総数と引用論文一覧を、元ページで一貫して照合するところまでは到達できなかった。［8］

Semantic Scholar についても、arXiv 書誌ページからの連携照会先を確認対象としたが、書誌整備時には照会結果を取得できず、最新の被引用数・引用一覧を再確認できなかった。したがって、本書では総被引用数を数値で確定しない。［9］

前回答では、ResearchGate の検索結果に「Citations (0)」という表示があったことにも触れた。しかし、対象論文の個別ページを再特定・照合できなかったため、本書では、この表示を PatRe の被引用数を示す確認済み情報から除外する。

本調査では、PatRe を引用して独自の実験を行った後続論文を、本文まで確認して特定することはできなかった。また、Hugging Face 上の類似論文推薦や、PatRe が引用している先行研究を、「PatRe を引用した論文」と数えることもできない。Hugging Face のコメント欄には、類似論文を推薦する自動ボットの投稿が明示されている。［3］（Community 欄）

したがって、現時点の適切な表現は、**「被引用総数は未確定。独立した後続検証の広がりは、本調査では確認できていない」**である。

3.3 英語圏――研究紹介としての取り上げは確認できる

英語圏では、AI Native Foundation の“AI Native Daily Paper Digest – 20260506”に掲載され、480 案件、複数段階の相互作用、OA と反論の非対称性が紹介されている。ただし、内容は論文の要約であり、独自の再実験や実務検証ではない。［10］（第 4 項）

日付の補記：前回答では「5 月 6 日付」としたが、記事題名の日付は 20260506、ページ上の公開日表示は 2026 年 5 月 7 日である。本書では両者を区別した。［10］

また、AI Research Brief の 5 月 8 日付記事は、PatRe を注目研究の一つとして取り上げ、静的分類を超えて審査通知と反論の反復を扱った点を評価している。こちらも短い研究紹介である。［11］（Also Notable 欄）

確認できた英語圏の評価軸は、主に**「特許審査を多段階の相互作用として扱う着想の新しさ」**であり、採点方法や実務上の性能の検証が中心ではない。［10］［11］

3.4 中国語圏――肯定的な紹介と短評があるが、実証評価とは別

中国語圏では、次の 2 件が比較的明確な反応である。

表 6　中国語圏で確認できた主な紹介・短評

媒体・日付	内容・評価	証拠としての位置づけ
鯨鯢のブログ「AI Paper Daily」／2026 年 5 月 7 日	審査を多段階の対抗的過程として扱う発想と、審査・答弁の難易度差に実用上の意味があると肯定的に紹介。	論文ダイジェスト上の短評。Hugging Face を情報源とする。
科技行者／2026 年 5 月 12 日	「神探」の比喩を用いて、AI が審査の問題発見に苦戦すること、反論の方が得意なこと、過剰拒絶や条文適用の問題を解説。	科学技術ニュースとしての詳しい紹介。独自の再現実験ではない。

出典：各掲載ページ。［12］（第 15 項）［13］

ブログには生成者を示す“Generated by”表記もあり、記名の特許専門家による査読的評価と同一視すべきではない。［12］

科技行者の記事は、論文を全面的な成功物語として扱うのではなく、過剰拒絶や法的推論の不安定さも説明している。一方、人手評価と自動採点の差や、OA・反論の採点基準の非対称性を独立に検証した記事ではない。［13］

本調査の検索範囲では、中国の特許事務所・企業知財部門による具体的な運用評価や、再現実験の報告までは確認できなかった。

3.5 日本語圏――紹介は存在する。ただし独立した専門評価は限定的

AIDB は、日本語タイトル、ポイント、Abstract の訳を掲載している。ページ自身が、それらは AI による自動生成であると明示している。したがって、日本語でアクセスしやすくなっている証拠ではあるが、人間の専門家が検証した評価の証拠ではない。［14］

Synapse Flow は、5 月 7 日付で紹介し、弁理士・審査官を支援する AI ツールの開発や、出願手続の効率化への期待を述べている。肯定的な展望を含むニュースダイジェストであるが、実験結果を追試したものではない。［15］

alphaXiv の日本語解説は、方法・結果を比較的詳しく紹介している。ただし、重要な数値の取り違えが見つかった。同解説は、r＝0.7285、τ＝0.5861 を、人間と LLM 採点との一致として説明している。［16］（「結論と今後の影響」）

原論文では、この数値は**人間評価者間の一致**である。LLM 採点と人手評価の相関として表 5 に示されている Pearson の値は、r＝0.6808 である。［1］（p.10、表 5；p.15、表 11）

この取り違えは、公開後の解説を読む際にも、**「論文紹介が多いこと」と「評価が独立に裏付けられたこと」は別である**と示す具体例である。

なお、添付された日本語 PDF は、同論文の日本語化資料として扱い、日本語圏の第三者による独立評釈とは区別した。数値と評価設計の確認には、英語原論文を用いている。［17］（pp.2–6）［1］

3.6 SNS・開発者コミュニティの反応

調査時に取得できた公開ページの表示では、GitHub は **10 stars、1 fork、公開中の Issues 0 件・Pull requests 0 件**であった。Hugging Face の論文ページは **7 upvotes** である。［18］［3］

Hugging Face で確認できたコメントは、著者による 5 月 6 日の紹介・投票依頼と、5 月 7 日の類似論文推薦ボットが中心である。また、同ページでは、論文にリンクするモデル・データセット・Spaces はいずれも 0 件と表示されている。これは Hugging Face 上のリンク集計であり、学術的被引用数ではない。［3］

X については、arXiv 番号を含む本論文の紹介投稿が検索インデックスで見つかった。しかし、投稿本文・日時・反応の全体を安定して取得できず、拡散数や評価の傾向を確定する材料にはしていない。［19］（検索結果の抜粋のみ確認）

これらから読み取れるのは、**小規模ながら関心と紹介は存在するが、開発者コミュニティで広く検証・改良が進んでいるとまでは確認できない**という段階である。

3.7 データ公開の到達点にも注意が必要

論文はコードとデータセットの公開を述べている。実際、コードの公開は確認できた。一方、プロジェクトサイトの Data リンクは Hugging Face の IPIntelligence 組織ページにつながり、同ページの表示は「datasets 0」「None public yet」である。［1］（p.1、Abstract）［20］［21］

公開設定ファイルは、ローカルの test_set_new.json や参照文献プールを必要としている。そのため、**第三者がそのまま取得して論文の実験を再現できる完成データ一式の公開状況までは、本調査では確認できなかった。**［22］

「コード公開済み」と「データ・評価出力を含めて完全に再現可能」は、分けて表記する必要がある。

4. 特許実務への含意

本章は、論文から得られる示唆に基づく本報告書の評価・提案であり、PatRe が直接実証した成果とは区別する。

PatRe が最も有用なのは、「AI に意見書を全部書かせるべきか」という二択の議論ではなく、**AI に任せる作業と、人間が検証すべき作業を分解するための材料**として使うことである。

実務上は、拒絶理由の論点一覧化、クレームと引用箇所の対応表、反論候補の複数案、応答漏れの点検といった支援工程から検証するのが適切である。一方、引用例に特徴が「ない」との断定、補正の出願当初開示による裏付け、必要以上の限定の回避、事業上必要な権利範囲の維持は、文章の流暢さや論点対応率とは別の評価軸として扱う必要がある。

表 7　特許実務で混同を避けるべき評価軸

混同しやすいもの	実務上、区別すべきもの
拒絶理由の各項目に文章がある。	各項目に対する反論が技術的・法的に成立する。
引用文献番号が正しい。	引用箇所が主張を裏付ける。
補正によって拒絶を解消できる。	事業上有用な権利範囲を維持できる。
実際の意見書に似ている。	その案件で最も適切な応答戦略である。

本報告書による実務上の整理・提案。

PatRe を実務導入の根拠として用いるなら、生成文書の点数だけでなく、**重大誤りの件数、確認・修正に必要な時間、不要な減縮の有無、専門家による採用率**を測る追加評価が必要である。この論文だけでは、そのような実務成果までは示されていない。［1］（§3.2、§4、付録 C）

5. 総合評価

PatRe は、**「AI が特許審査・中間処理を代替できることを示した論文」ではなく、「特許審査・応答における AI の能力を、どのように測るべきかを一歩進めた論文」**と位置づけるのが妥当である。

その価値は、OA、反論、クレームの変遷を結び付け、専門的な文体と実質的な推論能力の隔たりを示した点にある。一方、反論生成の高得点には、採点基準の違い、与えられる手掛かりの多さ、人手評価との隔たりがあるため、実務能力をそのまま表す数値としては扱えない。［1］（pp.7–10、表 12・図 10・図 11）

公開後の評価については、**英中日での紹介・肯定的な期待は確認できるものの、学術的な追試と実務的な妥当性検証によって評価が固まった段階とは確認できない。**［10］［11］［12］［13］［14］［15］

本調査で残った主要な未確定事項は、最新の総被引用数、学会採択の有無、完成データの取得可能性、実験時の採点コードと現在の公開コードの対応である。

> **評釈の中心命題**
> PatRe の意義は、生成 AI が「もっともらしい特許文書」を書けることと、「証拠に裏付けられた適切な審査・権利化判断」を行えることの間に、なお検証すべき大きな隔たりがあると示した点にある。反論生成の高得点は、その隔たりが解消された証拠ではない。

編集上の補記

本書は、前回答の本文・表を基に引用番号と参考文献を整備したものである。原論文の実験結果・分析の骨格は維持し、出典照合に関する次の点を明記した。

日付の区別。 AI Native Foundation のダイジェストは、題名に 20260506 とある一方、公開日表示は 2026 年 5 月 7 日である。第 3.3 節および参考文献では両者を区別した。［10］

引用数に関する留保の明確化。 Google Scholar は検索結果の抜粋による掲載確認にとどまり、Semantic Scholar の照会結果は書誌整備時に取得できなかった。前回答で触れた ResearchGate の「Citations (0)」は対象論文の個別ページと再照合できなかったため、確認済み情報から除外した。被引用総数は未確定のままである。

SNS 情報の確認範囲。 X は個別投稿への検索結果を確認したが、投稿日時や反応全体は確定していない。このため、前回答の「公開直後」という時期の表現は採用していない。［19］

参考文献

文中の引用番号順に掲載。参照日（ウェブ資料の確認日・添付資料の確認日）は、特記のない限り 2026 年 9 月 12 日である。原論文の図表・節番号と本報告書の表番号は異なる。

原論文、公式資料、公開コードを事実確認の主たる根拠とし、ブログ・ニュース記事は「公開後の反響」の記録として用いた。照会結果を取得できなかった資料は、その制約を各項目に記載した。

［1］ **Wang, Qiyao; Chen, Xinyi; Chen, Longze; Wang, Hongbo; Alinejad-Rokny, Hamid; Lin, Yuan; Yang, Min.** PatRe: A Full-Stage Office Action and Rebuttal Generation Benchmark for Patent Examination. arXiv:2605.03571v1 [cs.CL; cs.AI]. 2026 年 5 月 5 日. 24 頁. DOI: 10.48550/arXiv.2605.03571.

https://arxiv.org/abs/2605.03571

https://doi.org/10.48550/arXiv.2605.03571

一次資料。本文の数値・図表・プロンプトの確認は、ユーザー提供の英語原論文 PDF『2605.03571v1(1).pdf』による。頁番号は同 PDF の印刷頁番号。arXiv ページは投稿履歴・書誌情報の確認に使用。 参照日：2026 年 9 月 12 日。

［2］ **Wang, Qiyao.** Qiyao Wang | NLPer - Homepage. 著者ホームページ. 公開・更新日表示なし.

https://wangqiyao.me/

著者による一次情報。「Manuscripts」「Publications」「News」欄を確認。PatRe の採択未確認は本論文の掲載区分に基づくもので、著者の他論文の採択状況とは区別する。 参照日：2026 年 9 月 12 日。

［3］ **Hugging Face; Qiyao Wang（投稿者・コメント投稿者）.** Paper page - PatRe: A Full-Stage Office Action and Rebuttal Generation Benchmark for Patent Examination. Hugging Face Papers. 論文公開日表示：2026 年 5 月 5 日. ページ投稿：2026 年 5 月 6 日.

https://huggingface.co/papers/2605.03571

コミュニティ反応の一次記録。7 upvotes、著者コメント、自動推薦ボット、モデル・データセット・Spaces のリンク表示を確認。表示数は学術的被引用数ではない。 参照日：2026 年 9 月 12 日。

［4］ **AIforIP / PatRe contributors.** PatRe / run_exp.py. GitHub 公開ソースコード. main ブランチ、参照時点の版.

https://raw.githubusercontent.com/AIforIP/PatRe/main/run_exp.py

一次資料。_extract_round_texts、_build_previous_round_context、および各ラウンドの入力構築を静的点検。論文実験時のコミットとの同一性は未確認。 参照日：2026 年 9 月 12 日。

［5］ **United States Patent and Trademark Office（USPTO）.** MPEP §2163: Guidelines for the Examination of Patent Applications Under the 35 U.S.C. 112(a) or Pre-AIA 35 U.S.C. 112, first paragraph, “Written Description” Requirement. Manual of Patent Examining Procedure. §2163 の改訂表示：[R-01.2024].

https://www.uspto.gov/web/offices/pac/mpep/s2163.html

公式資料。特に§2163.01、§2163.03、§2163.05、§2163.06 を参照。小節ごとに改訂表示が異なる。出願当初開示、補正の裏付け、新規事項との関係を確認。 参照日：2026 年 9 月 12 日。

［6］ **AIforIP / PatRe contributors.** PatRe / src/evaluator.py. GitHub 公開ソースコード. main ブランチ、参照時点の版.

https://raw.githubusercontent.com/AIforIP/PatRe/main/src/evaluator.py

一次資料。OA 分類の読取り、機械可読分類から本文抽出への切替え、採点処理を静的点検。論文実験結果の再計算は未実施。 参照日：2026 年 9 月 12 日。

［7］ **AIforIP / PatRe contributors.** PatRe / src/metrics_library.py. GitHub 公開ソースコード. main ブランチ、参照時点の版.

https://raw.githubusercontent.com/AIforIP/PatRe/main/src/metrics_library.py

一次資料。extract_disposition と disposition_from_oa_type 等を静的点検。本文抽出経路での final／non-final の判定順序と、別経路の存在を区別した。 参照日：2026 年 9 月 12 日。

［8］ **Google Scholar / Qiyao Wang.** Qiyao Wang — Google Scholar 著者プロフィール. Google Scholar. 公開・更新日表示未確認.

https://scholar.google.com/citations?hl=zh-CN&user=STze0QgAAAAJ

書誌照合先。検索結果の抜粋で arXiv:2605.03571 の掲載を確認。元ページの最新被引用数・引用論文一覧は安定取得・照合できず、被引用数の確定には使用していない。 参照日：2026 年 9 月 12 日。

［9］ **Semantic Scholar.** PatRe（arXiv:2605.03571）の arXiv 連携照会先. Semantic Scholar 書誌照会サービス. 照会結果未取得.

https://api.semanticscholar.org/arXiv:2605.03571

未確認の調査先として記録。arXiv 書誌ページのリンク先を照会したが、書誌整備時には結果を取得できなかった。内容確認済みの参考資料や被引用数の根拠としては扱わない。 参照日：2026 年 9 月 12 日。

［10］ **AI Native Foundation.** AI Native Daily Paper Digest – 20260506. AI Native Foundation, Papers. ページ上の公開日：2026 年 5 月 7 日. 第 4 項「PatRe」.

https://ainativefoundation.org/ai-native-daily-paper-digest-20260506/

論文紹介。題名の日付 20260506 と、ページ上の公開日 5 月 7 日を区別。研究結果の独立検証ではない。 参照日：2026 年 9 月 12 日。

［11］ **AI Research Brief.** 10.6k SFT Trajectories Match Full RL Pipeline; Mamba Beats LZMA. Buttondown 掲載ニュースレター. 2026 年 5 月 8 日.「Also Notable」欄の PatRe 紹介.

https://buttondown.com/ai-brief-en/archive/106k-sft-trajectories-match-full-rl-pipeline/

論文紹介・短評。特許審査を多段階の OA と反論として扱う点を評価。追試の報告ではない。 参照日：2026 年 9 月 12 日。

［12］ **鯨鯢 WhaleDolphin（ブログ運営者）.** AI Paper Daily | 2026-05-07. 鯨鯢的博客. 2026 年 5 月 7 日. 第 15 項「PatRe：专利审查全流程生成基准」.

https://yf-cheng.com/zh-cn/p/daily-paper-2026-05-07/

中国語の論文ダイジェスト・短評。ページ末尾に「Generated by 爱弥斯」の表示。独立した特許実務上の検証とは区別する。 参照日：2026 年 9 月 12 日。

［13］ **科技行者.** 专利审查的"神探"挑战：中科院深圳先进院等机构推出 PatRe，测试 AI 能否胜任最难法律推理任务. 科技行者. 2026 年 5 月 12 日 10:11.

https://www.techwalker.com/2026/0512/3186603.shtml

中国語の科学技術ニュース解説。タスク非対称性、過剰拒絶、条文適用の問題を紹介。記事自体による再現実験は確認できない。 参照日：2026 年 9 月 12 日。

［14］ **AIDB.** 特許審査における全段階の審査官通知・反論生成ベンチマーク「PatRe」. AIDB Daily Papers. ページの公開日欄：2026 年 5 月 5 日（原論文公開日と一致。紹介ページ自体の作成日は未確定）.

https://ai-data-base.com/paper/2605-03571

日本語の論文紹介。日本語タイトル・ポイント・Abstract の日本語訳は AI による自動生成と明記されている。 参照日：2026 年 9 月 12 日。

［15］ **Synapse Flow 編集部.** PatRe: 特許審査におけるオフィスアクションと反論生成の完全段階ベンチマーク. Synapse Flow. 2026 年 5 月 7 日 13:00.

https://synapseflowlab.com/article.php?id=1530005

日本語の論文紹介・編集部短評。弁理士や審査官の業務支援、手続の効率化への期待を掲載。追試の報告ではない。 参照日：2026 年 9 月 12 日。

［16］ **alphaXiv.** PatRe: 特許審査のための全工程の拒絶理由通知及び応答書生成ベンチマーク. alphaXiv 日本語ページ、AI Overview. 論文投稿日表示：2026 年 5 月 5 日. 解説の公開・更新日は未確認.

https://www.alphaxiv.org/ja/abs/2605.03571

日本語の AI 解説。「結論と今後の影響」に記載された相関係数の説明を、原論文の表 5・表 11 と照合。解説の記述を原論文の結論と同一視しない。 参照日：2026 年 9 月 12 日。

［17］ **作成者・翻訳者未確認.** PatRe：特許審査における、オフィスアクションと反論書作成の全段階ベンチマーク. ユーザー提供 PDF. 55 頁. 作成日不明.

同名の添付ファイル。原論文の日本語化資料として参照。第三者による独立評釈とは扱わず、数値・評価設計は英語原論文で照合した。公開 URL は未確認。 参照日：2026 年 9 月 12 日。

［18］ **AIforIP / PatRe contributors.** AIforIP/PatRe: PatRe: A Full-Stage Office Action and Rebuttal Generation Benchmark for Patent Examination. GitHub リポジトリ. main ブランチ、参照時点の表示.

https://github.com/AIforIP/PatRe

コード公開と開発者コミュニティ反応の一次記録。10 stars、1 fork、公開中の Issues 0 件・Pull requests 0 件を確認。表示値は更新され得る。 参照日：2026 年 9 月 12 日。

［19］ **@HEI.** PatRe: A Full-Stage Office Action and Rebuttal Generation …（検索結果に表示された論文紹介投稿）. X. 投稿 ID：2051918235176468606. 投稿日時未確認.

https://x.com/HEI/status/2051918235176468606/photo/1

検索結果の抜粋のみ確認。投稿本文・反応全体は安定取得できず、拡散数や投稿時期の確定、実務家による評価の根拠としては使用していない。 参照日：2026 年 9 月 12 日。

［20］ **PatRe 研究チーム.** PatRe — Project Website. 研究プロジェクト公式サイト. 公開・更新日表示なし.

https://patre.wangqiyao.me/

一次資料。論文・コード・Data への案内を確認。Data リンク先の公開状況は Hugging Face 組織ページで別途確認した。参照日：2026 年 9 月 12 日。

［21］ **IP Intelligence.** IPIntelligence（IP Intelligence）. Hugging Face 組織ページ. 参照時点の表示.

https://huggingface.co/IPIntelligence

データ公開状況の一次記録。datasets 0、None public yet の表示を確認。他の場所での公開まで否定するものではない。参照日：2026 年 9 月 12 日。

［22］ **AIforIP / PatRe contributors.** PatRe / config/eval_config.yaml. GitHub 公開設定ファイル. main ブランチ、参照時点の版.

https://raw.githubusercontent.com/AIforIP/PatRe/main/config/eval_config.yaml

一次資料。ローカルの test_set_new.json および参照文献プール等への依存を確認。これだけでは完成データ一式の取得可能性は確認できない。 参照日：2026 年 9 月 12 日。