Gemini 4 Neon深掘り調査レポート：技術仕様・経済性検証と知財業務への戦略的活用考察

Gemini 4 Carbon

GoogleおよびGoogle DeepMindは2026年11月18日（米国時間）、第4世代フロンティアモデル群「Gemini 4」シリーズの新モデルとして「Gemini 4 Neon（モデルID：gemini-4-neon）」を発表し、同日より提供を開始した[1]。Google DeepMindシニアバイスプレジデント兼GoogleチーフAIアーキテクトのKoray Kavukcuoglu氏による公式発表およびモデルカードによれば、Neonは最上位モデル「Gemini 4 Carbon」から直接派生・調整されたモデルであり、超長期・高難度プロジェクト向けの「Carbon」と高速・日常処理向けの「Gemini 4 Hydrogen」の中間（Googleがパレートフロンティアと位置づける領域）を担う[3]。入出力統合で最大1,048,576トークン（約100万トークン）のコンテキストウィンドウと最大1,048,576トークン（入力分差し引き）の継続出力（ロングデコード）を備え、標準API単価は入力100万トークンあたり3.00ドル・出力15.00ドル（暗黙的キャッシュ読込は95%割引の0.15ドル）に設定されている[1]。また、Geminiアプリやエージェント型開発環境「Google Antigravity」において最上位のGoogle AI Ultraプランに限定されていたCarbonとは異なり、Neonは中位の「Google AI Pro」プランから利用可能となったことで、実務展開の敷居が大きく引き下げられた[2]。

知的財産（知財）業務の観点から見ると、Gemini 4 Neonは従来のLLMが抱えていた「長文脈における参照構造の分断」「出力上限によるクレームチャートの途絶」「厳格な要件充足判断における見落とし」という3つの実務上のボトルネックを同時に緩和・解消し得るモデルである。具体的には、現役弁護士が作成・査読した多段階の米国法調査ベンチマーク「Vals AI Legal Research Bench」（2026年11月18日時点リーダーボード）において世界1位（全必須項目完全充足率56.25%）を獲得し、Google公表のベンチマーク比較では長文コンテキスト処理「GraphWalks」で95.2%と競合他社モデル（Claude Opus 5.5の66.8%、GPT-6.2 Solの80.1%）を大きく引き離している[8]。さらに、PDFや画像・動画に加えてGoogle Cloud環境ではMicrosoft Office文書（Word、Excel、PowerPoint）の直接入力に対応し、作業途中の指示修正への追従性（Steerability）で世界2位、Google Antigravity内部テストで同社過去最低の情報捏造率（ハルシネーション率）を記録した[3]。これにより、審査経過書類（包袋）全体の一括解析に基づく拒絶理由通知（OA）対応、先行技術・FTO（クリアランス）調査における網羅的な構成要件対比表の生成、およびNeonを指揮役（Orchestrator）・Hydrogenを実行役（Subagent）とする低コストな多段階エージェント運用が現実的な選択肢となる[3]。

もっとも、2026年11月19日時点の導入判断にあたっては4つの重要な実務上の留意点が存在する。第一に、最上位のCarbonが2026年11月30日まで半額の発売記念価格（入力2.00ドル／出力10.00ドル）で提供されているため、Neonの対Carbonでのコスト優位性（標準単価比25%減、標準思考レベルでのGoogle試算タスク単価32%減）が実質的に発効するのは、Carbonが定価（4.00ドル／20.00ドル）に戻る2026年12月1日以降である点だ[5]。第二に、第三者評価機関Artificial Analysisの検証において、図表を含むPDF理解（GDP.pdf）では思考レベルを既定のhigh（33.4%）から最高のxhigh（28.0%）に引き上げると逆に精度が低下する現象が確認されており、特許図面解析では思考レベルの過剰設定を避ける必要がある[12]。第三に、学習データの知識カットオフは2026年3月であ

り、公開モデルカードには日本語固有の法務・特許ベンチマークが含まれていないため、直近の法改正・審査基準運用や国内裁判例の補完と専門家による検証が不可欠である[6]。第四に、日本弁理士会の「弁理士業務AI利活用ガイドライン」が要請する守秘義務（弁理士法第30条）および新規性喪失防止を満たすには、Google Cloud（Gemini Enterprise Agent Platform）上でInteractions APIの非保存設定（store = false）や永続キャッシュの制御、エンタープライズ向けWebグラウンディングの選択など、厳密な「データ保持ゼロ（Zero Data Retention：ZDR）」構成を実装する必要がある[15]。

Gemini 4シリーズにおける系譜とモデルポジショニング

Google DeepMindは2026年秋からのわずか7週間のうちに、元素周期表の名称を冠した4つのGemini 4世代モデルを相次いで投入し、モデルラインナップの全面的な再編を進めた[14]。先行投入された「Gemini 4 Argon」（2026年10月13日一般提供）を皮切りに、長期的推論とサイバー防御能力を高めた最上位モデル「Gemini 4 Carbon」（11月10日一般提供）、従来のGemini 3.8 Flashを置き換え2.5倍の高速化と5分の1以下の捏造率低減を謳う「Gemini 4 Hydrogen」（11月12日発表）、そして今回の「Gemini 4 Neon」（11月18日発表）である[5]。

公式モデルカード（PDF）の記述によれば、Gemini 4 Neonは最上位のGemini 4 Carbonから直接派生（adapted）して開発されたモデルであり、Carbonのアーキテクチャ、学習データセット、データ処理パイプライン、および利用ポリシーを継承している[5]（なお、Google Cloud Model Gardenの一部の概要文では「Hydrogenをベースに深い推論を付加した」旨の表現も見られるが、DeepMind公式モデルカード上の技術的系譜はCarbon派生と明記されている[6]）。一方で、フロンティアモデルの重大リスクを評価する「Frontier Safety Framework」においては、Argonからの一般能力の伸びが詳細な再評価を要する「重大な能力変化（Material Capability Change：MCC）」の閾値未満と判定されたため、Argonの安全性評価ベースラインが引き継がれた[5]。その結果、生物分野（CBRN - Biological）およびサイバーセキュリティ分野における重大能力レベル（CCL）、ならびに制御喪失リスクのうちステルス性・状況認識の閾値（TCL）については「到達している可能性を否定できない」との慎重な判定が付されており、信頼できる防御側組織向けの特化派生版「Gemini 4 Neon for cyber」が「Fairwind Program」を通じて数日中に別途提供される体制が採られている[3]。

このモデル再編に伴い、初期主力モデルであった「Gemini 4 Argon」は事実上の移行対象となった。Neon発表時点で、Google公式の比較表、開発者向けドキュメントのモデル比較、およびGeminiアプリのProプラン選択メニューからArgonが姿を消しており、Google Cloudの開発者ガイドでも一般的なコーディングや企業向け推論ワークフローについて、Argon（およびGemini 3.1 Pro、Gemini 3.8 Flash）からNeonへの移行が公式に案内されている[1]。しかも2026年12月1日以降はArgonのAPI標準単価が4.00ドル／20.00ドルへと倍増し、ベンチマーク性能で上回るNeon（3.00ドル／15.00ドル）よりも約33%割高になるため、通常のAPI運用においてArgonを継続選択する経済的メリットは乏しい[2]。もっとも、例外となり得るのがエージェント開発環境「Google Antigravity」上の利用枠（クォータ）管理であり、技術解説メディアの報告によれば、NeonおよびOrchestrator (Neon)が5時間周期のリセット枠で管理されるのに対し、Argonは別枠の週次リセットとして扱われるため、Neonの上限到達時における予備枠（サブ）として温存する運用が推奨されている[2]。

比較項目	**Gemini 4 Hydrogen**	**Gemini 4 Neon**（本調査対象）	**Gemini 4 Carbon**	**Gemini 4 Argon**（移行元）

モデルID	gemini-4-hydrogen[21]	gemini-4-neon[4]	gemini-4-carbon[21]	gemini-4-argon[21]
一般提供開始日	2026年11月12日[21]	2026年11月18日[2]	2026年11月10日[21]	2026年10月13日[21]
中核的な役割・系譜	日常作業・高速実行・サブエージェント向けの軽量主力（3.8 Flash後継）[18]	Carbon派生の高速・高効率フロンティア推論および指揮役（Orchestrator）[3]	超長期・最高難度の研究および複雑課題向けの最上位フラッグシップ[5]	初期フロンティア推論モデル（現在はNeonへの移行を公式案内）[1]
統合コンテキスト上限	1,048,576トークン[21]	1,048,576トークン（Antigravityでは最大512Kとの報告）[2]	1,048,576トークン[21]	1,048,576トークン[21]
最大出力トークン	同期64K／継続最大1M[23]	同期65,536（64K）／継続最大1,048,576[4]	同期64K／継続最大1M[23]	同期64K／継続最大1M[23]
思考レベル設定	low, medium, high（既定）, xhigh[21]	LOW, MEDIUM, HIGH（既定）, XHIGH[4]	low, medium（既定）, high[21]	low, medium（既定）, high[21]
知識カットオフ	2026年3月[25]	2026年3月[1]	2026年3月[26]	2026年3月（系列共通）[6]
標準API単価（入力／出力）	1.50ドル / 7.50ドル（期限なし・1M全域均一）[21]	**3.00**ドル / **15.00**ドル[3]	11/30まで 2.00ドル / 10.00ドル 12/1以降 **4.00**ドル / **20.00**ドル[11]	11/30まで 2.00ドル / 10.00ドル 12/1以降 4.00ドル / 20.00ドル[11]
Batch API単価（入力／出力）	0.75ドル / 3.75ドル[13]	**1.50**ドル / **7.50**ドル[11]	11/30まで 1.00ドル / 5.00ドル 12/1以降 2.00ドル / 10.00ドル[11]	11/30まで 1.00ドル / 5.00ドル 12/1以降 2.00ドル / 10.00ドル[11]

暗黙的キャッシュ読込単価	0.075ドル / 1M[11]	**0.15**ドル / **1M**（95%割引）[3]	11/30まで 0.10ドル / 1M 12/1以降 0.20ドル / 1M[11]	11/30まで 0.10ドル / 1M 12/1以降 0.20ドル / 1M[11]
API無料枠と学習利用	無料枠あり（無料枠は学習利用**Yes**／有料枠はNo）[13]	無料枠なし（有料枠は学習利用**No**）[13]	無料枠なし（有料枠は学習利用**No**）[13]	無料枠なし（招待制Contributor枠のみ学習利用Yes）[13]
Geminiアプリ／**Antigravity**	全ユーザー（AntigravityはPro/Ultra）[21]	**Google AI Pro** / Ultra / Enterprise[2]	**Google AI Ultra** / Enterprise限定[2]	Pro選択欄から消滅（AntigravityはUltra）[5]

アーキテクチャ・API仕様の深掘りと開発上の破壊的変更

Gemini 4 Neonの技術仕様において実務システム設計に直結する特徴は、「入出力統合100万トークン空間におけるロングデコード機構」「Office文書のネイティブ直接処理」「サンプリングおよび思考パラメータの厳格化」、そして「オーケストレーター主導のマルチエージェント設計」の4点に集約される。

第一に、コンテキストと出力の統合管理機構である。Neonのコンテキスト上限である1,048,576トークンは、入力プロンプト・添付ファイルだけでなく、モデル内部の思考トークン（thinking tokens）、ツール・関数呼び出し（function calls）、および最終回答テキストのすべてを合算した統一上限（Unified Token Limit）として機能する[1]。1回のフォアグラウンド同期セグメントあたりの最大出力は65,536トークン（モデルカード上の概数表記では64K）に制限されているが、GenerateContent API（v1beta1エンドポイント必須）およびInteractions APIにおいては、セグメント完了時にstatus="requires_continuation"とともに返却されるcontinuation_tokenを引き継ぐことで、最大1,048,576トークンから入力分を差し引いた全領域まで途切れなく長文生成を継続できる「ロングデコード（Long Decodes）」が実装された[4]。Google Gen AI SDKの対応バージョン（Python google-genai >= 2.34.0、JS/TS @google/genai >= 2.32.0、Go >= 1.77.0、Java >= 1.80.0）ではこの継続処理が自動化されており、またbackground=trueを指定したサーバー管理型の非同期バックグラウンド生成にも対応している[4]。

第二に、企業実務文書のネイティブ入力拡張である。従来のマルチモーダル入力（テキスト、画像、音声、動画、PDF、プレーンテキスト）に加え、Google Cloudのエンタープライズ基盤（2026年にVertex AIの生成AI基盤から再編・統合が進められた「Gemini Enterprise Agent Platform」）上のNeonでは、Microsoft Office形式（.docx, .dotx, .doc, .pptx, .ppt, .xlsx, .xls）およびOpenDocument形式（.odt, .odp, .ods）をインラインまたはCloud Storageから直接投入できる[4]。特筆すべきは、これらのOfficeファイルやPDF、画像を関数呼び出しの戻り値（マルチモーダルFunctionResponseパート内部）に直接格納してモデルへ返却できる点であり、エージェントが社内文書管理システムや特許データベースから取得したExcelリストやWord原稿を、前処理テキスト変換による構造崩れなしにそ

のまま推論ループへ組み込める[4]。
第三に、旧世代Geminiモデルからの移行時にシステム障害を引き起こしやすいAPIパラメータの破壊的変更である。Neonは推論の深さをLOW、MEDIUM、HIGH（デフォルト）、XHIGH（Agent Studio上の表記はExtra high）の4段階の文字列列挙型thinking_levelで制御する[4]。この際、Gemini 3.x世代以前で使われていた数値指定のthinking_budgetパラメータやthinking_level="MINIMAL"を送信すると、APIは**HTTP 400**バリデーションエラーを返して失敗する[4]。さらに、temperature、top_p、top_kといった従来のサンプリング制御パラメータはモデル側で自律管理される仕様へと変更され、カスタム値を渡しても無視される[4]。したがって、知財調査のように出力の再現性や定型性が求められる業務では、temperature=0に頼るのではなく、システム指示での明示的なルール定義と、OpenAPI準拠のresponseSchema（特に生成順序を制御するpropertyOrderingフィールド）を用いた構造化出力（Structured Output）によって出力形式を拘束する設計への転換が求められる[7]。なお、発表直後の2026年11月18日夜から19日にかけて、開発者フォーラムではAntigravity CLI（v1.5.11）環境で--model gemini-4-neon-xhighがコード400で拒絶される事象や一時的な思考トークン減少の報告、ならびにHacker Newsコミュニティでは長期ベンチマーク最適化の影響でモデルが自ら確認質問を発して自己回答して処理を進める「自己問答（Self-Answering）」挙動が指摘されており、本番パイプラインへの適用前にはSDKバージョンとハーネス挙動の事前検証が不可欠である[32]。
第四に、キャッシュ機構と長時間エージェント制御の高度化である。Neonは明示的キャッシュ（Explicit Context Caching）をサポートせず、4,096トークン以上の共通プロンプトプレフィックスに対して自動適用される「暗黙的キャッシュ（Implicit Caching）」のみを採用している[4]。またGoogleの発表によれば、長時間にわたる自律エージェント実行において、同一ファイルの重複読み込みを抑止し、作業進捗を逐次共有しながら蓄積した作業履歴を圧縮（Context Compaction）する信頼性が高められた[1]。さらにGoogle AntigravityのOrchestrator (Neon)モードやInteractions APIのprevious_interaction_id引き継ぎを用いることで、Neonが全体方針の立案・アーキテクチャ設計・難所の検証を担い、複数のGemini 4 Hydrogenサブエージェントに並列でファイル探索や定型実行を委譲し、サブエージェントがエラーや摩擦に直面した際のみ制御をNeonへ戻して計画を修復する「Planner / Executor分離型オーケストレーション」が標準機能として提供されている[1]。

ベンチマーク徹底検証：法務・長文推論の優位性と視覚推論の課題

Gemini 4 Neonの実力を客観的に評価するため、Google公式発表の数値と、第三者評価機関であるArtificial AnalysisおよびVals AIの独立検証データを統合して検証する[3]。全体像として浮かび上がるのは、Neonが単なる廉価版の中位モデルにとどまらず、実務エージェント遂行能力・法務調査・長文脈理解において最上位のCarbonや他社フラッグシップモデルを上回る指標を示す一方、特定条件下での思考レベル設定や実行環境（ハーネス）に明確な依存性を持つという点である。

Gemini 4 Neonは法務調査と長文脈推論で世界首位級を達成しつつ処理速度を倍増

AA総合知能指数 (xhigh)	Legal Research Bench	出力生成速度	GraphWalks 長文脈処理
61 pt	**56.25%**	**約110 tok/秒**	**95.2%**
世界総合3位 (Carbon: 60 pt)	世界総合1位 (完全充足率)	Carbon (約57 tok/秒) の約2倍	100万トークン統合文脈

Vals AI Legal Research Bench（全要件完全充足率・コスト・所要時間）

米国連邦法・州法208問の実務調査タスクにおける All-Pass Accuracy（0%〜70%スケール）

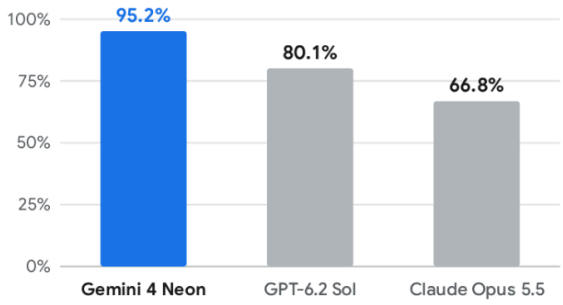
■ Neon ■ Gemini 4 系列 ■ 他社モデル

順位・モデル	完全充足率 (All-Pass Accuracy)	費用/回	所要時間
#1 Gemini 4 Neon	56.25%	$7.37	35分40秒
#2 Gemini 4 Carbon	55.77%	$8.74	51分53秒
#6 Gemini 4 Argon	54.81%	$6.45	31分21秒
#8 Gemini 4 Hydrogen	53.85%	$6.27	32分44秒
#11 Claude Opus 5.5	50.48%	$23.61	1時間52分
#17 Claude Sonnet 5.5	48.08%	$13.66	1時間33分
#32 GPT-6.2 Astra	42.79%	$12.84	1時間19分
#54 GPT-6.2 Sol	37.02%	$3.77	1時間09分

0% 10% 20% 30% 40% 50% 60% 70%

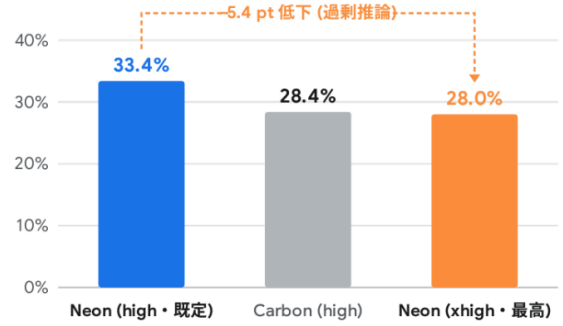
長文コンテキスト処理（GraphWalks）

100万トークン文脈における関係性推論精度（0%〜100%）



視覚・図表PDF理解（GDP.pdf：思考レベル別の逆転現象）

図表・グラフを含む資料理解スコア（0%〜40%）



Vals AIの法務調査ベンチマーク（米国法208問・全必須ルーブリック完全充足率）では、Neonが最上位のCarbonやClaude Opus 5.5を上回り首位を獲得した。長文脈推論（GraphWalks）でも圧倒的な優位を示す一方、図表を含む資料理解（GDP.pdf）では思考レベルを最高の*xhigh*に上げると既定の*high*より精度が5.4ポイント低下する逆転現象が確認されており、用途に応じた思考制御が不可欠である。

Data sources: Google, Vals AI, Artificial Analysis, GIGAZINE, XenoSpectrum

総合知能および処理速度の面では、Artificial Analysisの独立測定（AA Intelligence Index v5.0.1）

において、Neonは最高思考レベルxhigh設定時に61ポイントを記録し、GPT-6.2 Astra（max）およびClaude Fable 5.2（max）に次ぐ世界総合3位にランクインした[2]。注目すべきは、ベースモデルである最上位のGemini 4 Carbon（high設定：60ポイント）を1ポイント上回り、Google公表の比較表におけるClaude Opus 5.5（59ポイント）やGPT-6.2 Sol（57ポイント）を抑えている点である（既定のhigh設定同士でもNeonはCarbonと同等のスコアに到達している）[8]。しかも出力生成速度は毎秒約110トークンとCarbon（毎秒約57トークン）のほぼ2倍に達し、タスク1件あたりの出力トークン数がCarbon比で約15%増加しているにもかかわらず、平均タスク完了時間は約36分から約21分へと約40%短縮されている[12]。他方で、開発者コミュニティ（Hacker News）で指摘されている通り、研究規模の深い科学的推論を測る「CritPt」評価においては依然としてCarbonがNeonを上回っており、超長期の未知の科学探索ではCarbon、高頻度で回す高度な分析・実務エンジニアリングではNeonという役割分担が裏付けられている[33]。

知財・法務関係者が特に注目すべきは、Vals AIが実施する「Legal Research Bench」（2026年11月18日時点）における総合1位の獲得である[3]。同ベンチマークは単なる条文知識の暗記テストではなく、現役弁護士が実務上の論点から作成・査読した非公開テストセット208問（全体データセット413問のうち評価専用セット）に対し、エージェントが判例検索・Web検索・文書検索ツールを駆使して多段階の法的推論を行う評価である[9]。とりわけその採点基準は厳格であり、1問あたり平均9.35個（1〜31個、最頻値10個）設定された必須評価項目（ルーブリック）をLLM審査員が採点し、すべての必須項目を満たした場合のみ正解（100%）、1項目でも欠落すれば0%とする「All-Pass Accuracy（全要件完全充足率）」を主指標としている[9]。この基準下でNeonは56.25%を達成して全モデル中1位に立ち、2位のCarbon（55.77%）をわずかに上回るとともに、11位のClaude Opus 5.5（50.48%）に対して5.77ポイントの差をつけながら、1テストあたりの費用を約3分の1以下（7.37ドル対23.61ドル）、所要時間を約3分の1（35分40秒対1時間52分）に圧縮した[9]。さらに複雑な制約条件の遵守やスプレッドシートモデリングを測るSurge AIベンチマーク群の大半や税務エージェント評価「Tax Agent Bench」、業務自動化「Zapier AutomationBench」（56.3%）でも首位、実務総合指標「Vals Index」で2位（70.9%）、ビジネス資料作成評価「AA-Briefcase」の成果物品質（Presentation Quality）で首位（1,771 Elo）、4,200セッション超の実運用評価「Agent Arena Index」でも世界2位（1,336ポイント・途中修正対応力でも2位）を記録している[3]。ただし、Legal Research Benchの評価対象は米国の連邦法・州法（行政・規制法、商事法、民事訴訟等8分野）であり、日本の特許法や審査基準を直接測定したものではないほか、公開モデルカードにも日本語固有の言語ベンチマーク単体スコアは掲載されていない点には留意が必要である[6]。それでも「多数の必須構成要件を1つも漏らさず事実関係に当てはめて論証する」という情報処理タスクは、特許実務における文言侵害判断（オールエレメント・ルール）や進歩性判断と共通する構造を持っており、知財業務への高い応用可能性を示唆している。

一方で、ベンチマークの詳細分析からは実務導入における2つの重要な技術的課題が浮かび上がる。第一は「視覚資料理解における過剰推論（Overthinking）による精度低下」である。Artificial Analysisによる図表・グラフ混在PDF理解ベンチマーク「GDP.pdf」において、Neonは既定のhigh設定では33.4%とCarbon（28.4%）を5ポイント上回る最高水準のスコアを記録したにもかかわらず、思考レベルを最高のxhighに引き上げるとスコアが28.0%へと低下した[12]。また高密度チャートの読み取り「Chartography」でもNeonは73.5%を記録したものの、Google公表の比較表ではGPT-6.2 Solが優勢とされている[3]。すなわち、視覚情報や図面を含むPDF解析では、推論トークンを過剰に増やせば必ずしも成績が上がるわけではなく、high設定を基本とする運用が合理的である[12]。第二は「実

行ハーネスへの強い依存性」である。ソフトウェア修正能力を測る「**SWE-sweep**」において、軽量な**mini-SWE**環境での**Neon**の成功率は**12.8%**にとどまるが、サブエージェント連携やコンテキスト管理を備えた**Google Antigravity**環境で評価すると**3**倍の39.0%（Google比較表でのClaude Opus 5.5は10.1%、GPT-6.2 Solは8.3%）へと跳ね上がる[3]。これはモデル単体への単発プロンプトにとどまらず、周辺ツール群やオーケストレーション基盤（ハーネス）と統合することで実務性能が大きく引き出されることを示している[12]。

安全性・信頼性の面では、Google社内のAntigravity内部テストにおいてGeminiシリーズ過去最低の情報捏造率を記録したほか、エージェント誠実性評価（Agent Honesty Evaluation）で99.2%合格、意図からの逸脱行動率（Agentic Safety Suite）で0.09%、テストの抜け穴を突く不正行為率（Terminal-Bench 4.0 Cheating Eval）で0.0%、近道行動率（Shortcut rate）でもGemini 4シリーズ最低の2.45%と報告されている[3]。もっとも、第三者メディアの検証記事でも指摘されている通り、これらの誠実性・捏造率の数値はGoogleの内部テストに基づくものであり、公式モデルカード上でも一般的なLLMの限界としてハルシネーションの発生可能性が明記されているため、外部独立評価とは区別して捉える必要がある[6]。また、Hydrogenとの社内比較では正当なプロンプトに対する過剰拒否率（Unjustified-refusals）が3.00ポイント改善（低下）しており、医療・バイオテクノロジーや化学分野の文献を扱う際の不要な回答拒否が軽減されている[5]。加えて、外部文書に仕込まれた悪意ある命令への耐性を測る「Gray Swan間接プロンプトインジェクション（IPI）ベンチマーク」では、最高思考レベルとリアルタイムセーフガードの併用時にベンチマーク上の全攻撃を防御（攻撃成功率0.0%）したとGoogleは報告している[3]。

トークン経済性と2026年12月1日価格逆転の構造分析

Gemini 4 NeonのAPI価格体系を正確に評価するには、名目のトークン単価だけでなく、思考トークンを含む実効タスク費用、Batch推論・暗黙的キャッシュの割引構造、そして**2026**年**11**月**30**日を境に発生するシリーズ内の価格逆転現象を時系列で把握する必要がある[11]。

2026年12月1日のCarbon割引終了によりNeonのコスト優位性が本格発効

PANEL 1・API 100万トークン単価比較 (PAID TIER / GLOBAL)

12月1日の定価移行による単価逆転構造

Standard プラン | Batch プラン (50% OFF)

モデル・適用期間	入力単価 ($/1M) 上限 $5.00	出力単価・思考含む ($/1M) 上限 $25.00
Gemini 4 Hydrogen 常時均一	$1.50 (Cache読込: $0.075)	$7.50
Carbon / Argon (11/30迄 割引) 期間限定半額	$2.00 (Cache読込: $0.10)	$10.00
Gemini 4 Neon 新標準主力	**$3.00** (Cache読込: $0.15)	**$15.00** 12/1以降 Carbon比 -25%
Carbon / Argon (12/1以降 定価) 通常価格	$4.00 (Cache読込: $0.20)	$20.00

$0 $2.50 $5.00 | $0 $12.50 $25.00

価格逆転のポイント：11月30日まではCarbon/Argonの発売記念半額割引により一時的に割安だが、12月1日以降は**Neonが入力・出力ともに25%安価**となる。

PANEL 2・実測ベンチマーク 1タスク平均費用 ($ USD / TASK)

Artificial Analysis 第三者検証による実効タスクコスト比較

共通スケール：$0 ～ $18.00 / タスク

AA総合評価 1タスク平均費用（Gemini 4 シリーズ内比較）

Intelligence Index: Neon (xhigh) 61点 vs Carbon (high) 60点

Carbon (high・11/30迄割引換算)	$3.88 11/30迄 限定半額
Gemini 4 Neon (xhigh)	**$6.44** Carbon定価比 -17%
Carbon (high・12/1以降定価)	$7.76 12/1以降 標準単価

Coding Agent Index 1タスク平均費用（他社最上位比較）

同等の成功率（スコア約58～59）を約半額で達成

Gemini 4 Neon (xhigh / Antigravity) 指数 58.6 (#1)	**$6.86** 最安・最高水準
Claude Opus 5.5 (max / Claude Code) 指数 約58	$12.61 Neonで -46% 削減
Claude Fable 5.2 (xhigh / Claude Code) 指数 約58	$15.47 Neonで -56% 削減

$0 $6.00 $12.00 $18.00

2026年11月30日まではCarbonの発売記念半額割引が適用されるため一時的にCarbonの方が安価となるが、12月1日の定価移行後はNeonが単価で25%安く、標準思考レベルでのタスク費用で32%安価となる。またエージェント型コーディングの1タスク平均費用では、同等のスコアを持つClaude Opus 5.5の約半額（46%減）で稼働する。

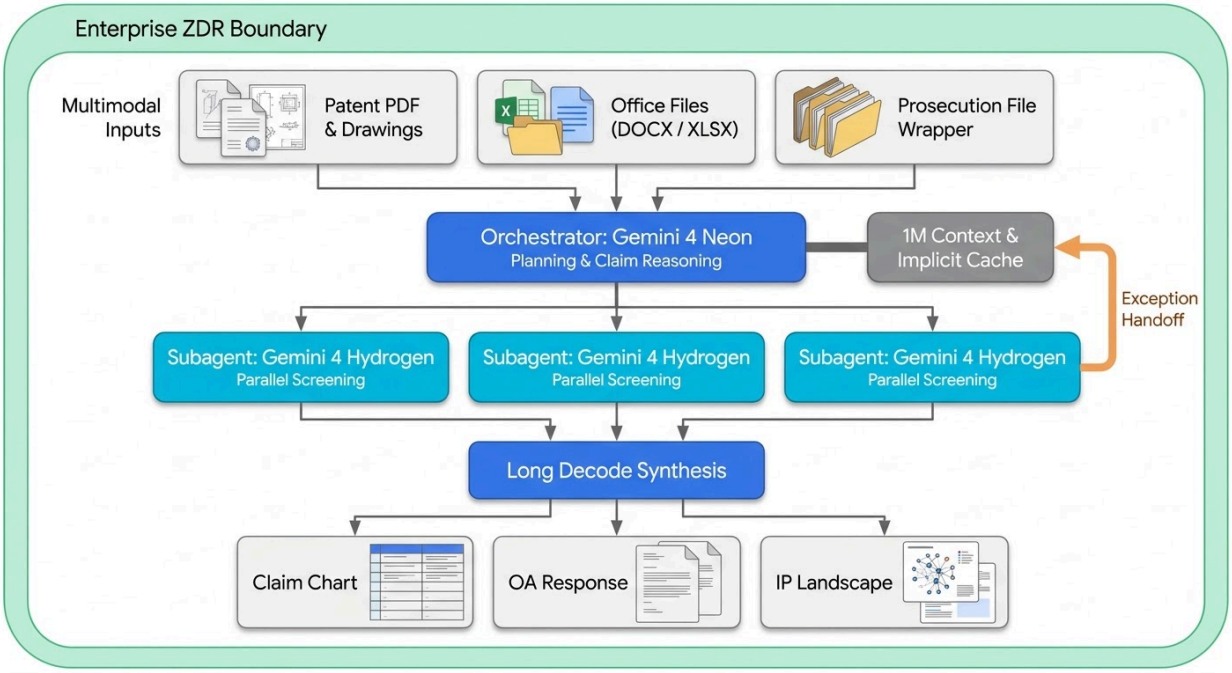
Data sources: Google AI for Developers (Gemini API Pricing), Google Cloud (Model Garden Pricing), Artificial Analysis, XenoSpectrum

Gemini APIおよびGoogle Cloud（Model Garden）の公式料金表によれば、Gemini 4 NeonのGlobalリージョン標準（Standard）単価は、入力100万トークンあたり3.00ドル、思考トークンおよび内部ツール宣言を含む出力100万トークンあたり15.00ドル、暗黙的キャッシュ読込が95%割引の0.15ドルである[11]。非同期一括処理を行うBatch推論（Gemini API料金表およびGoogle Cloud Batch Inference）では全単価が50%割引（入力1.50ドル／出力7.50ドル／キャッシュ0.075ドル）となり、Hydrogenの標準プランと同一単価でNeonの推論を利用できる（なお、AI Studioのレート制限ドキュメントでは発表直後時点でNeonのBatch上限値の記載が未反映となっているため、大規模Batch運用はGoogle Cloud側のBatch Inference基盤を軸に検証することが望ましい）[11]。一方、Google Cloudにおいて特定のリージョン固定（usやeuなどのNon-global）を選択する場合は10%のリージョン加算（標準入力3.30ドル／出力16.50ドル／キャッシュ読込0.165ドル）が適用されるほか、優先スループットを提供するPriority推論はGlobalで入力5.40ドル／出力29.70ドル、キャッシュ書き込み（Cache write）は標準入力の1.25倍（Global標準で3.75ドル／Non-globalで4.125ドル、より長期の保持は2027年1月12日以降追加料金対象）として設定されている[4]。またGoogle検索やGoogleマップによるグラウンディングは、Gemini 3および4シリーズ合算で月間5,000クエリまで無料、超過分は1,000クエリあたり14.00ドルで課金される[11]。

ここで実務上の予算策定において押さえるべきなのが、**2026**年**11**月**18**日から**11**月**30**日までの期間に限って生じている過渡的な価格逆転である。GoogleはCarbonおよびArgonに対し、11月30日まで50%相当の発売記念特別価格（入力2.00ドル／出力10.00ドル）を適用している[11]。そのため、11月中に限り同一トークン量を処理した場合の単価はCarbonの方が安く、Artificial Analysisの評価タスク平均費用でもNeon（xhigh：6.44ドル）に対し割引換算時のCarbonは約3.88ドル（Neonより約40%安価）、またHacker Newsで報告された43問のデータ分析テストでもHydrogenが6ドル、Carbonが11ドル、Argonが13ドルに対しNeonが17ドルと最も高くなる結果が観測された[12]。しかし、**2026**年**12**月**1**日に**Carbon**と**Argon**が標準定価（入力**4.00**ドル／出力**20.00**ドル）へ移行すると、コスト構造は逆転する[11]。定価ベースではNeonはCarbonよりトークン単価で25%安く、Artificial Analysis測定のxhighタスク平均費用で約17%安価（6.44ドル対7.76ドル）、さらに既定のhigh思考レベル同士ではGoogle発表値でタスクあたりコストがCarbon比32%減（Artificial Analysisのhigh同士比較でも定価Carbonの半額未満）となる[5]。さらに他社最上位モデルとの比較では、Coding Agent Indexで同水準のスコア（約58〜59ポイント）を出すためにClaude Opus 5.5が平均12.61ドル、Claude Fable 5.2が平均15.47ドルを要するのに対し、Neonは6.86ドルと約半額（それぞれ約46%減、約56%減）で完遂しており、12月以降の継続運用における費用対効果は極めて高い[12]。

知財業務への活用考察：実務ユースケース別の適合性と設計指針

以上の技術特性・ベンチマーク結果・経済性を踏まえると、Gemini 4 Neonは知財部門および特許事務所におけるAI活用を「単発の要約・翻訳補助ツール」から「一気通貫で調査・対比・起案を遂行する自律型マルチエージェント基盤」へと発展させる中核モデルとなる。とりわけ費用対効果と処理速度を両立する設計パターンは、Google AntigravityやGemini Enterprise Agent PlatformのWorkflow Builder／Interactions APIを活用し、高度な法的判断と全体設計を**Neon**（**Orchestrator**）が担い、大量文献の一次スクリーニングや定型抽出を並列稼働する**Hydrogen**（**Subagent**）に委譲するハイブリッド構成である[7]。



1. 先行技術調査・FTO（クリアランス）調査とクレームチャート作成

FTO調査や無効資料調査における最大の難所は、対象製品仕様または本願発明と数百～数千件の候補公報を突き合わせ、独立項・従属項のすべての構成要件（構成要件A、B、C…）について文言充足性および均等論・設計変更の可能性を漏れなく検証する作業である。従来のLLM運用では、公報をRAG（検索拡張生成）で細切れのチャンクに分割するケースが多く、「請求項1の『前記制御部』が明細書段落[0048]のマイコン20を指し、図3のフローチャートステップS14で動作する」といった文書横断的な参照関係が分断され、誤判定の原因となっていた。

Neonは長文コンテキスト処理「GraphWalks」で95.2%という高い関係追跡スコアを示し、さらにLegal Research Benchの全必須項目充足（All-Pass）評価で世界1位を記録している[8]。実務フローとしては、まずNeon（HIGHまたはXHIGH）が対象製品の技術仕様書（.docx／.xlsx）を読み込んで構成要件の分説と判定ルーブリックを設計し、母集団となる1,000件規模の公報（要約＋請求の範囲）の一次スクリーニングを並列のHydrogen（Batch推論またはサブエージェント）へ委譲する[4]。その後、抽出された重要公報30～50件については、製品仕様書（約2万トークン）をプロンプト先頭に固定してNeonの暗黙的キャッシュ（95%割引）を効かせた状態で、各公報の全文PDF＋全図面を丸ごと投入する[3]。さらに対比結果を出力する際、ロングデコード機能（最大約100万トークン継続出力）を活用すれば、全請求項×全引用文献の網羅的な構成要件対比表（クレームチャート）を出力上限で途中省略させることなく一括生成させる運用が可能となる[4]。

2. 中間処理（拒絶理由通知・OA対応）および包袋禁反言の横断検証

特許庁からの拒絶理由通知（Office Action）への対応や訴訟における無効抗弁の検討では、本願明細書、拒絶理由通知書、審査官が引用した主引例・副引例3～5件（特許公報および非特許文献PDF）、さらには米国・欧州・中国等の対応外国出願における審査経過書類（包袋／File Wrapper）をすべて整合的に読み解く必要がある。これら一式の総トークン数は10万～30万トークン規模に達するが、Neonの1,048,576トークン統合コンテキストであれば、包袋全体を単一セッション内にそのまま展開できる[3]。

ここで効果を発揮するのが、Neonの2つの特性の組み合わせである。第一に、審査官の認定における「主引例に副引例を組み合わせる動機付け（進歩性論理付け）の飛躍」や「外国ファミリー出願での過去の意見書主張と矛盾する補正（包袋禁反言／Prosecution History Estoppelリスク）」を、複数文書を横断して検証できる点である[1]。第二に、Agent ArenaのSteerability（作業途中の指示修正への対応力）で世界2位を獲得している点である[3]。実務では弁理士や知財担当者が「請求項3の限定ではなく段落[0065]の数値範囲で補正案を組み直し、サポート要件と新規事項追加の観点で再検討せよ」といった方針転換を対話的に繰り返すが、Neonは暗黙的キャッシュによって長大な包袋コンテキストを0.15ドル／1Mトークンの低コストで保持したまま、修正指示に柔軟に追従して補正案・意見書ドラフトを再構成できる[3]。

3. 特許明細書ドラフティング・外国出願展開と自己問答挙動の制御

発明者から提出される実験データ（.xlsx）、研究発表スライド（.pptx）、試作メモ（.docx）をGoogle Cloud上のNeonへ直接マルチモーダル投入し、背景技術・課題・複数の実施形態・変形例・請求の範囲を含む長大な明細書ドラフトを生成する用途でも、65,536トークンの単一セグメント枠を超えられるロングデコードが大きな強みとなる[4]。また、日英中韓独などの多言語ファミリー展開（PCT国際出願の各国移行翻訳）においても、大規模なバイオ・半導体・通信規格明細書を分割せずに一括処理できるため、チャンク分割に起因する技術用語やクレームの先行詞対応の揺らぎを抑えやすい（ただし、公式モデルカードには多言語安全性スコアとして対Hydrogen比-0.40ポイントが記載されているのみで、特許翻訳精度の定量ベンチマークは公表されていないため、専門翻訳者・弁理士によるクレーム表現の精査は必須である）[6]。さらに、Hydrogen比で過剰拒否率が3.00ポイント低減されているため、創薬・遺伝子工学、機能性化学、サイバーセキュリティ分野の発明を扱う際に安全フィルタの過剰反応で生成が拒否される頻度も緩和される[5]。

一方で注意すべきは、開発者コミュニティで報告されている「モデルが自ら確認質問を作って自己回答して進めてしまう挙動（Self-Answering）」である[33]。発明発掘や明細書ドラフティングにおいてこの挙動が生じると、実験データに存在しない変形例や作用効果をAIが推測で補完してしまい、実施可能要件違反や発明者の意図しない限定を招くリスクがある。これを防ぐため、システム指示（System Instruction）において「開示資料に明記されていない構成・数値範囲・作用効果を推測で補完してはならず、不明点は必ず[発明者確認事項]として列挙して出力を停止せよ」と明示的にルール化することが重要である[7]。

4. 図面・実験グラフ・UI動画のマルチモーダル知財解析

知財業務ではテキストだけでなく、特許図面中の引き出し線・符号と明細書本文の整合性チェック（符号抜け・誤記監査）、化学・材料特許における実験グラフ（実施例と比較例のプロット図）の数値読み取り、さらには画像意匠（GUI意匠）やソフトウェア特許の侵害立証における製品操作動画の解析が頻繁に発生する[1]。Neonは高密度グラフ読取「Chartography」で73.5%を記録し、Google Cloud環境では動画理解機能（Agentic Video Understanding：GA提供、デフォルトはオフのため明示的有効化が必要）も備えている[3]。

ここで徹底すべき運用方針が、前述したベンチマーク検証に基づく思考レベルの使い分けである。図表やレイアウトを含むPDF資料の解析（GDP.pdf）では既定のHIGH（33.4%）が最高のXHIGH（28.0%）を上回っているため、特許図面の符号監査やグラフ読み取りタスクでは思考レベルをHIGHに設定し、文字ベースの複雑なクレーム論理検証や外部文献取り込み時の間接プロンプトインジェクション防御においてXHIGHを選択する設計が合理的である[3]。また、複雑な三次元機械構造の断面図や重なり合う部材の空間的把握についてはハルシネーションが生じる余地が残るため、AIの出

力は一次監査補助と位置づけ、最終的な図面確認は必ず人間が行う必要がある[6]。

5. IPランドスケープ構築およびソフトウェア知財・OSS監査

経営層・事業部門への提言を行うIPランドスケープ業務では、特許母集団データ（.xlsx）、競合他社の決算説明資料・統合報告書（PDF）、および最新の市場動向（Web Grounding for Enterprise）を単一コンテキストで統合分析し、構造化されたレポートやプレゼン構成を生成できる[3]。NeonはSurge AIのスプレッドシートモデリング・財務分析で首位、AA-Briefcaseのプレゼン成果物品質でも世界1位（1,771 Elo）を獲得しており、知財情報と財務・市場情報を掛け合わせた分析レポート作成との親和性が高い[3]。

さらにソフトウェア企業や電機・自動車メーカーの知財部門にとって有用なのが、Coding Agent Index世界1位（58.6ポイント）およびリポジトリ横断解析能力の知財応用である[1]。自社のソースコードリポジトリをGoogle Antigravityまたは専用エージェントで走査させ、「最新コミット群から特許出願候補となる新規アルゴリズム・制御ロジックを抽出する（ソフトウェア発明発掘）」、あるいは「M&A対象企業のコードベースを解析し、他社特許クレームとの関連性や脆弱性リスクを洗い出す」といった、コード解析と知財実務を連結させたワークフローを従来の上位競合モデルの約半額のタスクコストで構築できる[10]。

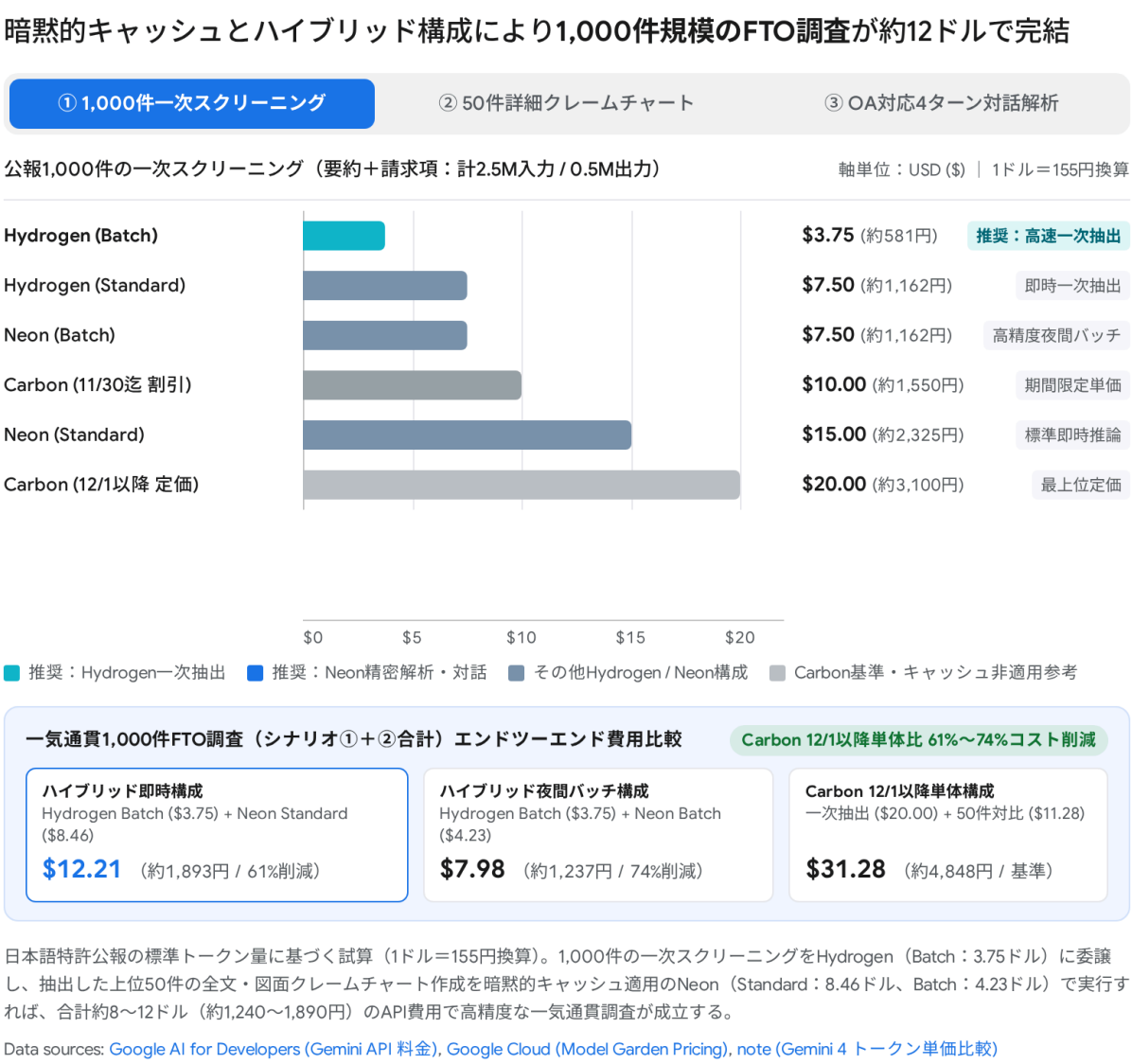
知財業務領域	推奨モデル構成・役割分担	推奨思考レベル	活用すべき主要**API**機能	実務上の留意点・リスク管理
大量母集団の一次スクリーニング（先行技術・SDIウォッチング）	**Hydrogen**（Batchまたは並列Subagent）＋**Neon**（検索式・評価基準設計）[7]	Neon: HIGH Hydrogen: MEDIUM / HIGH[4]	Batch推論（50%割引）、responseSchemaによる判定理由・ランクのJSON構造化出力[7]	Hydrogenの無料枠（**Free Tier**）は学習利用されるため、未公開製品情報を含む調査では必ず有料枠／Enterpriseを使用する[13]。
FTO詳細調査・無効資料対比・クレームチャート作成	**Neon** 単体、または超長期・最難関案件で **Carbon** 併用[6]	テキスト論理: XHIGH 図面参照主体: HIGH[12]	暗黙的キャッシュ（製品仕様固定）、ロングデコード（全請求項・全引例の一括出力）[4]	外部公報・Web取り込み時はXHIGH＋リアルタイムセーフガードで間接プロンプトインジェクションを防御する[3]。
拒絶理由通知（**OA**）対応・包袋	**Neon**（対話型マルチターン推論）[3]	HIGH ～ XHIGH[4]	100万トークン統合文脈、暗黙的キャッシュ、	2026年3月以降の最新審査基準・裁判例は知

一括解析			Interactions API（store=false）[7]	識カットオフ外のため、別途最新資料をコンテキスト注入する[6]。
特許明細書ドラフティング・多言語翻訳	**Neon**（起案・用語統一監査）[4]	HIGH[4]	Officeファイル（.docx, .pptx, .xlsx）直接入力、ロングデコード継続出力[4]	勝手な実施形態補完（Self-Answering挙動）を防ぐため、不明点を[発明者確認事項]として停止させるシステム指示を徹底する[7]。
図面符号監査・実験グラフ・**UI**意匠動画解析	**Neon**（マルチモーダル監査）[3]	**HIGH**基本（XHIGH回避）[12]	PDF／画像入力、Agentic Video Understanding（明示的にオン設定）[7]	GDP.pdf検証でXHIGH時に精度低下が見られるためHIGHを維持し、立体構造の重なりは必ず人間が目視確認する[12]。
IPランドスケープ・ソースコード発明発掘	**Orchestrator (Neon)** ＋ **Hydrogen** サブエージェント群[7]	HIGH ～ XHIGH[10]	.xlsx直接解析、Web Grounding for Enterprise、リポジトリ横断解析[4]	通常のGrounding with Google Searchはクエリが最大3日間保存されるため、機密テーマではEnterprise版を使用する[16]。

知財実務におけるコスト試算とオーケストレーション**ROI**

知財部門や特許事務所がGemini 4シリーズをAPI組み込みする際の具体的な投資対効果（ROI）を検証するため、日本の特許実務における標準的な文書分量と公式API単価（Globalリージョン）に基づく3つの業務シナリオを試算する[11]。日本語テキストは概ね1トークン＝約1～2文字に相当するため、①一次スクリーニング用の公報1件（要約＋請求の範囲）を入力2,500トークン・思考込み出力500トークン、②全文クレームチャート用の公開特許公報1件（請求の範囲＋明細書全文で約1.5万～2.5万文字に図面5～8頁のPDF視覚トークンを加算）を入力25,000トークン・出力6,000トークン

（共通の製品仕様プレフィックス20,000トークンは2件目以降の49件で暗黙的キャッシュ読込適用）、③OA対応1案件の包袋一式（本願25,000＋拒絶理由通知4,000＋引用文献3件60,000＝約90,000トークン）に対する4ターン対話解析（初回90,000入力＋8,000出力、2〜4ターン目は95,000キャッシュ読込＋2,000追加指示入力＋各8,000出力）とモデル化し、参考円換算レートを1ドル＝155円とした[11]。



各シナリオの試算から導かれる実務上の示唆は明快である。第一に、「一次スクリーニングと精密対比の二段階分離」による大幅なコスト圧縮である。1,000件の特許母集団すべて（入力計2.5M・出力計0.5Mトークン）を12月1日以降のCarbon標準プランで処理し（2.5 × 4.00ドル + 0.5 × 20.00ドル = 20.00ドル）、さらに上位50件のクレームチャート（通常入力1.27M・キャッシュ入力0.98M・出力0.30Mトークン）を作成すると合計31.28ドル（20.00ドル＋11.28ドル、約4,848円）を要する[11]。これに対し、1,000件の一次スクリーニングをHydrogenのBatch推論（2.5 × 0.75ドル + 0.5 × 3.75ドル = 3.75ドル、約581円）で処理し、絞り込んだ重要50件のみを対象製品仕様20,000トークンの暗黙的キャッシュを効かせたNeon標準プラン（1.27 × 3.00ドル + 0.98 × 0.15ドル + 0.30 × 15.00ドル =

8.46ドル、約1,311円）で精密対比すれば、合計**12.21**ドル（約**1,893**円、全工程**Batch**なら**7.98**ドル＝約**1,236**円）で**1,000**件規模の**FTO**調査プロセスが完結し、**12**月以降の**Carbon**単体構成比で**61%〜74%**のコスト削減となる[11]。

第二に、中間処理（**OA**対応）における暗黙的キャッシュの経済効果と**ZDR**設定とのトレードオフである。約9万トークンの包袋一式を投入して4ターンの対話的検討を行うケースでは、キャッシュ非適用時には毎回9.5万トークン前後の入力課金が発生してNeon標準プランで1.620ドル（約251円）かかるが、4,096トークン以上の共通プレフィックスに95%割引の暗黙的キャッシュがヒットすれば、2ターン目以降の包袋読込コストが圧縮され、1案件4ターン合計で0.811ドル（約126円、約50%削減）へと半減する[7]。ただし後述の通り、Google Cloud上でデータ保持ゼロ（ZDR）を構成する際、インメモリ揮発キャッシュ（retentionType: "EPHEMERAL"）を選択すれば短時間の対話セッション内でのキャッシュ恩恵を維持しやすい一方、キャッシュ自体を完全無効化（disableCache: true）する厳格運用を採った場合にはキャッシュ割引が効かず通常入力単価（上記OA例で1.620ドル、50件クレームチャート例で11.25ドル）が適用される点は、セキュリティ要件とコストの設計判断として事前に織り込んでおく必要がある[4]。

ガバナンス・法的留意点とデータ保持ゼロ（ZDR）実装チェックリスト

どれほどモデルの推論性能と経済性が優れていても、知財業務では出願前の発明情報が外部へ漏洩・公知化すれば新規性を喪失し、権利取得が困難となる重大な法的リスクを伴う[15]。日本国内の知財部門および特許事務所がGemini 4 Neonを導入する際は、日本弁理士会が策定した「弁理士業務AI利活用ガイドライン」（令和7年4月発行）の規範と、Google Cloud（Gemini Enterprise Agent Platform）のデータガバナンス仕様を突き合わせた運用統制を構築しなければならない[15]。

日本弁理士会ガイドラインに基づく法務・倫理原則

法務面の前提として、弁理士法第30条の守秘義務および企業間の秘密保持契約（NDA）、ならびに特許等の新規性喪失リスクへの対応が最優先事項となる[15]。外部事業者の生成AIにおいて入力データがモデル学習に利用される環境へ秘密情報や出願前発明を入力する行為は、第三者への秘密開示（守秘義務違反・NDA違反）や他者への出力を通じた漏洩・新規性喪失につながる恐れがある[15]。この点で実務上特に注意を要するのが、「NeonとHydrogenを組み合わせる際のプラン混在リスク」である。NeonやCarbonにはAPI無料枠が存在せず有料枠のみ（製品改善への学習利用：No）であるのに対し、サブエージェントとして併用されやすいHydrogenにはAPI無料枠（Free Tier）が存在し、無料枠（および招待制**Contributor**枠）では入力データが**Google**の製品改善（学習）に利用される（**Yes**）仕様となっている[13]。したがって、知財業務システムでは無料APIキーや個人向け無料アカウントの利用を技術的・規程的に遮断し、すべてのモデル呼び出しを学習非利用保証が適用される法人向け環境（Gemini Enterprise Agent PlatformまたはGemini in Workspace等）に統一する必要がある[13]。

また、同ガイドラインが明記している通り、「AI側での学習利用の有無にかかわらず、クライアントから得た秘密情報を生成AIに入力する場合は、あらかじめクライアントの合意を得るべき」である点を見落としてはならない[15]。委託元であるクライアント企業自身が情報資産保護の観点から独自に生成AI利用を制限している場合があるため、特許事務所においては受任時の契約や事前合意書において利用基盤とデータ保護条件を開示し、承諾を取得する運用が求められる[15]。

さらに、民法第644条に基づく善管注意義務（ファクトチェック義務）および他人の知的財産権への配

慮も不可欠である[15]。NeonはGoogle社内テストで低い捏造率を示したとはいえ、公式モデルカードにはハルシネーションの限界と**2026**年**3**月の知識カットオフが明記されている[3]。実在しない公報番号や誤った段落引用、および2026年4月以降の制度改正（AI関連発明やネットワーク関連発明等の産業財産権制度見直しを含む）・最新判例の欠落については、必ず弁理士・担当者自身が一次情報に基づいて検証・判断しなければならない[15]。なお、先行技術調査等の情報解析目的で他者の著作物をプロンプト入力する行為は原則として著作権法第30条の4により適法であるが、元著作物の表現を享受・類似出力させる目的での入力や、生成物による他者権利の侵害がないかの確認には注意が必要である[15]。あわせて、一般提供（GA）版のGeminiモデルをGoogle CloudのGemini Enterprise Agent Platform APIやGemini in Workspace上で利用する場合、Googleの生成AI補償対象サービス（Generative AI Indemnified Services）に含まれるものの、補償の具体的な適用要件や対象となる知的財産権の細目はサービス固有規約（Service Specific Terms）で個別に確認する必要がある[38]。

Google Cloudにおけるデータ保持ゼロ（ZDR）実装の4大チェックリスト

システム設計者が陥りやすい誤解は、「有料APIやGoogle Cloudを使えば初期設定のままでサーバー上に一切データが残らない（ZDRになる）」という思い込みである。Google Cloud公式のZDRドキュメントによれば、事前の許可なく顧客データがモデル学習に使われないことは保証されているものの、キャッシュやデバッグログを含めた厳格な「データ保持ゼロ（ZDR）」を実現するには、以下の4つの技術設定と機能選択を実装しなければならない[7]。

1. **Interactions API**における状態保存の明示的オフ（**store = false**）: マルチターン対話やオーケストレーションに用いるInteractions APIでは、storeパラメータを指定しない場合、すべてのモデルでデフォルトの**true**が適用され、プロンプト・レスポンス・会話状態が保存される[16]。ZDRを実現するには、APIリクエスト時に明示的にstore = falseを設定するとともに、BigQueryへのリクエスト／レスポンスロギング機能を有効にしない（デフォルトの無効状態を維持する）必要がある[16]。
2. 永続コンテキストキャッシュ（**Durable Caching**）の無効化またはインメモリ限定化: Google Cloud上のNeonでは、デフォルトで永続キャッシュ（Durable Caching）が有効になっている[4]。永続ストレージへのデータ保持を排除するには、Agent Platform管理者ロール（roles/aiplatform.admin）を持つ管理者がプロジェクトのcacheConfigエンドポイント（https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT_ID/cacheConfig）に対してPATCHリクエストを実行し、キャッシュ機能の無効化（disableCache: true）またはインメモリ揮発キャッシュのみへの制限（retentionType: "EPHEMERAL"）を構成する必要がある[16]。
3. 標準**Google**検索グラウンディングの回避と**Enterprise**版への切り替え: 未公開発明の周辺調査などで通常のGrounding with Google Searchを使用すると、システムデバッグ目的でプロンプト由来のクエリやコンテキスト情報が最大**3**日間保存され、この保存は無効化できない（Grounding with Google Mapsでは30日間保存）[16]。ZDRが求められる機密知財業務で外部Web情報を参照させる場合は、公式ドキュメントが推奨するWeb Grounding for Enterpriseを採用するか、機密情報を含まない公開検索語のみを切り出して別系統で検索する設計を採るべきである[16]。
4. フロンティアモデルの不正使用監視ログ確認と外部文書からの**IPI**防御: Neonはサービス固有規約第34条（Abuse monitoring）および高度なAI安全性に関する追加条項の対象となるGoogle Frontier Modelであり、不正使用検知のためにプロンプトやレスポンスのロギングが

行われ、標準設定のままではZDRにできない場合があるため、完全なZDR適用にあたってはGoogle Cloudの担当アカウントチームへの確認・申請手続きを行う必要がある[7]。また、外部の公開公報PDFや第三者Webページを自律エージェントに取り込ませる際は、外部文書に埋め込まれた間接プロンプトインジェクション（IPI）による機密コンテキスト流出を防ぐため、ツール実行権限の最小化・サンドボックス化とあわせて、リアルタイムセーフガードおよび高思考レベルの適用を組み合わせた多層防御を実装することが推奨される[3]。

引用文献

1. Gemini 4 Neon の概要 | npaka - note, https://note.com/npaka/n/n95ac3fce7543
2. Gemini 4 Neonとは？ArgonやCarbonとの違い・利用枠・他社比較, https://snowsystem.net/ai/gemini-4-neon/
3. Introducing Gemini 4 Neon - Google Blog, https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-4-neon/
4. https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/4-neon
5. Gemini 4 Neon」発表 「Carbon」と「Hydrogen」の中間を担う, https://www.itmedia.co.jp/news/article/2611/19/2000003315/
6. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-4-Neon-Model-Card.pdf
7. https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/guides/gemini-4-neon
8. https://gigazine.net/news/20261119-google-gemini-4-neon/
9. https://www.vals.ai/benchmarks/legal_research
10. https://antigravity.google/blog/gemini-4-neon-in-google-antigravity/
11. Agent Platform Pricing | Google Cloud, https://cloud.google.com/model-garden/pricing
12. https://xenospectrum.com/gemini-4-neon-model-comparison/
13. https://ai.google.dev/gemini-api/docs/pricing?hl=ja
14. Google continues to accelerate its release cadence with Gemini 4, https://artificialanalysis.ai/articles/gemini-4-neon-intelligence-gain
15. https://www.jpaa.or.jp/cms/wp-content/uploads/2025/04/AIservices-guideline.pdf
16. https://docs.cloud.google.com/gemini-enterprise-agent-platform/resources/zero-data-retention?hl=ja
17. After Gemini 4 Carbon, Google may be launching a Hydrogen model, https://www.indiatoday.in/technology/news/story/after-gemini-4-carbon-google-may-be-launching-a-hydrogen-model-3037253-2026-11-12
18. Google Rilis AI Gemini 4 Hydrogen, Bisa Bikin Kode, Analisis Data, dan Minim "Halu", http://tekno.kompas.com/read/2026/11/13/13030027/google-rilis-ai-gemini-4-hydrogen-bisa-bikin-kode-analisis-data-dan-minim-halu-
19. Gemini 4 Argon：フロンティア知能の新時代 - Google Blog, https://blog.google/intl/ja-jp/company-news/technology/gemini4argon/
20. Gemini 4 Carbon is now available - Google Blog,

https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-4-carbon-now-available/
21. Gemini 4 Argon vs Carbon vs Hydrogen: Which to Choose? - RuneLab, https://runelab.ai/en/blog/post/which-gemini-4-model-to-choose
22. Google just released Gemini 4 Neon, "near-Carbon quality at Flash, https://www.reddit.com/r/GeminiAI/comments/1xzq3hn/google_just_released_gemini_4_neon_nearcarbon/
23. Models - Gemini API | Google AI Studio, https://aistudio.google.com/docs/models
24. Latest Gemini models - Gemini API | Google AI Studio, https://aistudio.google.com/docs/latest-model
25. Gemini 4 Hydrogen - Model Card - Google DeepMind, https://deepmind.google/models/model-cards/gemini-4-hydrogen/
26. Gemini 4 Hydrogen vs Gemini 4 Carbon: Price, API & Specs (2026), https://meetcody.ai/models/compare/gemini-4-hydrogen-vs-gemini-4-carbon/
27. Introducing Gemini 4 Hydrogen for Agentic Workflows - Google Blog, https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-4-hydrogen/
28. Gemini 4 Hydrogen Pricing: Free Tier, $1.50/7.50 per 1M, https://www.ayautomate.com/blog/gemini-4-hydrogen-pricing
29. Zero Data Retention for Gemini and Vertex AI: A Life Sciences Deep, https://intuitionlabs.ai/insights/guides/zero-data-retention-gemini-vertex-ai-life-sciences
30. Google Cloud 2026年上半期アップデート総まとめ #AI - Qiita, https://qiita.com/asayan_mana/items/5d040d2390224535353a
31. Google AI plans - Gemini API, https://ai.google.dev/gemini-api/docs/google-ai-plans
32. Gemini 4 Hydrogen / Carbon / Argon / Neon think 3-4x less since, https://discuss.ai.google.dev/t/gemini-4-hydrogen-carbon-argon-neon-think-3-4x-less-since-nov-18-evening-utc-xhigh-rejected-cli-1-5-11-ultra/202317
33. https://news.ycombinator.com/item?id=50552424
34. Batch inference with Gemini | Gemini Enterprise Agent Platform, https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/capabilities/batch-inference
35. Rate limits - Gemini API - Google AI Studio, https://aistudio.google.com/docs/rate-limits
36. Gemini Enterprise release notes | Google Cloud Documentation, https://docs.cloud.google.com/gemini/enterprise/docs/release-notes
37. An Easy Guide to Gemini 4's Hydrogen, Argon, and Carbon - note, https://note.com/glad_quail5544/n/n85dc2880e947?hl=en
38. Google Cloud Generative AI Indemnified Services, https://cloud.google.com/terms/generative-ai-indemnified-services
39. 【特許庁ステータスレポート2026】知財出願動向とAI時代の知財戦略, https://www.migalio.com/jpo-status-report-2026-ai-strategy-88/
40. Gemini Zero Data Retention: Paid API vs Free Tier - Meetily, https://meetily.ai/llm-privacy/gemini

41. How to request Zero Data Retention for Vertex AI? - Gemini API, https://discuss.ai.google.dev/t/how-to-request-zero-data-retention-for-vertex-ai/200809