ChatGPT

Gemini 3.5 Flashのエージェント性能向上と特許調査実務への含意

エグゼクティブサマリー

Gemini 3.5 Flashは、Googleが2026年5月に公開した「Flash」系の最新中核モデルであり、Google自身はこれを「持続的なフロンティア性能を持つ、エージェント実行とコーディング向けの最も知的なFlashモデル」と位置づけています。公式ブログでは、Gemini 3.1 ProをGDPval-AAで**1656 Elo**、Terminal-Bench 2.1で**76.2%**、MCP Atlasで**83.6%**上回るとされ、出力速度は「他のフロンティアモデルより4倍速い」と主張されています。公開仕様上は**1Mトークンのコンテキスト**、**64K出力**、**Function calling**／**Structured output**／**Search as a tool**／**Code execution**をサポートし、長文・多段・ツール呼び出し型の業務設計に向いています。もっとも、詳細なアーキテクチャ差分やパラメータ規模は公開されておらず、性能向上の根拠は主として評価結果と製品機能から読み解く必要があります。 [1]

IP実務の観点では、この進化は特に**先行技術検索の前処理・探索設計・候補統合・根拠抽出・報告書草案化**に効きます。理由は、特許調査が本質的に「長文読解」「多ソース探索」「分類体系の取り扱い」「証拠の抜粋」「反証・見落とし確認」を要するからです。JPOはすでにAIによる**概念検索、ランキング表示、検索語・分類の提案、画像検索**を導入・実証段階に進めており、WIPOも27庁で70件超のAI施策、うち13件が先行技術検索関連と整理しています。すなわち、**“特許検索をAIで強化する方向性”自体は、研究室の仮説ではなく、各国庁の実装トレンド**です。 [2]

ただし、**GDPval-AA 1656**をそのまま「特許調査の自動化水準」と読むのは過大評価です。GDPval-AAは、OpenAIのGDPval公開ゴールドセット220タスクを、Artificial Analysisの**Stirrup**ハーネス上で**Web検索とシェルアクセス**付きエージェントとして解かせ、**Gemini 3.1 Pro Previewによる盲検ペア比較**から**Bradley–Terry/Elo**でランキング化する評価です。つまり、これは「経済価値のある知識業務における汎用エージェント性能」の代理指標であって、**特許・無効資料・FTO・クレームチャート**のような法的高精度業務を直接測っているわけではありません。しかも元のGDPval論文でも、公開ゴールドセットは220件に限られ、評価の基準は本来**人間専門家のペア比較**です。 [3]

本報告の結論は明確です。**Gemini 3.5 Flashは、複数エージェントによる特許調査・先行技術検索ワークフローの“実用化可能性”を確かに引き上げる**一方、**完全自律の法的判断系ワークフローを直ちに許容するものではない**、という評価が最も妥当です。PoCの主眼は「人を置き換えること」ではなく、**見落とし低減、一次探索の広化、根拠整理の高速化、再現性の向上**に置くべきです。特に導入初期は、**人間レビューを前提とした“証拠付き検索補助エージェント”**として運用するのが妥当です。 [4]

調査の前提と結論

まず前提として、Googleの公開資料は**Gemini 3.5 Flash固有の詳細内部構造**をほとんど開示していません。Model Cardは「Gemini 3.5 Flash is based on Gemini 3 Flash」と述べ、Gemini 3 Flash Model Cardはさらに「Gemini 3 Flash is built off of the Gemini 3 Pro reasoning foundation」と説明しています。さらにGemini 3 Pro側では、**疎なMixture-of-Experts型のネイティブ・マルチモーダルTransformer**であること、TPU・JAX・ML Pathwaysで訓練されたことが明記されています。したがって、本件でいう「アーキテクチャ改善」は、**3.5 Flashで新規に何が変わったかを詳細分解する**というより、**3 Pro → 3 Flash → 3.5 Flashの継承関係と評価改善から推定可能な範囲で論じる**のが限界です。 [5]

また、Googleのドキュメント群には**公開チャネル間の記載差**があります。AI for Developersの「What's new」ではGemini 3.5 Flashを**GA stable**としていますが、DeepMindモデルページの取得時点表示では**Preview**でした。出力上限も、What's newでは**65k max output tokens**、Model Card/モデル情報ページでは**64K output**と記載されています。本報告では、こうした表記差を踏まえ、**API実装時には必ず実際の利用リージョン・課金ページ・モデルエンドポイント仕様を再確認する**という前提を置きます。 6

そのうえで総合判断を述べると、Gemini 3.5 Flashの価値は、単純な「賢いチャットボット」ではなく、**長文文脈下で、外部ツールを呼び出しながら、多段の探索・統合・修正を続けられるエージェント基盤**にあります。特許実務では、これは「検索式を立てる」「分類を組み合わせる」「並列に文献探索する」「根拠段落を抜く」「別エージェントで反証する」「報告草案を作る」という**分業型ワークフロー**に直結します。ゆえに実務導入の論点は「3.5 Flashが単体で弁理士の代わりになるか」ではなく、**人間の知財担当者が監督する多段検索パイプラインの中核として使えるか**です。 7

技術的背景とGemini 3.5 Flashの改善点

技術的な改善点の第一は、**推論基盤の継承と"thinking"制御の成熟**です。Gemini 3 Flashは3 Proの推論基盤をベースとし、**thinking levels**で品質・コスト・レイテンシの混合を制御する設計とされています。3 Proは疎MoE・ネイティブマルチモーダルです。3.5 Flashはその3 Flash系を継承しつつ、GoogleはブログとModel Cardで、**agentic execution、coding、long-horizon tasks**の前世代改善を前面に出しています。つまり、Googleが公開している改善点は、内部パラメータ変更よりも、**推論時の持続性能、ツール利用性能、長期タスク安定性**に重点があると読むのが妥当です。 8

第二は、**速度と価格体系の再設計**です。Googleは公式ブログで3.5 Flashについて「output tokens per secondで他のフロンティアモデルの4倍」と主張し、長期エージェント作業を「しばしば他のフロンティアモデルの半分未満のコスト」でこなせると述べています。価格面では、Gemini Developer APIの標準課金で**入力$1.50/100万トークン、出力$9.00/100万トークン**、Batch/Flexで**入力$0.75、出力$4.50**、Priorityで**入力$2.70、出力$16.20**です。Flexは公式に**標準比50%コスト削減**と説明されています。したがって、対話的な検索方針設計には標準、夜間の再評価や候補文献のバッチ再スコアリングにはBatch/Flexを回す、という分離が実務的です。 9

第三は、**長文文脈の維持**です。Gemini 3.5 Flashは**1Mトークン入力**をサポートし、知識カットオフは**2025年1月**です。特許実務では、発明概要、請求項、明細書、引用候補、審査経過、社内技術メモ、分野別シソーラス、分類候補を一度に抱えられるかが重要ですが、1Mトークン級の窓はこの点で有利です。ただし、MRCR v2の1M pointwiseスコアは**26.6%**であり、長文窓そのものが万能ではないことも公式評価が示しています。つまり、**広い窓は"入る"ことを意味するが、"全部を正しく扱える"ことは意味しない**ため、実装上は依然として分割・要約・再検索が必要です。 10

第四は、**ツール連携の標準化**です。DeepMindのモデル情報ページでは、3.5 Flashは**Function calling、Structured output、Search as a tool、Code execution**をサポートします。Gemini APIのManaged agentsは、**単一API呼び出しでLinuxサンドボックスをプロビジョニングし、推論・コード実行・ファイル管理・Web閲覧**を行えると説明されています。さらにDeep Research Agentは、デフォルトで**Google Search、URL Context、Code Execution**を持ち、**background=true**で長時間タスクを非同期実行し、**MCPサーバ接続**も可能です。これは、社内ナレッジ・外部特許DB・PDF処理・表計算生成・比較表作成を跨ぐIPワークフローと相性がよい設計です。 11

第五は、**マルチエージェントの実装可能性**です。GoogleはブログでAntigravity上の**collaborative subagents**を紹介し、複数エージェントが資産整理やゲーム開発を分担する例を示しています。公開の開発基盤側でも、ADKは**複数の専門エージェントを階層的に配置**でき、**SequentialAgent、ParallelAgent、LoopAgent**、さらにグラフベースの柔軟なワークフローを提供します。ParallelAgentは独立タスクを並列化してレイテンシを削減し、LoopAgentは改善条件が満たされるまで反復でき、Artifactsは**バージョン付きファ**

イルとして保存できます。これは、特許調査で必要な**分担、並列探索、相互検証、改稿の反復、証拠保全**にほぼそのまま対応します。 [12]

ただし制約もあります。Googleは明示的に、**Computer UseはGemini 3.5 Flashでは現時点未対応**としています。そのため、J-PlatPatや各庁ポータルを人間の代わりにブラウザ操作で完全自動巡回する、という発想は現時点では主戦略にしにくく、**API・検索ツール・URL取得・ファイル検索中心の設計**が現実的です。特許調査系では、まず**検索API/データコネクタ経由**で候補収集し、必要箇所だけ人間がポータルUI確認する構成がよいでしょう。 [13]

モデル比較表

項目	Gemini 3.5 Flash	Gemini 3 Flash Preview	Gemini 3.1 Pro Preview	IP実務上の意味
系譜	Gemini 3 Flashベース	Gemini 3 Pro推論基盤ベース	Gemini 3系上位推論モデル	3.5 Flashは「軽量＝低性能」ではなく、上位系譜を継ぐ高性能Flash
入力コンテキスト	1M	1M	1M	明細書・請求項・引用候補・社内資料の同時保持に有利
出力上限	64K〜65K表記差あり	64K	64K	長い比較表・報告書草案を一回で生成しやすい
標準ツール	Function calling / Structured output / Search / Code execution	同系統	同系統	検索・抽出・構造化出力の自動化に直結
標準料金	入力$1.50 / 出力$9.00	入力$0.50 / 出力$3.00	入力$2.00〜4.00 / 出力$12.00〜18.00	3.5 FlashはProより安く、旧3 Flashより高い
Batch/Flex	いずれも標準比ほぼ半額	Batchあり	Batchあり	夜間再評価や候補再スコアリング向き
エージェント訴求	持続的エージェント性能・coding	speed重視の高性能Flash	最上位推論・マルチモーダル	3.5 Flashは「速度×エージェント」最適化が強み

出典はGoogle DeepMind/Google AI for Developersのモデルカード、モデル情報、価格ページ、リリースノートです。なお3.5 Flashの出力上限と状態表示には公式ページ間で差があり、実装時は実エンドポイントの仕様確認が必要です。 [14]

GDPval-AA Elo 1656の評価

GDPvalの原典はOpenAIの評価研究で、**米国GDP上位9セクター・44職種・1320タスク**を対象とし、業界専門家が実務成果物をもとに作成したタスクで構成されます。専門家の平均経験年数は**14年**で、公開されたゴールドセットは**220タスク**です。元論文における主要評価は、**人間専門家による盲検ヘッド・トゥ・ヘッド**

比較です。この点が重要で、GDPvalは学力試験型でなく、**“現実の知識労働成果物の品質比較”**に重心があります。 [15]

GDPval-AAは、このGDPval公開セットをArtificial Analysisが**共通ハーネスで再実装**したものです。Methodologyページによれば、評価対象モデルは**Stirrup**というオープンソース・エージェントハーネス上で動かされ、**Web Fetch、Web Search、View Image、Run Shell、Finish**の5種ツールを用い、各タスクごとに新しいサンドボックスが与えられます。提出物の比較は、**Gemini 3.1 Pro Previewが盲検で2出力を比較**し、その結果を**Bradley–Terryモデルに当てはめ、ブートストラップ信頼区間付きのElo**に変換します。つまりGDPval-AAは、**“同じ外部環境下でのエージェント性能比較”**に価値がある一方、**評価者も人間ではなくLLM**であることを忘れてはいけません。 [16]

Gemini 3.5 Flashの**1656 Elo**は、GoogleのModel Cardでは**Gemini 3 Flashの1204**、**Gemini 3.1 Proの1314**を明確に上回り、一方で**Claude Sonnet 4.6の1676**、**Claude Opus 4.7の1753**、**GPT-5.5の1769**はなお上にあります。したがって1656は、**Gemini系としては大幅前進**だが、**市場全体の最上位ではまだ僅差〜中差で追う位置**です。「注目すべき向上」という元の問題意識は妥当ですが、「すでに最強だから採用すべき」というほど単純ではありません。実際には、**速度・コスト・ツール連携・ガバナンス**との総合判断が要ります。 [17]

さらに、GDPval-AAの**1656**は“生の法務実務品質”ではなく、**共通ハーネス上の相対順位**です。元のGDPval論文では、公開ゴールドセットの自動評価器は**人間との平均一致65.7%**、人間同士の一致は**70.8%**でした。また論文は、タスク数がまだ限定的であること、評価コストが高いこと、自動評価器にはインターネット依存タスクや非Python環境などの制約があることを明確に認めています。Epoch AIも、GDPval-AAは**人間成果物との絶対的なwin+tieを直接出せない**こと、**ライブWeb依存のタスクは経時ドリフトしうる**ことを指摘しています。したがって、1656は有用なシグナルですが、**特許調査の品質保証指標として単独採用すべき数値ではありません。** [18]

ベンチマーク比較表

ベンチマーク	Gemini 3.5 Flash	Gemini 3 Flash	Gemini 3.1 Pro	Claude Sonnet 4.6	Claude Opus 4.7	GPT-5.5
GDPval-AA Elo	1656	1204	1314	1676	1753	1769
Terminal-Bench 2.1	76.2%	58.0%	70.3%	—	66.1%	78.2%
MCP Atlas	83.6%	62.0%	78.2%	69.5%	79.1%	75.3%
Toolathlon	56.5%	49.4%	—	—	—	55.6%
MRCR v2 1M	26.6%	22.1%	26.3%	—	—	—

この表が示すのは、Gemini 3.5 Flashが**Gemini系内部では大幅改善**であり、特に**エージェント・ツール利用系**で伸びていることです。一方で、GDPval-AA単独では競合最上位を決定づけるほどではなく、**「速度・価格を含めた最適解」**として検討するのが現実的です。 [19]

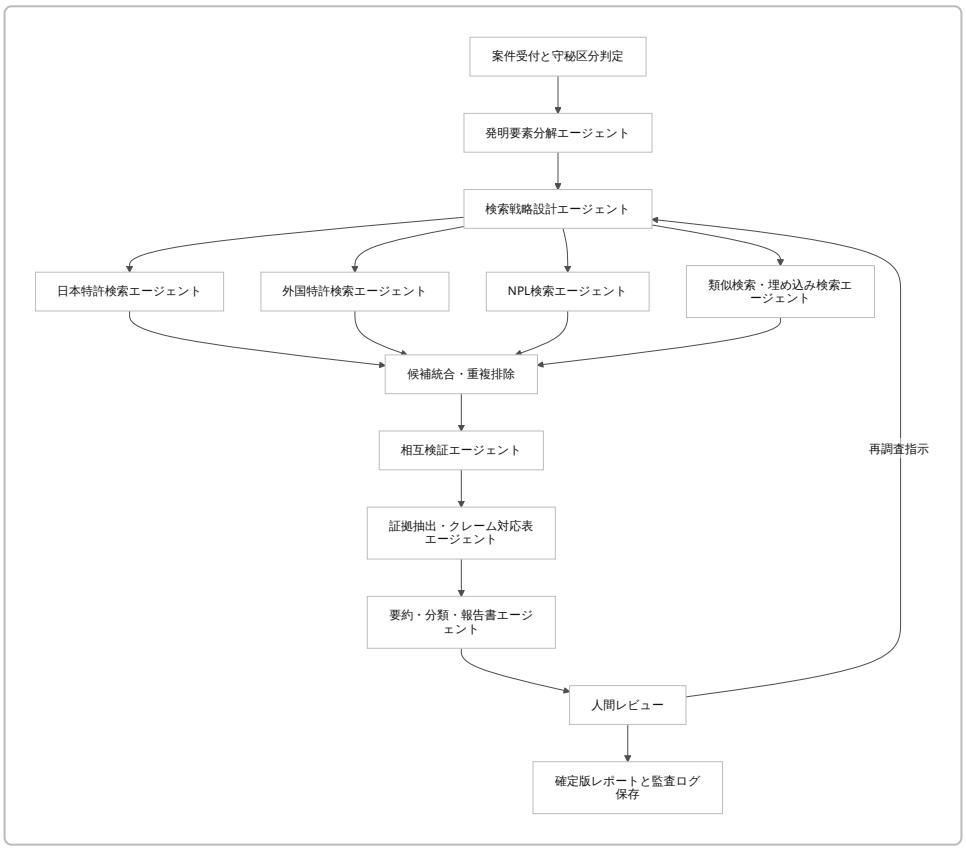
特許調査と先行技術検索への適用設計

特許調査にこのモデルが効く理由は、実務がもともと**多段・多ソース・多観点**だからです。J-PlatPatは、先行技術調査の文脈で**外国文献や論文等も調査可能**であり、現行のINPIT資料群でも**論理式、ワイルドカード、近傍検索**が重要な検索手法として説明されています。PATENTSCOPEはPCTを中核に世界のコレクション検索を提供し、EPOは世界最大級の前技術コレクションと**NPL引用27.8%**という数字を示しています。つまり、

先行技術検索は**特許文献だけ見ればよい業務ではなく、分類・キーワード・非特許文献・外国文献を束ねる業務**です。ここにマルチエージェントの分担運用が適合します。 [20]

さらに、JPO自身がAIについて、**概念検索、ランキング表示、検索手法高度化、分類付与、検索語・分類提案、画像検索**を導入・実証していることは重要です。JPAA誌の記事でも、先行技術調査では**類似文章ランキング**がBM25より語順・同義語処理で優位性を示しうる一方、入力長制限など実務課題が残ること、生成AIでは**精度評価の難しさ、課題の陳腐化、ハルシネーション**が課題になることが整理されています。したがって、Gemini 3.5 Flashの役割は、**既存検索の置換**よりも、**既存検索を上流・下流から補強する“オーケストレータ”**として設計する方が成功確率が高いと言えます。 [21]

推奨ワークフロー図



この構成は、Google ADK/Cloud Architectureの**Sequential → Parallel → Loop**の発想と整合します。構造化された段取りは順次実行し、候補探索は並列化し、論点争いが出たら再探索ループに戻す、というのが最も自然です。 [22]

推奨ワークフロー表

段階	主担当	主な処理	主なデータ/ツール	成果物	重点管理点
受付・分解	発明要素分解エージェント	課題・構成要件・作用効果・代替表現を抽出	発明メモ、請求項草案、図面、社内用語集	構成要件表、同義語表	機密区分、対象法域

段階	主担当	主な処理	主なデータ/ツール	成果物	重点管理点
戦略設計	検索戦略設計エージェント	キーワード、同義語、FI/Fターム/CPC/IPC候補、除外条件を設計	JPO/INPIT分類知識、過去案件、分類辞書	検索式案、分類候補	網羅性とノイズ率の均衡
並列探索	複数検索エージェント	日本、外国、PCT、NPL、類似検索を同時実行	J-PlatPat、PATENTSCOPE、EPO系DB、社内RAG	候補文献集合	由来、タイムスタンプ、再実行性
相互検証	クリティークエージェント	見落とし仮説、反対解釈、ノイズ除外検証	候補文献、検索ログ	再調査指示、差分リスト	反証手順の明示
根拠抽出	証拠抽出エージェント	クレーム要素ごとの対応箇所抽出	明細書、候補文献PDF、画像	クレームチャート草案	引用箇所の誤対応防止
報告・承認	要約エージェント＋人間	主要文献、対比、残課題、追加調査要否を整理	全ログ・証拠・評価メモ	調査報告書	最終判断は人間が行う

出典の基礎は、JPOのAI検索高度化方針、INPITの検索式・外国文献/NPL導線、Google ADKの並列・順次・ループ型ワークフロー、File Search/Artifactsのバージョン管理・引用機構です。表中の具体的な役割分担は、これらを踏まえた本報告の実装提案です。 [23]

期待効果は四つに整理できます。第一に**精度**で、分類・語彙・構成要件・NPLを別エージェントが持つことで、単一検索式の偏りを弱められます。第二に**網羅性**で、ParallelAgent型の同時探索により、国内・外国・NPL・社内知識の取りこぼしを減らせます。第三に**効率**で、収集・要約・表形成・クレーム対応表草案の機械化により、調査担当者は“読むべき数十件”に集中しやすくなります。第四に**コスト**で、探索方針の設計や理由付けだけ3.5 Flashを使い、広い候補群の再スコアリングはFlash-LiteやBatch/Flexに落とす構成にすると、品質とコストの両立余地があります。 [24]

実装要件としては、**データソース、API、検索式、メタデータ管理**が鍵です。Google API側では、File Search Storeに**custom metadata**を付与でき、埋め込みは**削除まで保持**、元ファイルは48時間後に削除されます。Grounding metadataから引用やPDFページ番号を返せるため、社内の過去調査報告、拒絶理由通知、意見書、技術資料を**証拠付きRAG**として再利用しやすい設計です。対外文献はJ-PlatPat/PATENTSCOPE/EPO等から収集し、社内では少なくとも**案件ID、法域、分類、検索式バージョン、取得日時、モデルID、プロンプト版、文献ファミリーID、引用箇所、担当レビューア**を永続保存すべきです。後者は本報告の推奨事項ですが、前者のAPI機能自体はGoogle公式が提供しています。 [25]

リスクとガバナンス

最大の実務リスクは、**もっともらしい誤り**です。Gemini 3 Pro Model Cardは明示的に、モデル一般の限界として**hallucinations**や時に**slowness/timeout**を挙げています。加えて、JPAA誌は生成AIを特許審査に使う際の課題として、**精度評価の困難さ、課題の陳腐化、ハルシネーション**を挙げています。特許調査では、誤りが「事実の誤認」ではなく、**見つけるべき引用文献の見落とし**として現れやすいため、通常のチャット用途より高い損失を生みます。 [26]

次に重要なのは、**バイアスと評価バイアス**です。GDPval-AA自体がGemini 3.1 Pro Previewを judge に使っており、原典GDPvalの人間評価と同一ではありません。元GDPvalでも、自動評価器は人間との一致が**65.7%**で、人間同士の**70.8%**を下回ります。よって、PoCでは単に「LLM-as-a-judge」の勝率を見るのではなく、**既知の引用文献をどれだけ再発見できるか、主要拒絶理由をどれだけ再構成できるか、クレーム要素対応表をどれだけ正しく埋められるか**のような業務メトリクスが必要です。 27

第三は、**証拠保全性・再現性・説明可能性**です。NIST AI RMFは、説明可能性の欠如がリスク測定を難しくし、説明可能なシステムは**デバッグ、モニタリング、文書化、監査**に有利だと述べています。Google ADKのArtifactsやGemini File SearchのGrounding metadataは、この要請への技術的部品になります。したがって特許調査用途では、最終レポートだけでなく、**どの検索式・どのモデル・どの文献断片・どの比較ロジックで結論に至ったか**を追跡できることが必須です。説明不能な“総合点だけ高い”エージェントは、IP実務では採用しにくいでしょう。 28

第四は、**機密情報管理**です。GoogleのTermsでは、**Unpaid Services**であるGoogle AI StudioやGemini API無償枠では、送信コンテンツと生成内容がGoogle製品・機械学習技術の改善に使われうると明記されています。一方、**Paid Services**ではプロンプトやレスポンスは製品改善に使われず、さらにGemini Developer APIでは承認制の**Zero Data Retention**も説明されています。Google Cloud側では、生成AI製品について**customer data is not used to train foundation models**とし、Vertex AIには**データ保持ゼロ**オプションもあります。結論として、**未公開発明や訴訟前提案件を扱うなら、無償枠や曖昧なデータ条件を避け、少なくともPaid＋必要ならZDR/Google Cloud系ガバナンスに寄せる**べきです。 29

最後に、**法的判断の代替はできない**という原則があります。EPOは、**prior art searchは法的に新規性の決定的証拠ではない**と明示し、検索は定期的に更新すべきだと述べています。これはAIでも同じで、検索が優れていても、**クレーム解釈、基準時、法域差、進歩性の組合せ論法、実施可能要件**まで機械に丸投げはできません。AIの役割は、**“見せる”“並べる”“根拠を抜く”“反証候補を出す”**までであり、**法的評価の責任主体は人間**であるべきです。 30

実務導入ガイド

PoCは、**業務全体の自動化実験**ではなく、**どの工程で品質を落とさずに工数を削減できるかを見極める設計**にするべきです。Google Cloudのagentic設計ガイドは、要求定義で**タスク特性、レイテンシ、コスト、人間関与**を先に決めるよう勧めています。特許調査は高リスク・高根拠性の業務なので、PoC対象はまず**先行技術検索の一次探索・候補統合・報告草案**に絞り、**最終結論は人間**とするのが正しい切り分けです。評価基盤としては、Gen AI Evaluation Serviceが**CSV/JSONL/BigQuery/Pandas**を受け、エージェント評価も可能です。 31

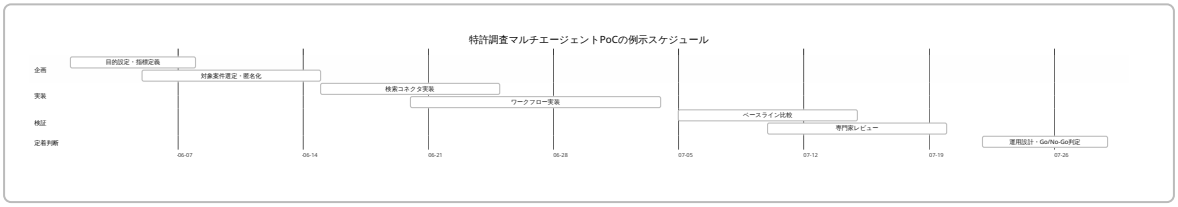
PoC設計の推奨値

項目	推奨設計
目的	既存調査フローの見落とし率を悪化させず、一次探索と根拠整理の工数を削減できるかを検証
対象案件	過去の実案件 50〜100件、少なくとも3技術分野、外国文献・NPL含有案件を含める
データセット	発明メモ、請求項、既存検索式、実際の引用文献、審査結果、社内レビューコメント
比較ベースライン	人手のみ、既存検索システムのみ、単一エージェント、マルチエージェント
定量指標	既知重要文献の再発見率、上位20件での関連文献率、主要引用の順位、NPL捕捉率、案件当たり時間、案件当たりコスト

項目	推奨設計
定性指標	クレーム対応表の妥当性、説明可能性、レビューしやすさ、誤引用の有無
成功基準	既知重要文献の再発見率を維持または改善しつつ、一次探索〜報告草案作成時間を有意に短縮
ガードレール	無償枠禁止、案件匿名化、全ログ保全、最終承認者固定、失敗事例レビュー会を運用

この表は本報告の推奨設計です。根拠となる考え方は、Google Cloudのエージェント設計・評価指針と、GDPval/特許実務の高リスク性にあります。 32

PoCタイムライン



上記は例示的な標準工程です。実際には、社内データ接続の難易度とレビュー体制で前後します。特に匿名化と権限制御に時間を取る方が、後戻りが少なくなります。 33

運用フローとしては、まず**Gemini 3.5 Flashをオーケストレータ**に置き、検索戦略設計・反証・統合の中核を担わせます。その下で、低コストな再スコアリングや大量候補の粗い前処理には**Gemini 3.1 Flash-Lite**や**Batch/Flex**を使うのが合理的です。特許文書理解については、Googleの公式Cookbookに**Patents Document Understanding with Gemini**が存在し、分類・エンティティ抽出・オブジェクト検出がサンプル化されています。したがって、Google製基盤だけでも、**文書理解・RAG・マルチエージェント・評価**まで一応の要素は揃っています。 34

コスト見積もりは、**トークン＋検索クエリ＋キャッシュ＋評価**の四要素で考えるのが実務的です。3.5 Flash標準の料金は**入力$1.50/100万、出力$9.00/100万**で、Grounding with Google Searchは**5000件/月超で$14/1000クエリ**です。たとえば、1案件で合計**入力25万トークン、出力2.5万トークン、Google Search 15クエリ**を使うと、概算は**入力$0.375 + 出力$0.225 + 検索$0.21 = 約$0.81/案件**です。Batch/Flexに落とせる工程が半分以上あれば、同等推論部分のコストはさらに圧縮できます。もちろんこれは**再試行、社内RAG基盤費、外部特許DBライセンス、人手レビュー費用を含まない概算**です。 35

事例と参考文献

公開事例として信用度が高いのは、まず**特許庁・国際機関・大規模公的機関**の動向です。JPOはAI活用アクション・プランで、**先行技術調査①（概念検索・ランキング）**を導入フェーズへ、**先行技術調査②（検索手法高度化）**も導入フェーズへ進めています。WIPOのAI施策インデックスは、JPOについて**検索語・分類候補提案、画像類似検索、文献ランキング**を説明しています。EPOは2026年に**ANSERA-based SEARCH**を40超庁・2500人超の審査官に展開し、巨大文献群の高速・構造化検索を前面に出しました。これらは、「先行技術検索の高度化は、単なる仮説ではなく制度運用側の重点領域」であることを示します。 36

産業側の公開事例は、現時点では**ベンダー公表ベース**が多く、慎重に扱う必要があります。日本ではTokkyo.aiが2025年末に**特許特化ディープリサーチ**を発表し、検索式や分析過程まで可視化する「ディープエージェント方式」を訴求しています。またAconnectも、2026年に**除外条件の言語化支援**という、実務上き

わめて重要な検索支援機能を公表しています。これらは独立検証済み成果ではありませんが、**市場ニーズが“単なる要約AI”から“検索設計・除外設計・根拠可視化”へ移っている**ことを示す参考にはなります。 37

学術面では、近年の潮流は明確です。OpenAIの**GDPval**は現実の知識労働タスク評価を押し上げ、Artificial Analysisの**GDPval-AA**は共通ハーネスでのエージェント比較を提供しました。特許領域では、**A Survey on Patent Analysis: From NLP to Multimodal AI**が、検索・分類・要約・マルチモーダル処理を俯瞰しています。さらに、**Towards Automated Patent Workflows: AI-Orchestrated Multi-Agent Conversational Framework for Patent Tasks**は多エージェントによる特許ワークフローの設計を、**Rethinking patent retrieval with language models**は埋め込みモデルの実務的統合を、**Is It Novel and Why?**はクレーム要素ごとの新規性評価を論じています。研究はすでに、**“単一LLMの会話”から“検索・比較・証拠抽出を伴う複合ワークフロー”へ主戦場を移している**と言えます。 38

参考文献の要点整理

区分	優先して参照すべき資料	主な用途
Google公式	Gemini 3.5 Flash blog / Model Card / Pricing / Agents / Deep Research / ADK / Evaluation docs	モデル性能、ツール、価格、実装方式、評価方法
ベンチマーク原典	OpenAI GDPval論文、OpenAI解説ページ	GDPvalの定義、人間評価、限界
ベンチマーク実装	Artificial Analysis GDPval-AA methodology / leaderboard	Eloの意味、judge、ハーネス、比較対象
日本のIP実務	JPO AI action plan、INPIT J-PlatPat資料、JPAA記事	検索業務の要件、分類・検索式・官庁の導入動向
国際実務	WIPO AI initiatives、EPO search quality / novelty page	前技術検索の現実的要件、NPLの重要性、法的限界
リスク管理	AI事業者ガイドライン、NIST AI RMF	説明可能性、監査、ガバナンス、機密管理

出典。 39

オープンクエスチョンと限界

本件でなお不確実なのは、**Gemini 3.5 Flashの内部アーキテクチャ差分、Googleの公開チャネル間の仕様差の最終整合、特許検索ポータルごとのAPI可用性、社内案件での機密区分に応じた最適なデータ保持設定**です。また、GDPval-AAは有力な指標ですが、**特許調査専用ベンチマークではない**ため、導入判断には必ず**自社案件ベースのPoC**が必要です。したがって、実務上の最適解は「GDPval-AA 1656だから採用」ではなく、**証拠保全と人間レビューを組み込んだ限定導入で、再発見率・順位・工数・レビュー容易性を実測し、Go/No-Goを決める**ことです。 40

---

1 7 9 12 39 https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/
https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/

2 21 23 36 https://www.jpo.go.jp/system/laws/sesaku/ai_action_plan/ai_action_plan-fy2025.html
https://www.jpo.go.jp/system/laws/sesaku/ai_action_plan/ai_action_plan-fy2025.html

3 15 18 38 https://arxiv.org/html/2510.04374v1
https://arxiv.org/html/2510.04374v1

4 30 https://www.epo.org/en/learning/learning-resources-profile/business-and-ip-managers/inventors-handbook/novelty-and-prior-art
https://www.epo.org/en/learning/learning-resources-profile/business-and-ip-managers/inventors-handbook/novelty-and-prior-art

5 17 19 https://deepmind.google/models/model-cards/gemini-3-5-flash/
https://deepmind.google/models/model-cards/gemini-3-5-flash/

6 https://ai.google.dev/gemini-api/docs/interactions/whats-new-gemini-3.5
https://ai.google.dev/gemini-api/docs/interactions/whats-new-gemini-3.5

8 14 https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf
https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf

10 11 https://deepmind.google/models/gemini/flash/
https://deepmind.google/models/gemini/flash/

13 https://ai.google.dev/gemini-api/docs/computer-use
https://ai.google.dev/gemini-api/docs/computer-use

16 27 40 https://artificialanalysis.ai/methodology/intelligence-benchmarking
https://artificialanalysis.ai/methodology/intelligence-benchmarking

20 https://www.inpit.go.jp/content/100583908.pdf
https://www.inpit.go.jp/content/100583908.pdf

22 31 32 33 https://docs.cloud.google.com/architecture/choose-design-pattern-agentic-ai-system
https://docs.cloud.google.com/architecture/choose-design-pattern-agentic-ai-system

24 https://adk.dev/agents/workflow-agents/parallel-agents/
https://adk.dev/agents/workflow-agents/parallel-agents/

25 https://ai.google.dev/gemini-api/docs/file-search
https://ai.google.dev/gemini-api/docs/file-search

26 https://deepmind.google/models/model-cards/gemini-3-pro
https://deepmind.google/models/model-cards/gemini-3-pro

28 https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

29 https://ai.google.dev/gemini-api/terms
https://ai.google.dev/gemini-api/terms

34 35 https://ai.google.dev/gemini-api/docs/pricing
https://ai.google.dev/gemini-api/docs/pricing

37 https://prtimes.jp/main/html/rd/p/000000388.000042056.html
https://prtimes.jp/main/html/rd/p/000000388.000042056.html