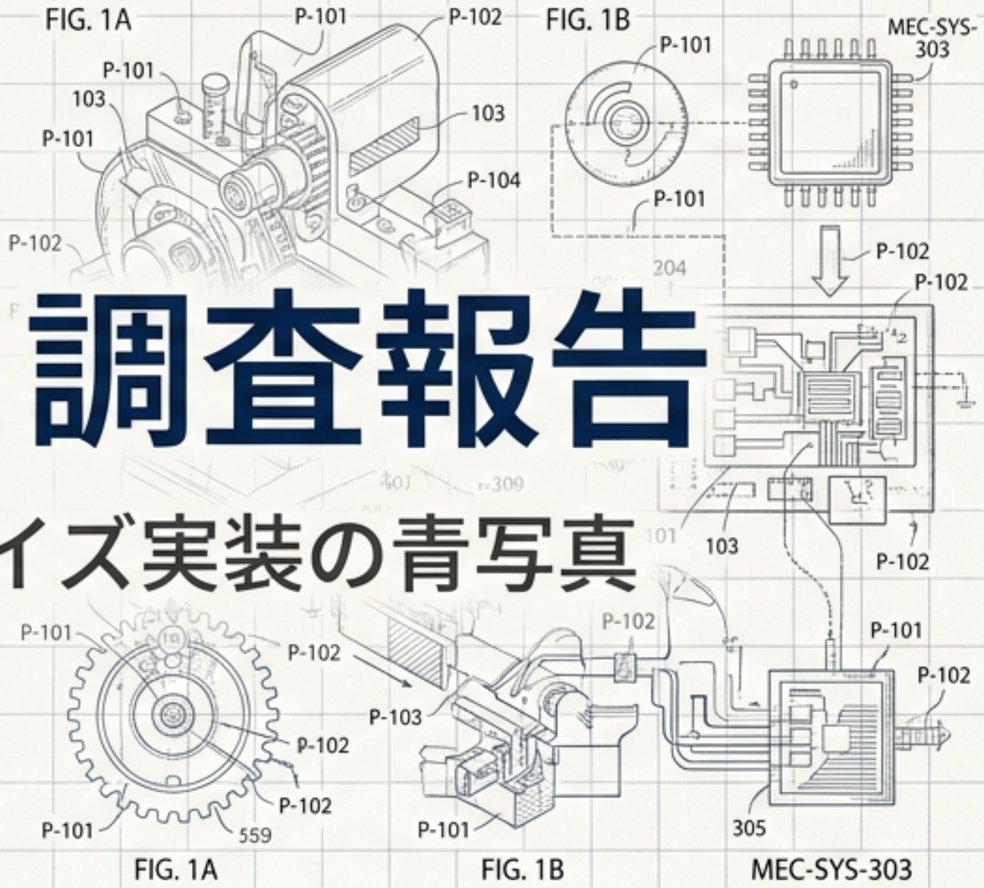
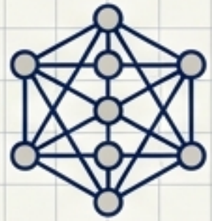
MiniMax M3 調査報告

知財実務とエンタープライズ実装の青写真



機密事項取扱基準・AI実装ガイダンス準拠



428B / 23B

総パラメータ数 428Bに対し、推論時の活性化パラメータを23Bに圧縮する高効率MoEアーキテクチャ。

1M Tokens

最大1,048,576トークンのコンテキストウィンドウ。特許明細書や大規模ポートフォリオの一括読込に対応。APIでは最低512Kを保証。

Native Multimodal

テキスト・画像・動画のネイティブ入力に対応。「Step zero」からの統合学習。（※音声は現在未対応）

Open Weights

Hugging Face等で重みを公開。オンプレミスや専有VPC環境の配備が可能であり、機密性の高い知財データに最適。

MiniMax M3 公開タイムライン

2026-05-27

M3と新 sparse attention の事前報道・ティーザー。

2026-06-01

MiniMax 公式ブログでM3 正式発表。1M context / MSA / native multimodality / benchmarkの公表。「10日以内に technical reportと open weights を公開」と告知。

2026-06 中旬

Hugging Face 公式組織に MiniMax-M3本体と MXFP8版が公開。

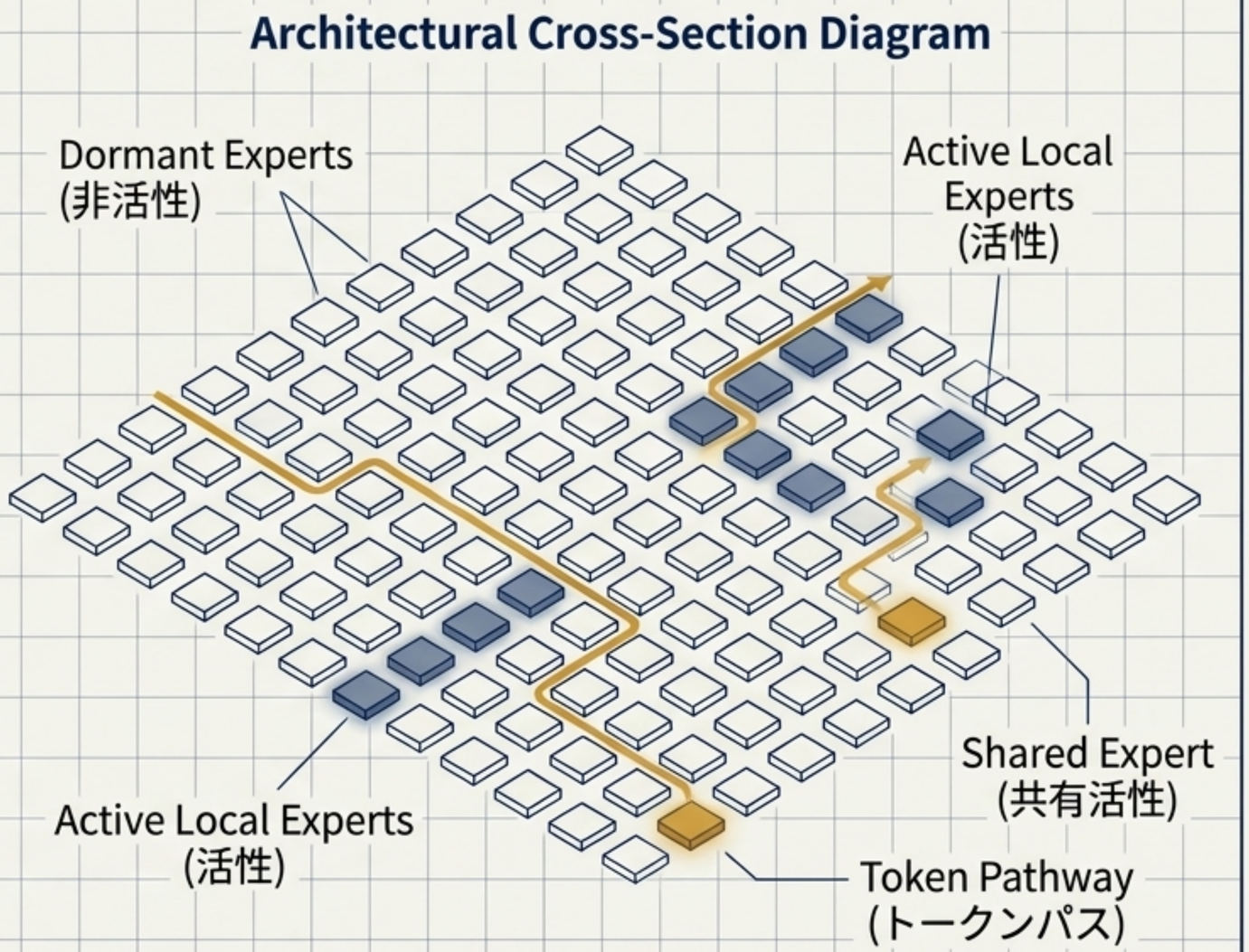
2026-06 中旬

Together AI / NVIDIA NIM / OpenRouter 系プロバイダで提供拡大。短期間で強力な実装エコシステムを形成。

アーキテクチャのDNA: 高度な疎結合ルーティング

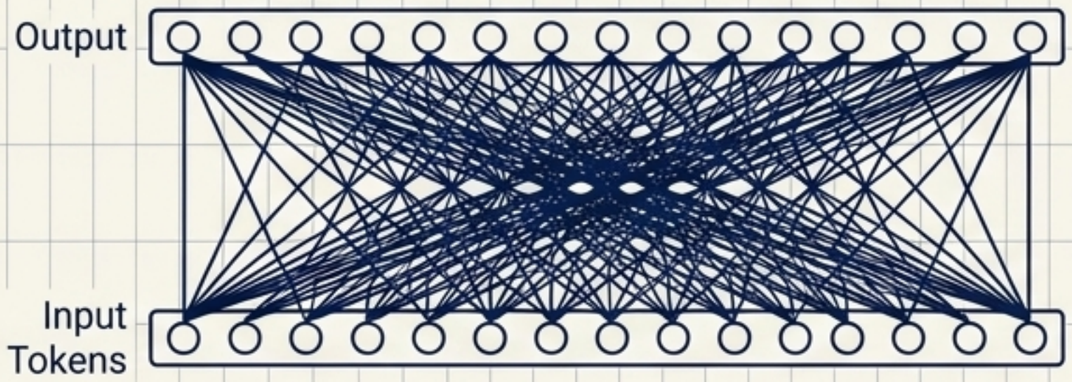
1	**テキスト層数: 60（最初の3層は Dense、残り57層が MoE）** • Attention heads: 64 / KV heads: 4
2	**Local experts: 128** • 活性化ノード: トークンあたり 4 Local Experts + 1 Shared Expert を選択。

全体は巨大（428B）だが、各トークン処理では極めて限定されたパス（23B）のみを使用。推論コストを総パラメータに対して大幅に圧縮する設計。



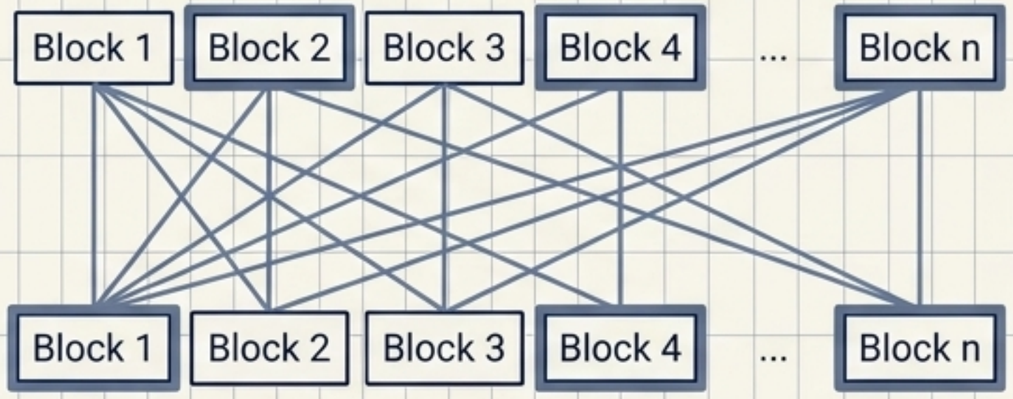
長文処理のブレイクスルー: MSA (MiniMax Sparse Attention)

Quadratic Full Attention



Inefficiency

MiniMax Sparse Attention (MSA)



Efficiency

Block size: 128 / Top-k blocks: 16

KVキャッシュをブロック単位で選別し、上位ブロックのみにDense Attentionを張る独自方式。

計算量削減	Prefill 高速化	Decode 高速化
1/20	**9倍超**	**15倍超**
1トークンあたりの計算量を前世代から圧縮		

モデル設計だけでなく、疎注意カーネル、Paged Attention、KV Cache管理のシステム全体最適化により1M文脈を実現。

パフォーマンスの実態: ベンチマークと「Scaffolding」の要件

SWE-Bench Pro:

59.0%

Terminal-Bench 2.1:

66.0%

BrowseComp:

83.5

Reality Check

1. Agent Harness 依存

公式スコアの多くは、MiniMax自社インフラと特定のスキャフォールディング（Terminus 2, Mini-SWE-Agent等）上で計測。「モデル単体のIQ」ではなく「複合システム性能」として読むべき。

2. 欠落データ

MMLU や HumanEval の単独スコアは公式未発表。

3. 思考遅延（TTFT）

出力速度は約 57.8 tokens/s と高速だが、Thinking（深い思考）を含む場合、Time to First Answer Token は 36.8〜41.3秒に達する。即レス型ではなく熟考型。

インフラストラクチャの負荷と経済性

Local Deployment (Hardware Weights)

BF16 本体版 (854GB) : 実稼働には
16基の 80GB GPU クラスが計画値の目安。

MXFP8 量子化版 (444GB): 重みだけで444GB。
最低でも 8基の 80GB GPU クラスが必要。

※フルSFT（微調整）はマルチノード前提の重い作業。
初期はRAG + 小規模アダプタを推奨。

Pay-As-You-Go API Costs

標準入力（512K以下）:

入力 $0.30 / 出力 $1.20
(per 1M tokens)

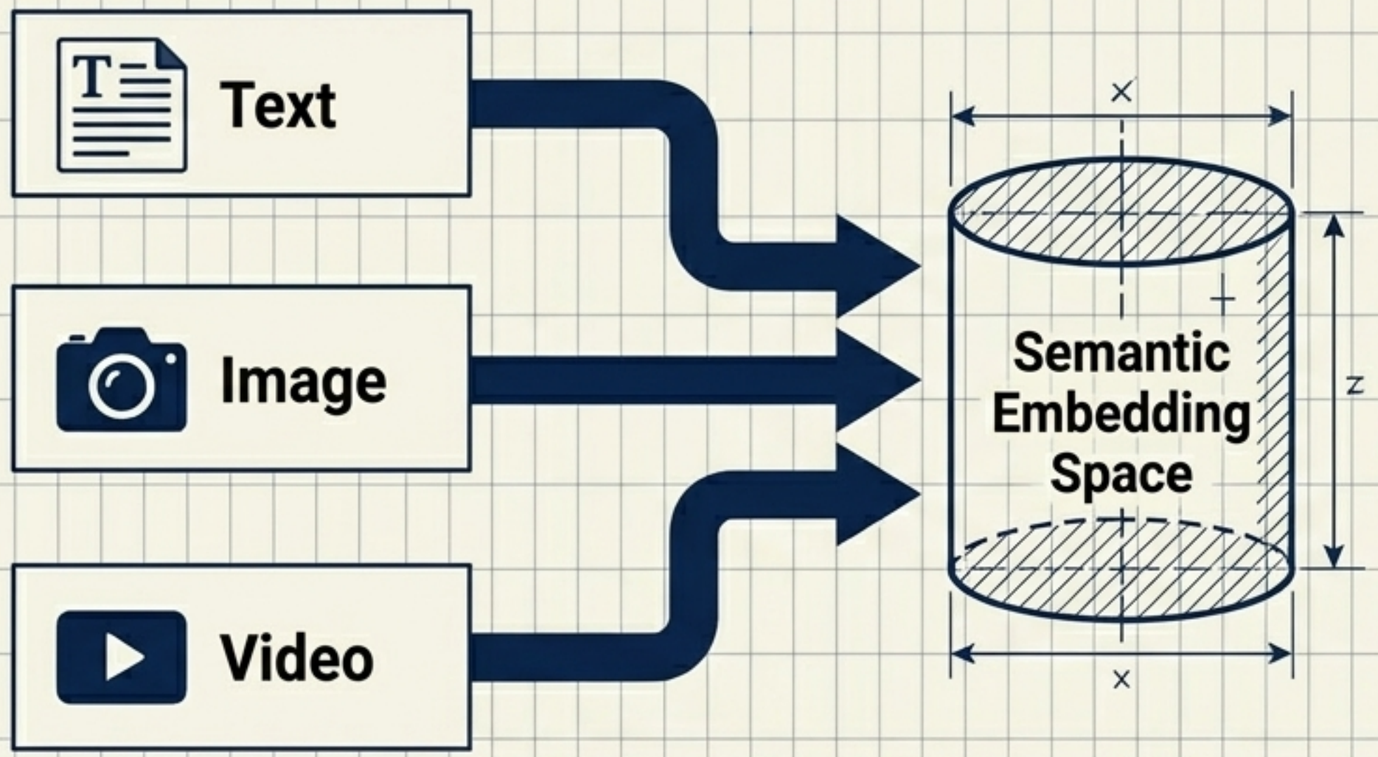
長文入力（512K超）:

入力 $0.60 / 出力 $2.40
(per 1M tokens)

Priority モード:

各 1.5倍。
知財ポートフォリオ一括読込（入力1M/出力贔8k）の概算コストは約$0.61〜$0.92。

ネイティブ・マルチモーダリティ: 後付けではない視覚統合



Key Technical Specs

- Vision Stack: CLIP系 ViT (32層 / hidden 1280)
- 動的解像度: 336 ～ 2016 (Dynamic Resolution)
- 動画対応: Video-MME 84.6 (512 frames条件)

The Adapter-Free Reality

「Step zero」からの Mixed-Modality Training。視覚を後付けのアダプタで接ぐのではなく、初期段階からテキストと視覚の意味空間を統合。

知財実務における価値: 複雑な機械図面やフローチャートの構造を、テキストクレームと高精度に紐付ける基盤となる。

エンタープライズ適合性マトリクス: オープン vs クローズド

	MiniMax M3	Gemini 3.1 Pro	GPT-5.5
ライセンス制限	minimax-community (商用表示義務/年商要件あり)	Proprietary	Proprietary
配備の自由度	完全オンプレ/VPC可	API/Cloud依存	API/Cloud依存
入力モダリティ	Text / Image / Video	Advanced Multimodal	非開示
SWE-Bench Pro	59.0% (公式)	54.2% (二次情報)	58.6% (二次情報)

絶対性能一本勝負ではなく、
「長文・視覚・配備自由度の総合値」がM3の競争優位。

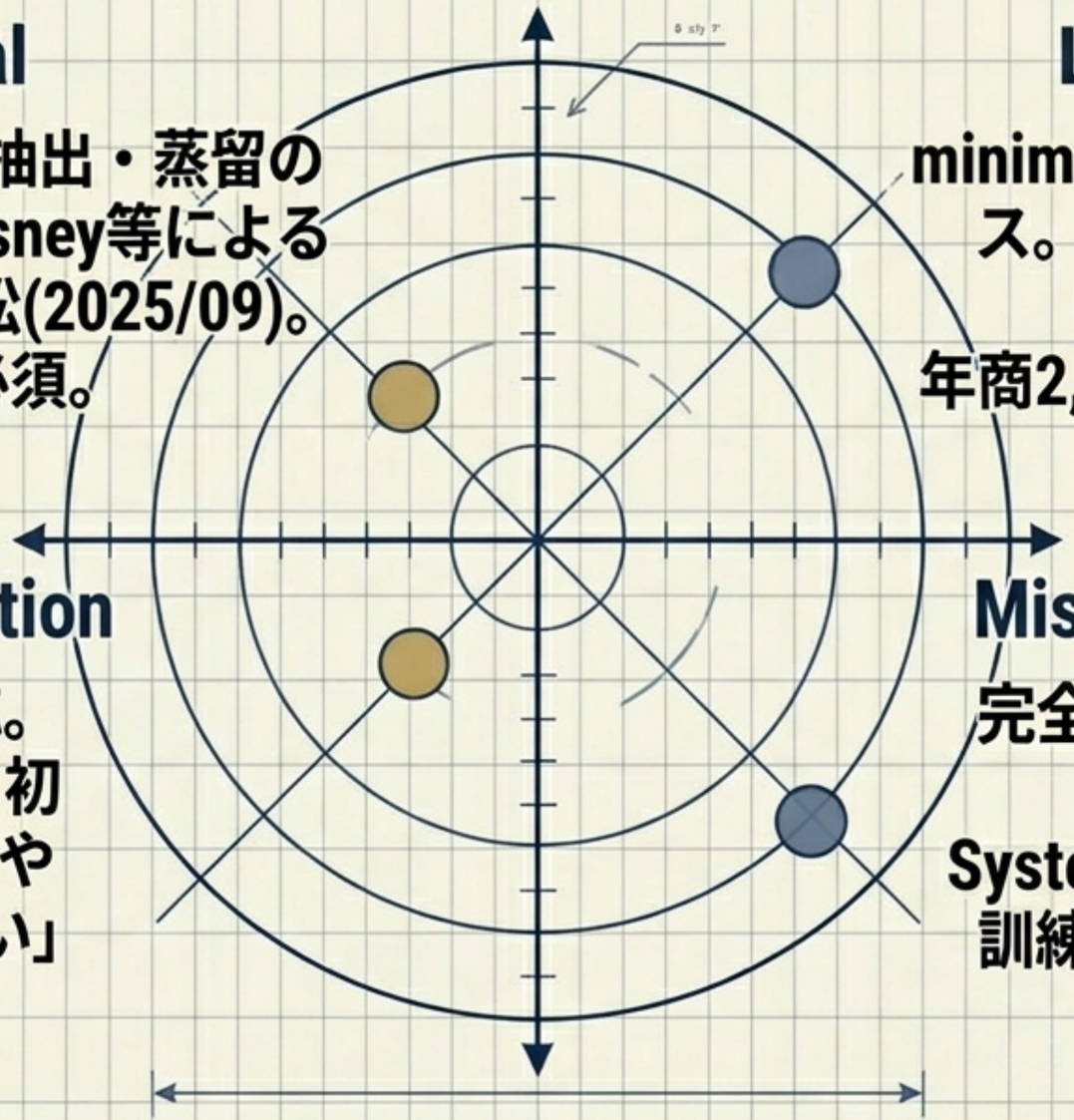
リスク＆ガバナンス・レーダー

Legal & Reputational

Anthropicによるデータ抽出・蒸留の主張報道(2026/02)、Disney等によるHailuo AIへの著作権訴訟(2025/09)。ベンダー審査での確認必須。

License Restrictions

minimax-community ライセンス。商用利用時の Built with MiniMax M3 表示義務。年商2,000万ドル超は事前許諾必須。



Implementation Friction

HF Communityでの反応。「広く触られているが、初期の詰まり（QAT版要求やindex.json欠落等）が多い」荒削りな実装フェーズ。

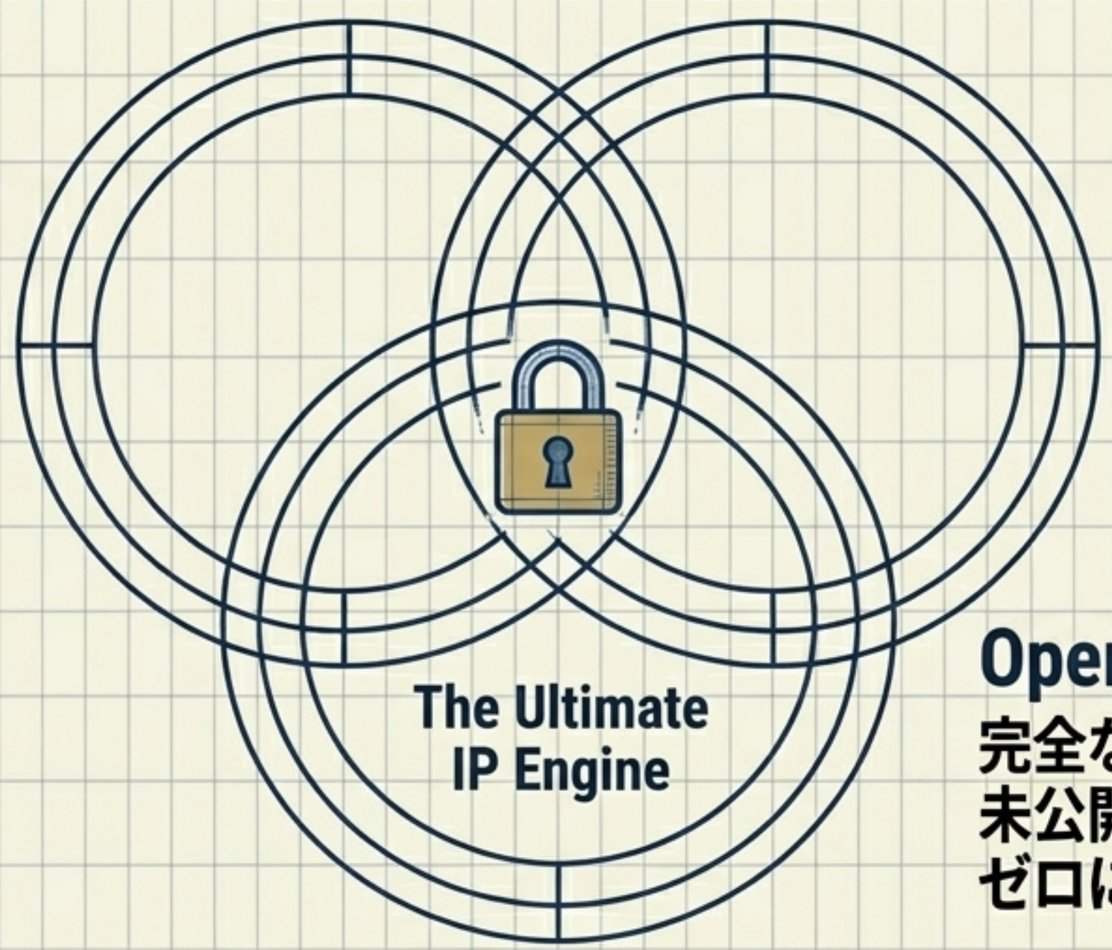
Missing Documentation

完全な Technical Report、Report、M3固有の System Card / Safety Card、訓練コーパス詳細が未公開。

戦略的シンセシス: なぜ M3 が「知財特化型」の最適解なのか

1M Context

請求項、先行技術集合、
ファミリー単位の長文・
長表を一度に保持。



Native Vision

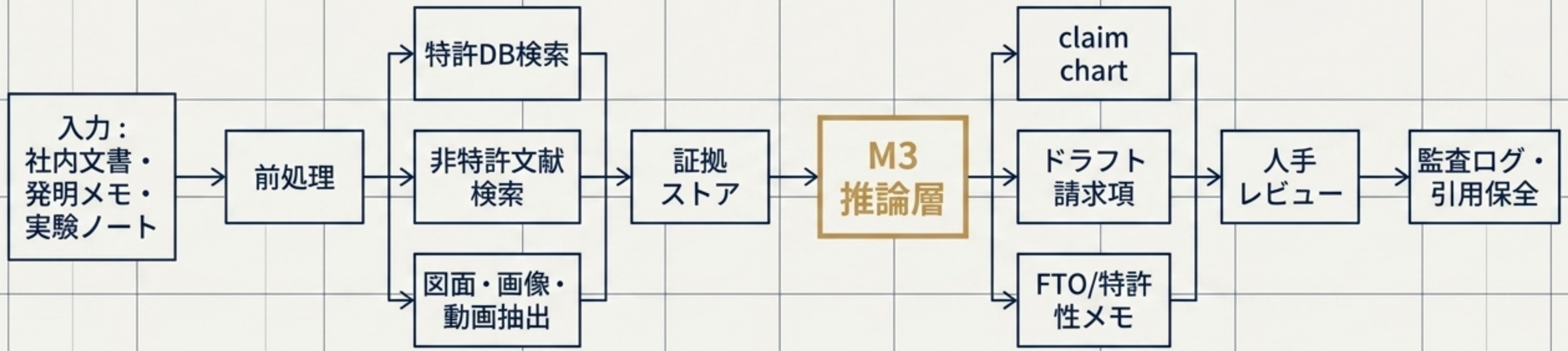
複雑な特許図面
(Drawings) を直接解釈し、
クレーム要素と対応付け。

Open Weights

完全なオンプレミス配備により、
未公開発明のデータ漏洩リスクを
ゼロに（API直投げの回避）。

**単独の「知能」で他社を圧倒するのではなく、知財実務に求められる
「巨大文脈＋視覚＋絶対的機密性」の構造的合致こそがM3の真の価値である。**

知財ワークフローのリファレンス・アーキテクチャ



検索と証拠抽出は確定的ツールに任せ、
整理・草案をM3に寄せるRAG構成が品質を安定させる。

知財ワークフローと「Human-in-the-Loop」検証マトリクス

用途	M3の強み	推奨プロンプト例	必須の人手検証
先行技術調査	1M文脈で複数公報保持	段落番号付きClaim Chart作成	原文の段落・図番号の照合。
特許性分析	差分抽出が高速	肯定/否定両面からの論点整理	進歩性判断は弁理士が最終レビュー。
クレームドラフト	長い実施形態の統合	独立・従属請求項案とサポート箇所明示	サポート要件・明確性の確認。
FTO予備評価	ファミリー構成比較	侵害リスクの暫定評価と不明確箇所の列挙	Claim Constructionを含む弁護士主導の最終判断。
図面・動画検証	ネイティブ視覚入力	構成要素抽出・動画の時系列イベント抽出	図面番号・実証動画の真正性確認。

各国特許庁のコンプライアンスと「発明者の境界線」

The Core Rule: AIは「ツール」であり、「発明者」は自然人に限られる。

JPO（日本）

発明者欄にAIを記載することは不可。

EPO（欧州）

発明者（Inventor）は人間（Human）でなければならない。

USPTO（米国）

2025年改訂ガイダンスに準拠。AIは人人間発明者の道具。既存ルールの遵守とリスク緩和が必須。

M3は発明者候補の列挙やマッピング補助には極めて有用だが、AI出力を無検証で出願書類へ流し込む自動化運用は法的に厳禁。常に最終責任は人間が負う。

エンタープライズ・パイロット導入チェックリスト

- ☑ **Target Data (対象データ): 公開公報、または社内で開示許可済みの案件のみに限定。**
- ☑ **Deployment (配備環境) : 専有VPC環境の構築。未公開発明の公開APIへの送信は厳禁**（NIST AI RMF準拠）。
- ☑ **Governance (ガバナンス) : 案件**（Matter）**単位のアクセス制御、入力ログの最小化、案件終了時のデータ削除ルールの策定。**
- ☑ **License Check (ライセンス) :** minimax-community **の表示義務確認。対外サービス化の場合は法務審査を必須化。**
- ☑ **Success/Withdrawal Metrics (撤退条件) : 「モデルの賢さ」ではなく、根拠引用率、誤引用率（ハルシネーション）、レビュー時間削減効果で測定。誤引用率が基準を超過した場合は即時撤退。**

確証的ツール(Deterministic tooling)との併用により、
M3は知財戦略の最強の補助エンジンとなる。