

生成AI特許調査サービス「サマリア」の「特許検索競技大会・再現率100%（上位30件）」主張の検証・分析報告書

エグゼクティブサマリー

結論から言えば、PR TIMES記事に記載された「2024年度特許検索競技大会アドバンストコースの3分野で、既知の正解特許26件をすべて上位30件以内に入れた」という“社内ベンチマークの観測結果”は、公開情報の範囲では整合的であり、記事中の数値に明白な矛盾は確認できなかった。電気4件、機械8件、化学・医薬14件の計26件をすべて捕捉したという内訳も開示され、サマリア公式マニュアルにも同じ結果が掲載されている。したがって、「パテント・インテグレーションがこの試験を行い、26/26だったと公表した」こと自体の事実性は高い。¹

しかし、この結果を「サマリアは特許調査一般で再現率100%を実証した」あるいは「人間の専門調査者より高性能であることが科学的に証明された」と読むのは適切ではない。本検証はベンダー自身による自己評価であり、第三者による監査・再実行は公開されていない。また対象は3課題、既知正解文献26件に限られ、会社自身も「統計的な有意性を主張する規模ではない」と明記している。さらに現在の母集団総件数も取得されていない。²

とりわけ重要なのは、見出しの「上位30件」自体は特許検索競技大会の公式評価基準ではないと考えるべき点である。IPCC職員による大会解説では、競技の検索設問では検索式とヒットリストを提出し、そのヒットリストを「再現率を評価するため」の採点に用いると説明されている。一方、「上位30件で100%」はサマリア側がAIスクリーニング後のランキングを評価するために設定した Recall@30 型のベンダー独自指標である。³

さらに、今回使われた2024年度過去問は、同社が2026年2月にすでに同じ問題で一度評価して失敗箇所を把握した後、機能を改良し、同じ問題で再評価したものである。2月時点では機械8件中7件、化学・医薬14件中12件だったところ、9月には「類語シソーラス拡張」導入後に26/26となった。これは製品改善としては有意義だが、機械学習・情報検索のベンチマーク論では独立したホールダアウト・テストというより、既知のテストセットを見ながら改善した後の再テストに近い。したがって外部妥当性を評価するには、新規・未見の問題を用いたブラインド試験が必要である。⁴

「偶然に26/26になった可能性」についても、単純な確率は算出できない。ランダムランキングなら候補総数 N と各分野の正解数 m が分かれば上位30件への全件混入確率を計算できるが、会社自身が今回の母集団総件数は未取得としているためである。逆に、26件を独立な二項試行と強引にみなした場合、26/26の95% Clopper-Pearson信頼区間は概ね 86.8%–100% となるが、同一問題内の文献は独立ではなく、「26文献」よりも実質的な独立サンプル数は「3課題」に近い。この計算を性能保証として使うべきではない。元データが小さいという会社自身の留意事項の方が妥当な解釈である。⁵

総合評価は次のとおりである。

評価対象	本報告の評価	理由
「会社が26/26、上位30件で100%と発表した」	確認済み	PR TIMESと公式マニュアルで一致 ⁶
「2024年度大会の実問題を使用した」	高い蓋然性	過去問2024の存在・発行元を確認。同社は問2を使用と具体的に説明 ⁷
「大会公式基準で“上位30件100%”を達成した」	支持されない	上位30件はサマリア側の評価方法であり、大会公式の30件基準は確認できない ⁸
「独立第三者によって100%が検証済み」	未確認	公開された独立再現試験を確認できず
「未知の特許調査でも100%を期待できる」	根拠不足	3課題・26正解、同一過去問での再テスト ⁹
「実務上有望な検索・スクリーニング技術である」	合理的に示唆される	2月の失敗から9月に改善し、特に語彙拡張＋ランキングで大幅改善。ただし外部検証が必要 ¹⁰

PR TIMES記事の事実確認と信頼性

対象記事の発表元は**パテント・インテグレーション株式会社**、公開日時は**2026年9月10日10時00分**である。PR TIMES上でも「パテント・インテグレーション株式会社のプレスリリース」と明示されている。PR TIMES自体は「プレスリリース・ニュースリリース配信サービス」であり、記事は第三者記者による検証記事ではなく、発表企業が提供した情報を配信する形式である。 ¹¹

原典：

<https://prtimes.jp/main/html/rd/p/000000028.000086119.html>

記事の中心的主張は、短く引用すれば、

「上位30件を確認するだけで全分野が100%に到達」

である。記事はこれを、2024年度特許検索競技大会アドバンストコースの電気・機械・化学医薬3分野について、AIスクリーニング後の上位30件に既知正解文献がすべて含まれた、という意味で用いている。 ¹²

公開された結果の内訳は明瞭である。

分野	正解特許数	検索集合で捕捉	AIスクリーニング上位30件	Recall@30
電気	4	4/4	全4件を含む	100%
機械	8	8/8	全8件を含む	100%
化学・医薬	14	14/14	全14件を含む	100%
合計	26	26/26	26/26	100%

出典はPR TIMES本文および同社公式マニュアルである。 ⁶

記事は、結果だけを宣伝しているわけではなく、限界もかなり明確に開示している。具体的には、対象が3分野・正解特許計26件であり「統計的な有意性を主張する規模ではない」こと、母集団総件数を取得していないこと、拒絶理由通知書の蓄積が少ない分野へ同じ効果を一般化できないことを自ら記載している。この自己限定はプレスリリースとしては評価できる。⁵

一方、**証拠の独立性は低い**。試験者、製品開発者、評価結果の公表者が同一企業であり、第三者機関による監査結果、実行ログ、全順位表、生成検索式、モデルの具体的なバージョン、乱数条件、全出力ファイル等は今回確認した公開資料からは得られなかった。そのため、「公表された観測値の確認」と「観測値が第三者にも再現できることの確認」は分ける必要がある。¹³

記事編集上の小さな注意点もある。同じPR TIMESページ内で、本文では受賞歴を「**第37回**中小企業優秀新技術・新製品賞」と記している箇所がある一方、会社概要欄では「**第27回**」と記しており、ページ内部で回次に不整合がある。性能試験の数字とは直接関係しないが、プレスリリースが学術論文のような査読済み資料ではないことを示す編集品質上の一例ではある。⁵

したがって、PR TIMES記事の信頼性は二層に分けて評価するのが適切である。「**企業がこの結果を発表した**」という一次資料としては信頼性が高いが、「**その性能が独立に立証された**」という証拠としては弱い。

特許検索競技大会の主催者・過去問・評価基準

「特許検索競技大会」は2007年に始まり、**2013年から一般財団法人工業所有権協力センター（IPCC）が主催**している。IPCC所属研究員がJapio YEAR BOOKに寄稿した大会解説にも、この経緯が記載されている。アドバンストコースは特許調査実務者を対象とし、電気、機械、化学・医薬などから選択問題を解く形式が採用されてきた。¹⁴

2024年度の過去問については、Japioの販売案内で『**特許検索競技大会 過去問 2024**』が確認でき、監修は特許検索競技大会実行委員会、発行はIPCC、内容は「2024年度の問題と解答例・解説」とされている。したがって、ベンダーとは独立した大会側の正解資料を入手する手段自体は存在する。¹⁵

ただし、今回の調査では**2024年度版の全問題・採点表そのものを自由公開された公式PDFとして取得することはできなかった**。詳細な2024年度採点配点や全正解文献リストについては、販売されている過去問冊子が原典となる。このため、「2024年度問2の各小問について何点をどのように割り当てたか」という粒度の公式採点基準は本報告では**未確認**である。¹⁵

一方、IPCC職員による過年度の問題解説を見ると、検索式を作成・実行し、スクリーニング対象とした検索式と、その検索式でヒットした特許文献を「ヒットリスト」として提出する方式であり、本文中に明確に**ヒットリストは再現率を評価する採点に使う**と記されている。次の設問では、そのヒット文献の中から新規性を否定できる文献を選ぶ。したがって大会が検索の網羅性を重要視していることは原典に近い資料から裏付けられる。¹⁶

情報検索でいう**再現率（recall）**は通常、

$$Recall = \frac{\text{検索された関連文献数}}{\text{存在する関連文献数}}$$

であり、日本弁理士会系の特許情報解説でも「情報全体から調査目的に適合する情報を拾い出すことができた割合」という意味で説明されている。¹⁷

今回のサマリヤ試験をこの形式に合わせれば、正解集合を G 、AIスクリーニングの上位 k 件を R_k とした、

$$Recall@k = \frac{|G \cap R_k|}{|G|}$$

と理解するのが最も正確である。サマリアの見出しは $k = 30$ であり、4/4、8/8、14/14だったため全分野で `Recall@30 = 1.00` となる。¹⁸

ここで最重要なのは、「30」という数字を大会公式基準と取り違えないことである。大会側の過去問解説ではヒットリストの再現率を採点するが、本報告で確認した大会資料に「上位30件に入れば合格」「30件を公式カットオフにする」という規定はない。上位30件というカットオフは、サマリアがAIランキングの効率を表すために設定した評価点である。⁸

また、「上位30%」という別の数字にも注意が必要である。過年度のIPCC解説ではゴールド・シルバー・ブロンズ等の認定者がおおむね参加者の上位30%程度だったと説明されるが、これは参加者の順位分布の話であり、「上位30文献」とは無関係である。¹⁹

サマリアのサービス・技術・料金・データ

「サマリア (Summaria)」はパテント・インテグレーション株式会社が提供する特許文書読解・調査支援サービスで、2024年5月27日から提供されている。公式資料では、検索式・検索集合の生成、特許集合のAIスクリーニング、クレームチャート、無効資料調査、侵害予防調査、分類、レポート、拒絶対応、明細書作成支援などを一つの環境に統合している。²⁰

調査機能では、IPC、FI、Fターム等を含む検索式を生成し、検索母集団を作った後、「記載評価」等で文献本文・図面まで参照して関連度順にランキングする仕組みが公開されている。2026年9月には、拒絶理由通知書10万件以上を解析し、審査官が本願と引用文献の概念を対応すると認定した事例を検索語拡張に利用する「類語シソーラス拡張」を導入した。類似する審査事例を最大15件参照し、その情報を検索要素作成、文献評価、スクリーニングまで引き継ぐ仕様である。²¹

同時期のリリース情報では、最大50個の検索集合を組み合わせる際にRRF (Reciprocal Rank Fusion) スコアを使う検索集合作成ロジックも導入されたと説明されている。したがって、「単にLLMへ質問して検索結果を出すサービス」ではなく、特許分類・キーワード検索、複数集合統合、審査官由来シソーラス、生成AIによる意味評価を組み合わせたハイブリッド型と見るのが妥当である。²²

ただし、今回の100%試験で使われた具体的LLMモデル、プロンプト全文、ランキング重み、RRFのパラメータ、検索式生成ロジックの完全仕様は未確認である。公式マニュアルにはOpenAI、Claude、Gemini系の多数の利用可能モデルが掲載されるものの、2026年9月の当該ベンチマークでどれをどの工程に用いたのかはPR TIMES本文では特定されていない。これは第三者再現性上の重要な欠落情報である。²³

関連する技術について同社は複数の特許番号を公式サイト上に掲げているため、知財として開示された部分は存在する。しかし、公開特許が製品の全アルゴリズム、検索順位決定式、プロンプト、モデル設定を再現可能なレベルで開示しているとまでは確認できない。したがって「アルゴリズム公開」は部分的と評価すべきである。²⁴

料金は2026年9月時点の公式ページで税抜き次のように公開されている。²⁵

プラン	対象	年間契約時	月間契約時	30日あたり指示数
フリー	個人	0円	0円	48時間ごとに4回
コーポレート	企業・法人	25,000円/月	30,000円/月	2,000回

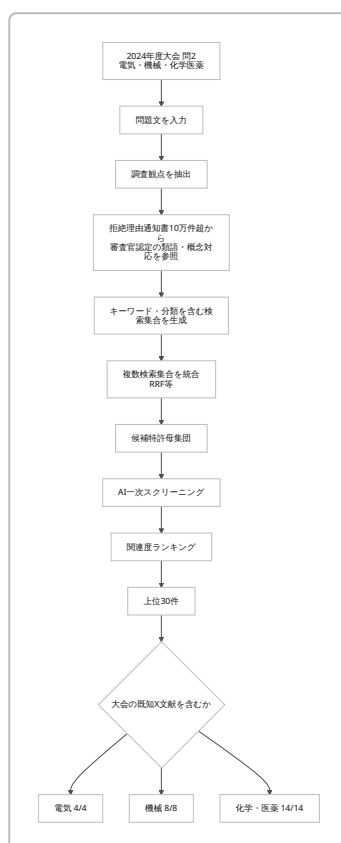
プラン	対象	年間契約時	月間契約時	30日あたり指示数
ビジネス	個人	15,000円/月	18,000円/月	1,000回
ミニマム	個人	8,000円/月	10,000円/月	400回

コーポレート／ビジネス等には有料機能が含まれ、追加指示の購入体系もある。月間契約については原則として最短3か月という条件が公式マニュアルに示されている。 ²⁶

外部連携としては、通常の公開REST APIの完全仕様は今回確認できなかった一方、「**サマリアMCP**」が公式に公開されている。Claude、ChatGPT、Copilot等からサマリアの特許データベースを検索し、本文・請求項・書誌・図面の取得や分類情報の参照等を行えるMCPサーバーで、コーポレートまたはビジネス契約では追加料金なしとされる。したがって「APIがない」というより、**公開確認できる外部機械連携インターフェースはMCPであり、汎用REST APIの一般向けドキュメントは未確認**、という表現が正確である。 ²⁷

特許データベースの公開収録範囲はかなり広く、日本、米国、欧州、PCT国際公開、ドイツ、中国、韓国、台湾、オーストラリア等が列挙されている。日本公開・登録公報は1971年以降を日次更新、米国や欧州・国際公開は週次更新等とされる。中国・韓国の日本語翻訳データの一部はJ-PlatPat経由で収録され、世界書誌データとしてDOCDBの利用も明記されている。 ²⁸

公開資料から再構成できる今回の処理イメージは、おおむね次のとおりである。 ²⁹



再現率100%の再現テストと統計的妥当性

公開資料から確認できるテスト設計は、2024年度特許検索競技大会アドバンストコース問2の電気、機械、化学・医薬を入力し、公式正解として扱われるX文献を検索・ランキングできるかを評価する、というものである。正解文献数は合計26件で、内訳は4、8、14件である。³⁰

2026年2月の最初の試験では、問2の設問(4)について、問題文・図表を入力し、検索集合上位200件に含まれる正解数を測定した後、検索集合上位2,000件をAIスクリーニングし、関連度上位200件を評価していた。人手による出力調整は行わなかったと説明されている。³¹

9月版では新たに「類語シソーラス拡張」を検索式作成に適用し、生成された検索集合を既存のAIスクリーニングへ送った。同社は「各分野の問題文をそのまま入力する以外、人手による調整は行っていない」と説明する。検索集合段階ですでに26/26となり、その後AIランキングの上位30件以内にも全26正解が入ったという。¹²

一方、以下の再現性に必要な情報は公開資料からは**未確認**である。

再現性項目	公開状況	重要性
使用した2024問2の正式問題・正解一式	過去問書籍として入手可能。本文の完全公開は未確認	入力とgold setの固定に必要 ¹⁵
実際に投入した入力文字列・画像	「問題文をそのまま」と説明。完全ログ未公開	入力差を排除するため重要 ¹¹
自動生成された全検索式	未確認	検索再現性の核心
9月試験の候補母集団総件数	未取得と会社自身が明記	偶然確率・precision評価に必要 ⁵
上位30件の全公報番号・順位	公開資料では未確認	26/26を第三者が照合するため必要
使用LLMのモデル・バージョン	未確認	モデル更新による変動を管理するため重要
temperature、seed等	未確認	確率の変動を評価するため重要
複数回実行による平均・分散	未確認	LLMのrun-to-run変動の評価に必要
シソーラス学習・構築データのcutoff	未確認	ベンチマーク汚染の排除に必要
独立第三者による再実行	未確認	自己評価バイアス排除に必要

特に問題になるのは**テストセットの再利用**である。2026年2月、同じ過去問に対して機械は8件中7件、化学・医薬は14件中12件しか到達しなかった。同社はその結果を分析し、「母集団を増やしても届かない漏れ」があると認識した。その後、語彙不足を解消する「類語シソーラス拡張」を実装し、9月に同じ過去問へ再適用して26/26を得た。³²

これは**製品改良のA/B評価としては説得力がある**。同一問題に対して変更前後を比較しているため、「新機能によりこの問題の失敗が解消された」ことは比較的理解しやすい。一方で、**未知データへの一般化性能を測**

る実験としては弱い。2月の失敗が開発チームに知られた後の9月再試験なので、過去問2024は事実上、開発中に参照された評価集合となっているからである。 33

また、類語シソーラスは過去の拒絶理由通知書10万件以上を原料としている。2024年度問題の対象特許や、意味的に極めて近い案件に関連する拒絶理由通知書がこの10万件に含まれていたか、それらをベンチマーク前に除外したか、データcutoffを設けたかは**未確認**である。含有していたと断定する根拠はないが、独立ベンチマークとして扱うなら、対象文献・ファミリー・引用ネットワーク・関連審査案件を学習／シソーラスコーパスから除外したことを示す必要がある。 34

統計的評価

26/26を単純に「26個の独立な成功」と扱うこと自体に強い仮定が必要だが、参考として二項モデルを当てると、26回すべて成功した際のClopper-Pearson 95%信頼区間下限は約**86.8%**となる。すなわち、26/26という観測だけから母集団性能が99%や100%と断定できるわけではない。

さらに実態として、26文献は独立にランダム抽出された26試験ではなく、**わずか3つの技術課題にクラスタリングされた正解文献**である。「独立な課題」を単位に3/3と見ると不確実性ははるかに大きい。会社自身も「統計的な有意性を主張する規模ではない」と明記しており、この点は同社の留意事項と一致する。 5

ランダムランキングによる偶然確率も、現在の公開情報では計算不能である。候補総数を N 、正解数を m とし、完全にランダムに順位を付けるなら、全正解が上位30件に入る確率は、

$$P = \frac{\binom{30}{m}}{\binom{N}{m}}$$

で表せる。しかし各分野の現在の N が公開されていないため、実際の確率は出せない。そもそもアルゴリズムはランダムランキングではないため、本質的な問題は「偶然」より**評価セットへの適応・漏洩・選択バイアス**である。母集団総数が未取得であることは同社自身も認めている。 5

「査読時間約70%削減」も実測結果ではない。記事では**1件3分で読むと仮定**し、100件なら約5時間、30件なら約1.5時間として算術的に70%削減している。したがってこれは「実際のユーザーで平均70%短縮した」というユーザースタディではなく、文献数に比例するという仮定上の試算である。 18

他の特許検索ツール・生成AIとの比較

今回の見出しと完全に同じ「2024問2・設問(4)・AIスクリーニング後Recall@30」条件で、Perplexity、ChatGPT、Gemini等を再実行した**比較結果は公開資料では未確認**である。この点は重要であり、9月の100%を競合製品に対する直接優位性として解釈してはいけない。

ただし同社は2026年2月、**同じ2024年度問題の設問(5)**について、サマリア、Perplexity Patents、GPT-5.2 Deep Research、Gemini 3.0 Deep Researchを比較している。条件は設問(4)のRecall@30とは異なるものの、比較表内では同じ正解集合に対する「正解文献を何件抽出できたか」という同一指標で評価している。同社は本来の大会では3件抽出のところを、評価安定性を理由に最大10件まで緩和している。 35

ツール／方式	電気 3件	機械 6件	化学・医薬 7件	3分野合計の正解 捕捉	マイクロ再現 率
サマリア・最終抽出工程	3/3 = 100%	3/6 = 50%	2/7 = 29%	8/16	50.0%

ツール／方式	電気 3件	機械 6件	化学・医薬 7件	3分野合計の正解 捕捉	マイクロ再現 率
Perplexity Patents	0/3 = 0%	0/6 = 0%	3/7 = 43%	3/16	18.8%
GPT-5.2 Deep Research	0/3	0/6	0/7	0/16	0%
Gemini 3.0 Deep Research	0/3	0/6	0/7	0/16	0%

数値は同社公開の比較画像から集計した。なお、同じ画像ではサマリアの「検索集合作成だけ」の結果も別途示され、最終抽出工程とは異なる。³⁶

この比較にも大きな注意書きがある。**比較実施者がサマリアの提供会社自身**であるうえ、同社自身が、GPT-5.2 Deep ResearchとGemini 3.0 Deep Researchはそもそも特許のコンプライトサーチ専用製品ではなく、Perplexity Patentsも同じ目的のサービスではないと明記している。したがって、「0%だから汎用LLMは特許調査に使えない」という一般化はできない。³⁷

また、この2月比較は**9月の見出しと同じ評価条件ではない**。整理すると次のようになる。

比較条件	サマリア	Perplexity Patents	GPT Deep Research	Gemini Deep Research
2024問2 設問(4)、9月版、 Recall@30	100% (3分 野)	未確認	未確認	未確認
2024問2 設問(5)、2月版、 最大10件抽出	公開あり	公開あり	公開あり	公開あり
独立第三者による同条件比 較	未確認	未確認	未確認	未確認

このため、「サマリアだけがRecall@30=100%で競合はそれ以下」とする比較広告を、現在の資料から作ることはできない。

学術研究と照らすと、今回のサマリア検証がどの程度のエビデンスレベルかも見えやすい。Helmetsらは特許先行技術検索について全文類似度を使った機械学習検索を評価し、ドメイン専門家による検索結果品質評価を組み込んでいる。³⁸

PatentMatchは、EPOの技術的専門性を持つ審査官によって、クレームと引用先行技術の文章ペアが新規性を害するほど対応するかをラベル付けしたデータセットである。このような**固定された外部ラベル付きベンチマーク**は、モデル間比較の再現性を高める代表的アプローチである。³⁹

より近年のPatentFinder研究では、化学構造を含む特許侵害評価という別タスクながら、**MolPatent-240**という明示的なベンチマークを構築し、LLMや化学ツールのベースラインとF1・accuracyで比較している。ただしタスクそのものがサマリアの一般的先行技術検索とは異なるため、その数値を直接Recall@30と比べることはできない。⁴⁰

つまり、学術的な意味での次の一步は、**問題・正解・候補集合を固定し、複数製品を同じk、同じ時点、同じ入力情報量で評価し、全順位とコード／設定を保存すること**である。現在のサマリア結果は有望な製品ベンチマークだが、その水準の比較試験にはまだ達していない。

潜在的バイアス・限界・誤解を招き得る表現

最も大きな問題は、「**再現率100%**」という表現の射程である。正確には、

「2024年度の特定3課題について、競技大会の既知X文献26件すべてが、この時点のサマリアによるランキング上位30件内に入った」

という主張である。 ⁶

これは、「未知の全関連特許を100%見つける」「実務の全案件で検索漏れがゼロ」「世界中の特許を完全検索できる」という意味ではない。正解集合は競技大会で定義されたX文献であり、現実世界に存在するすべての潜在的先行技術を網羅的にrelevance judgmentした集合とは性質が異なる。

主要な限界を整理すると、以下のようになる。

自己評価バイアス。 製品を開発・販売するパテント・インテグレーション自身が試験設計、実行、評価、公表を行っている。PR TIMESは配信プラットフォームであって、性能値を独立検証する査読機関ではない。

¹¹

テストセット再利用。 同じ2024年度問題を2月に試し、取りこぼした文献を認識した後、シソーラス機能を追加して9月に同じ問題を再試験している。これは「未知問題への外挿性能」ではなく、「既知の失敗ケースに対する製品改善」を強く測っている。 ³²

データ汚染を排除する手続きが未確認。 類語シソーラスの基になる10万件超の拒絶理由通知書に、評価対象文献やその近縁案件の情報が含まれているかどうか、その除外ルール、時間的cutoffは公開資料から確認できない。これは漏洩があったという意味ではなく、**漏洩がなかったことを独立に検証できない**という意味である。 ³⁴

サンプル数が少ない。 3分野、26正解文献のみで、会社自身も統計的有意性を主張しないと記している。技術分野も電気、機械、化学・医薬の1課題ずつであり、同じ「化学」でも材料、ポリマー、バイオ、医薬化合物などの難易度差を評価するには足りない。 ⁵

上位30件は大会公式指標ではない。 「特許検索競技大会で100%」という見出しから、大会主催者がサマリアを公式採点して満点と認定したように読める余地があるが、公開された資料が示すのは、過去問を利用してベンダー自身が作ったランキングを後から評価した結果である。 ⁸

母集団の大きさが不明。 会社自身が総件数未取得とするため、本来のprecision、false-positive rate、random baselineを厳密に求めにくい。上位30件に全正解があることは実務的に重要だが、それだけで検索空間全体の精度を特徴付けることはできない。 ⁵

正解集合の完全性。 競技大会のX文献は評価用gold setとして有用だが、「その発明について実際に存在するすべての関連先行技術」を網羅した完全集合と同一ではない可能性がある。そのため26/26は、厳密には「既知gold文献に対する100%」である。

生成AIの反復安定性が不明。1回の実行結果なのか、複数回実行してすべて100%だったのか、モデル更新や乱数で順位が変わるかは未確認である。現在のサマリアは複数のOpenAI、Claude、Geminiモデルを利用可能としているため、モデル・バージョン固定は再現試験に不可欠である。 41

時間削減70%はモデル計算。上位文献数が100から30に減ったことに、1件3分という一定速度を掛けた推計であり、実利用者をランダム化して計測した生産性試験ではない。 18

人的独立性についても透明化余地がある。サマリア公式サイトは技術顧問の角淵由英氏について、特許検索競技大会での受賞歴に加え、大会実行委員会との関係を紹介している。これは不正や不適切な情報利用を意味するものでは全くないが、競技大会の過去問を製品ベンチマークとして使うのであれば、独立性に疑義を残さないため、正解データへのアクセス時期、開発者が事前に関連した情報、ブラインド化手順を明示することが望ましい。大会側一次資料による同氏の現在の役職確認は本調査では**未確認**である。 42

反対に、今回のPRを過度に否定的に評価するのも適切ではない。会社は2月時点の不完全な結果（機械7/8、化学・医薬12/14）を隠さず公開し、9月記事でもサンプル数不足・統計的有意性なし・母集団総数未取得・適用範囲不明という重大な留意事項を自ら書いている。これは、単純に「100%だから万能」とだけ広告する場合より透明性が高い。 9

したがって問題は「100%という数字が虚偽らしい」というより、**100%という数字が何についての100%なのかを見出しだけでは誤読しやすいこと**にある。

結論と特許調査実務への示唆

本調査で最も妥当な判定は、「**主張は限定されたベンダーベンチマークとしては成立しているが、一般性能の独立実証には至っていない**」である。

サマリアが2024年度特許検索競技大会アドバンストコースの3分野で、既知のX文献計26件を検索集合にすべて含め、さらにAIスクリーニング後の上位30件以内にすべて順位付けした、という発表内容には、公開されたPR TIMES本文、公式マニュアル、2月の旧ベンチマークとの間で整合性がある。特に2月の7/8、12/14という失敗から、審査官認定語彙を利用するシソーラスを導入して26/26へ改善したという前後比較は、**検索語彙の拡張が特許検索の取りこぼし低減に寄与し得る**という技術的仮説と整合する。異なる専門語彙が先行技術検索を難しくすること自体は学術研究でも認識されている。 43

一方、今回の数字だけを理由に、専門調査者による検索式レビュー、FI/Fターム／IPC等の分類検索、引用・ファミリー追跡、審査経過確認、最終的な全文査読を省略することは推奨できない。同社自身も以前のベンチマークで、人間による最終確認の併用が望ましいとしている。 44

実務上は、サマリアを「**完全自動で答えを確定する装置**」ではなく、**高再現率の候補集合を作り、人間が読む順序を最適化する検索・トライアージシステム**として評価するのが適切である。この使い方なら、「正解が上位30件に集中した」という結果は非常に実務的な意味を持つ。100件読む代わりに30件を先に確認して重大文献を発見できるなら、たとえ100%という一般保証がなくても価値は大きい。 18

導入を検討する企業・特許事務所には、PR値をそのまま採用判断に使うのではなく、**自社案件での受入試験**を勧める。最も望ましい検証設計は次のようなものである。

推奨検証	方法	合格判断の例
未見案件テスト	サマリア開発側が未見の過去調査案件を30～100件用意	Recall@30 / @50を案件別に測る

推奨検証	方法	合格判断の例
時間分割	シソーラス構築cutoffより後の案件だけをテスト	データ漏洩を排除
ファミリー除外	gold文献と同一ファミリー・直接引用案件を学習側から除外	厳しい外部妥当性確認
競合同条件比較	サマリア、商用DB、人手、他AIに同じ入力・同じ期限を設定	Recall@k、Precision@k、MRR/nDCG等を比較
反復試験	AIを同条件で5～10回実行	Recall@30の最低値・平均・標準偏差を測る
ブラインド評価	正解を知らない担当者が検索を実施し、別担当者が採点	benchmark tuningを低減
工数計測	実際の調査者の操作・査読時間を記録	「70%削減」を実測値で検証
エラー分析	見逃し文献を語彙、分類、翻訳、図面、化学構造等に分類	製品が弱い案件を特定

特に第三者検証として説得力が高いのは、**2025年度以降の未使用問題や新たに専門家が作成した問題を、正解非開示のまま固定版サマリアに投入すること**である。問題入力、生成検索式、候補集合総数、全順位、モデル名・バージョン、実行時刻、シソーラスのデータcutoffをログとして凍結し、その後に正解を開封すれば、今回の最大の弱点である「同じ2024問題を改良前後に繰り返し使った」問題を解消できる。

さらに、本当に「上位30件で100%」の実務的価値を主張するのであれば、単なる26/26だけでなく、案件ごとの**Recall@10、@20、@30、@50、@100**の曲線と、**Precision@k、平均最終正解順位、最悪正解順位**を公開するとよい。これにより、単一カットオフを選択したことによる印象操作の余地も小さくなる。

学術的にも、PatentMatchのような審査官ラベル付きデータセットは、モデルを訓練・評価する際に固定された外部gold standardを持つことの重要性を示している。サマリアの「審査官が実際に対応すると認定した語彙を利用する」という発想そのものは、特許特有の表現差を扱う方法として合理的かつ興味深い、その技術的優位性を強く主張するには、より大規模な未見データでの検証が必要である。⁴⁵

最終的な評価を一文でまとめるなら、

「再現率100%」は虚偽と判断すべき数字ではないが、“2024年度の既知3課題・26正解文献に対するRecall@30が100%だった”という限定条件を外して読むべき数字でもない。

現段階で確認できる証拠は、サマリアの検索・ランキング能力が**有望であり、特に専門語彙の展開と人間の査読候補の圧縮に実務的価値があり得る**ことを示している。一方、一般的な「特許検索性能100%」を証明するには、第三者、未見問題、大規模サンプル、固定された母集団、完全な実行ログを用いた次段階の検証が必要である。⁴⁶

1 2 4 5 6 9 10 11 12 13 18 20 29 32 33 43 46 生成AIによる特許調査サービス「サマリア」、特許検索競技大会の過去問検証で電気／機械／化学・医薬の3分野で再現率100%(上位30件)を達成 | パテント・インテグレーション株式会社のプレスリリース

<https://prt-times.jp/main/html/rd/p/0000000028.000086119.html>

- 3 8 14 16 19 [japio.or.jp](https://japio.or.jp/00yearbook/files/2018book/18_1_07.pdf)
https://japio.or.jp/00yearbook/files/2018book/18_1_07.pdf
- 7 15 [特許検索競技大会 過去問](https://japio.or.jp/?utm_source=chatgpt.com)
https://japio.or.jp/?utm_source=chatgpt.com
- 17 [弁理士のための特許調査の知識](https://jpaa-patent.info/?utm_source=chatgpt.com)
https://jpaa-patent.info/?utm_source=chatgpt.com
- 21 30 34 [類語シソーラス拡張 | 特許文書読解アシスタント・サマリア](https://patent-i.com/summaria/manual/recognition_thesaurus)
https://patent-i.com/summaria/manual/recognition_thesaurus
- 22 [新機能「サマリアMCP」・「類語シソーラス拡張」機能のリリース ...](https://patent-i.com/summaria/manual/R_20260901?utm_source=chatgpt.com)
https://patent-i.com/summaria/manual/R_20260901?utm_source=chatgpt.com
- 23 41 [A I モデル一覧 | 特許文書読解アシスタント・サマリア](https://patent-i.com/summaria/manual/available_models)
https://patent-i.com/summaria/manual/available_models
- 24 42 [サマリア\(Summaria\) | 特許文書読解支援サービス](https://patent-i.com/summaria/)
<https://patent-i.com/summaria/>
- 25 26 [利用料金 | 特許文書読解アシスタント・サマリア](https://patent-i.com/summaria/fee)
<https://patent-i.com/summaria/fee>
- 27 [サマリアMCP | 特許文書読解アシスタント・サマリア](https://patent-i.com/summaria/manual/mcp)
<https://patent-i.com/summaria/manual/mcp>
- 28 [収録公報・公報番号の仕様 | 特許文書読解アシスタント・サマリア](https://patent-i.com/summaria/manual/database)
<https://patent-i.com/summaria/manual/database>
- 31 35 37 44 [特許検索競技大会の問題に「サマリア」が挑戦。集合上位100件について85%以上の再現率を達成！ / 導入実績360社超を誇る中小企業向け知財管理システム「root ip クラウド」とサービス連携を開始 | パテント・インテグレーション株式会社のプレスリリース](https://prt-times.jp/main/html/rd/p/000000018.000086119.html)
<https://prt-times.jp/main/html/rd/p/000000018.000086119.html>
- 36 [prcdn.freetls.fastly.net](https://prcdn.freetls.fastly.net/release_image/86119/18/86119-18-fbe8c52f6f98baa645f5c8d0a960dfbc-2516x1462.jpg?auto=webp&bg-color=fff&fit=bounds&format=jpeg&height=1350&quality=85%2C65&width=1950)
https://prcdn.freetls.fastly.net/release_image/86119/18/86119-18-fbe8c52f6f98baa645f5c8d0a960dfbc-2516x1462.jpg?auto=webp&bg-color=fff&fit=bounds&format=jpeg&height=1350&quality=85%2C65&width=1950
- 38 [Automating the search for a patent's prior art with a full text similarity search](https://arxiv.org/abs/1901.03136?utm_source=chatgpt.com)
https://arxiv.org/abs/1901.03136?utm_source=chatgpt.com
- 39 45 [PatentMatch: A Dataset for Matching Patent Claims & Prior Art](https://arxiv.org/abs/2012.13919?utm_source=chatgpt.com)
https://arxiv.org/abs/2012.13919?utm_source=chatgpt.com
- 40 [Intelligent System for Automated Molecular Patent Infringement Assessment](https://arxiv.org/abs/2412.07819?utm_source=chatgpt.com)
https://arxiv.org/abs/2412.07819?utm_source=chatgpt.com