2026年10月 / Executive Briefing & Technical Advisory

ELYZA「LLM-jp-4」推論モデル 総合技術評価

o1/o3との比較、エンタープライズ導入価値、および戦略的ポジション

Gemini Notebook

エグゼクティブサマリ：総合判定

日本語オープンウェイトとしての価値 - 高い

LLM-jp-4からの明確な性能向上。Apache 2.0、64Kコンテキスト、ツール呼びし対応により、自社運用モデルとして極めて高い技術価値を持つ。

o1/o3フルモデル超え - 証拠なし/下回る

汎用的な高度推論においてo1/o3を凌駕したという証拠はない (GPQA Diamondスコア等で劣後)。

オンプレ・閉域導入適性 - 高い

重みの自社管理が可能であり、機密性の高いRAGや社内文書QAに最適。API型フロンティアモデルとは競争軸が異なる。

即時の商用本番投入 - 条件付き

発表翌日時点での評価であり、実機性能、安全性評価、業務固有のE2E検証が必須。PoCは強く推奨される。

Gemini Notebook

メディアのナラティブと技術的事実の照合

Card 1

NOISE（誤解/過大評価）	SIGNAL（技術的事実）
「完全国産AI」：部材・データ・ソフトウェアまで100%国産である。	ELYZA側も訂正済み。ベースはNIIモデルだが、グローバルな技術スタックに依存。 「国内企業が作成した」が正確な表現。

Card 2

NOISE（誤解/過大評価）	SIGNAL（技術的事実）
「o1を超えた」：OpenAIのo1/o3を凌駕する最高性能モデルである。	共通ベンチマーク（GPQA Diamond）において、o1（75.7）に対しELYZA 33B（69.95）と下回る。 汎用推論ではo1/o3優勢が保守的かつ妥当な推定。

Card 3

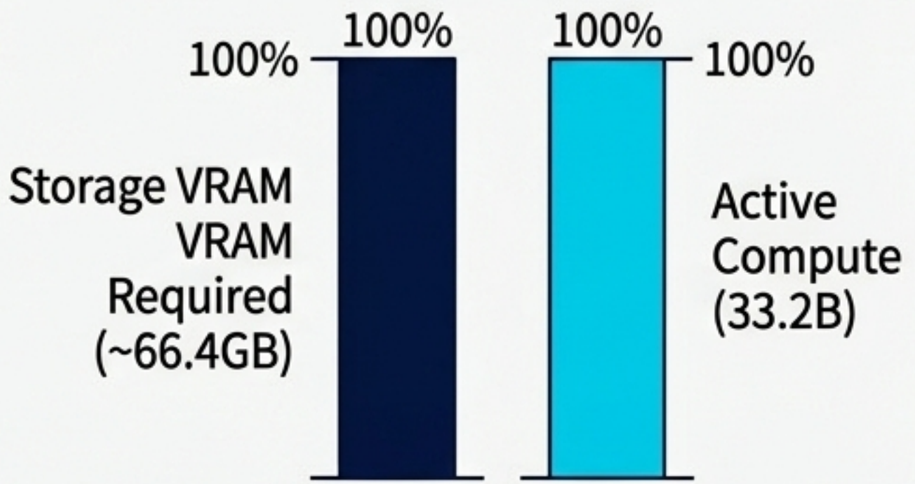
NOISE（誤解/過大評価）	SIGNAL（技術的事実）
「導入実績多数」：既に多くの企業で本番導入されている。	Hugging Faceの初動ダウンロード数（約400件）やELYZA Voice Agentの開発発表は、本モデルの「企業での本番採用」を直ちに意味するものではない。

Gemini Notebook

モデルラインナップとアーキテクチャの現実

Model A

ELYZA 33B Dense (The Heavy Lifter)

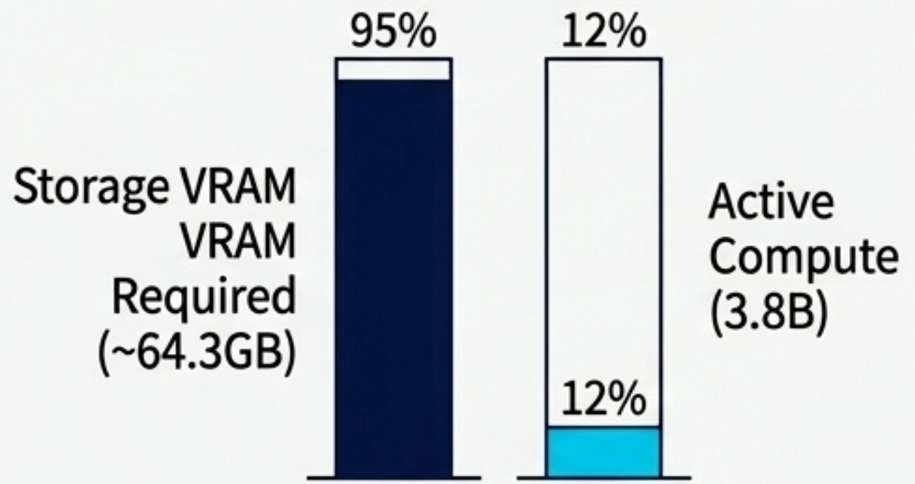


総パラメータ: 33.2B / Active Params: 約33.2B

特徴: 全ウェイトを毎トークン使用。最高品質だが推論時の計算コストが高い。

Model B

ELYZA 32B-A3B MoE (The Efficiency Play)



総パラメータ: 32.1B / Active Params: 約3.8B

特徴: 128エキスパート中8つを活性化。計算量は削減されるが、重み保持のため必要なVRAMは33Bとほぼ同じである点に注意。

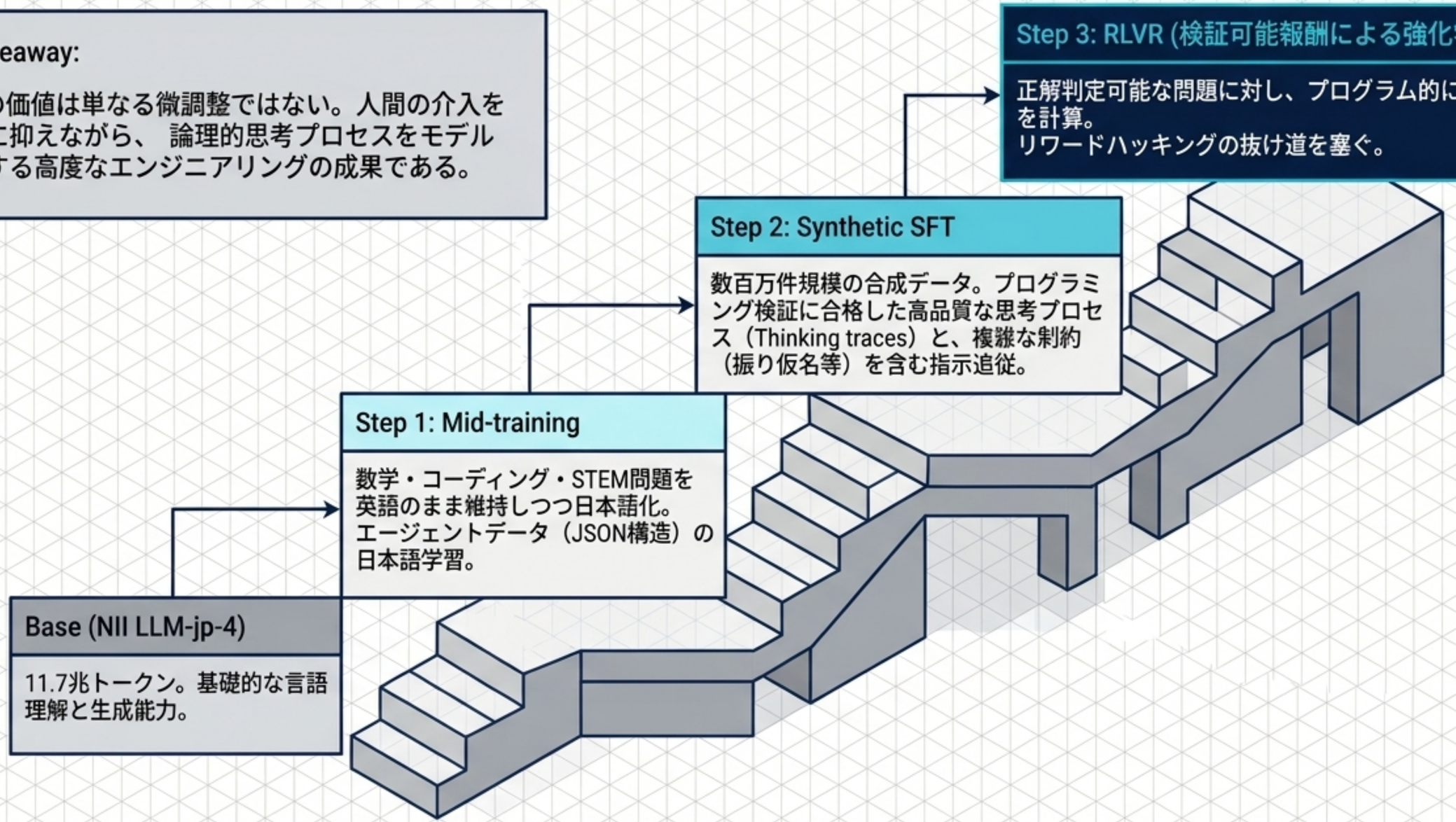
共通仕様: 65,536トークンコンテキスト、常時Thinking、Tool calling対応、Apache 2.0ライセンス。

Gemini Notebook

ベースモデルを推論モデルへ昇華させる3段階の学習パイプライン

Key Takeaway:

ELYZAの価値は単なる微調整ではない。人間の介入を最小限に抑えながら、 論理的思考プロセスをモデルに強制する高度なエンジニアリングの成果である。



Base (NII LLM-jp-4)

11.7兆トークン。基礎的な言語理解と生成能力。

Step 1: Mid-training

数学・コーディング・STEM問題を英語のまま維持しつつ日本語化。
エージェントデータ（JSON構造）の日本語学習。

Step 2: Synthetic SFT

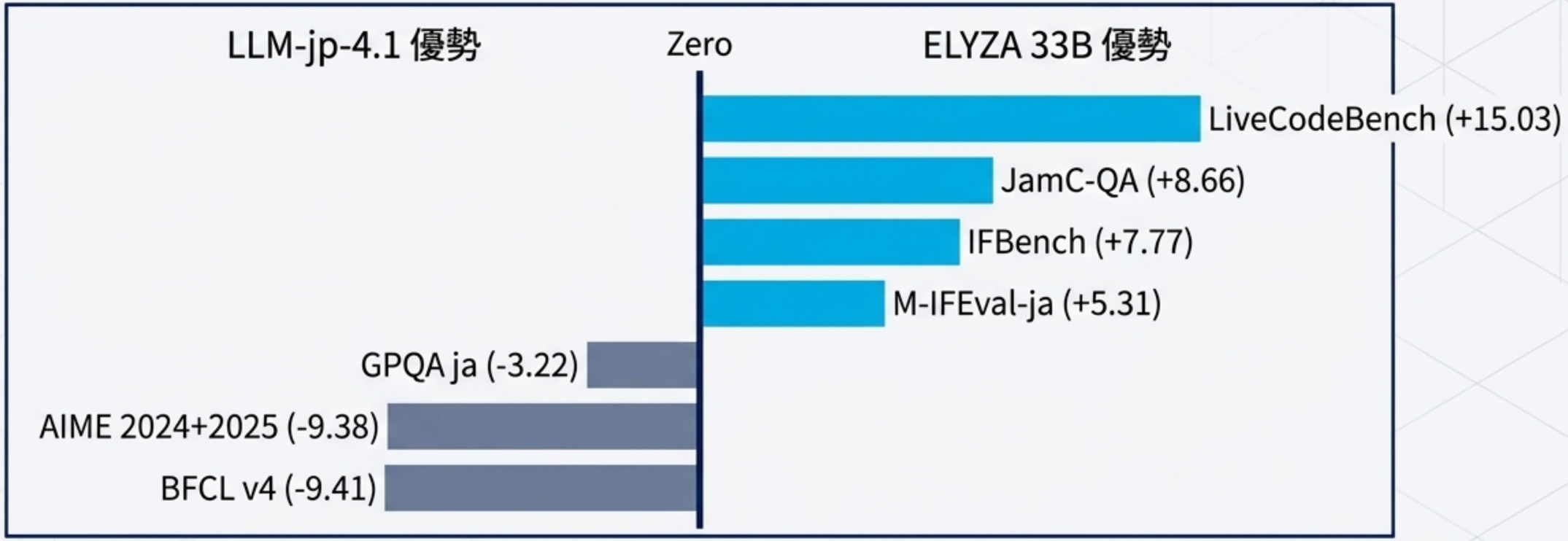
数百万件規模の合成データ。プログラミング検証に合格した高品質な思考プロセス（Thinking traces）と、複雑な制約（振り仮名等）を含む指示追従。

Step 3: RLVR (検証可能報酬による強化学習)

正解判定可能な問題に対し、プログラム的に報酬を計算。
リワードハッキングの抜け道を塞ぐ。

Gemini Notebook

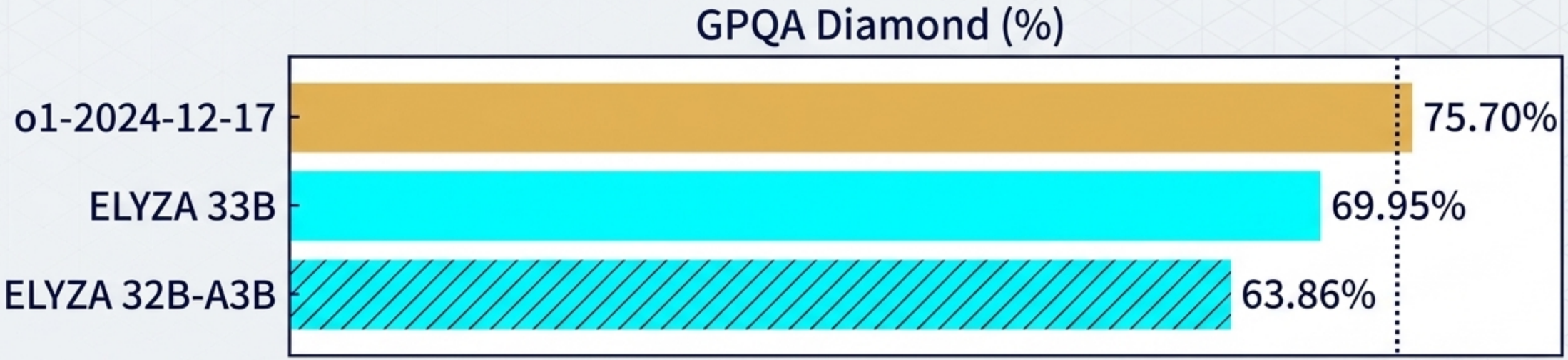
性能評価 I：進化の方向性（vs LLM-jp-4.1）



Insight:
全面的な勝利ではない。生粋の数学 (AIME) や純粋な関数呼び出しでは4.1に劣るが、日本語のQA、厳密な指示追従、コーディングにおいてELYZAは劇的な改善を見せ、総合スコア（Overall 61.69 vs 60.64) で逆転している。

Gemini Notebook

性能評価 II：フロンティアとの距離（vs o1/o3）



Analytical Context:
同一ハーネスではないため厳密な対照試験ではないが、「ELYZA 33Bがo1を上回った」という証拠にはならず、むしろ逆方向の証拠である。
OpenAI公式において「o3は同一レイテンシ・コスト条件のo1より高性能」とされている。

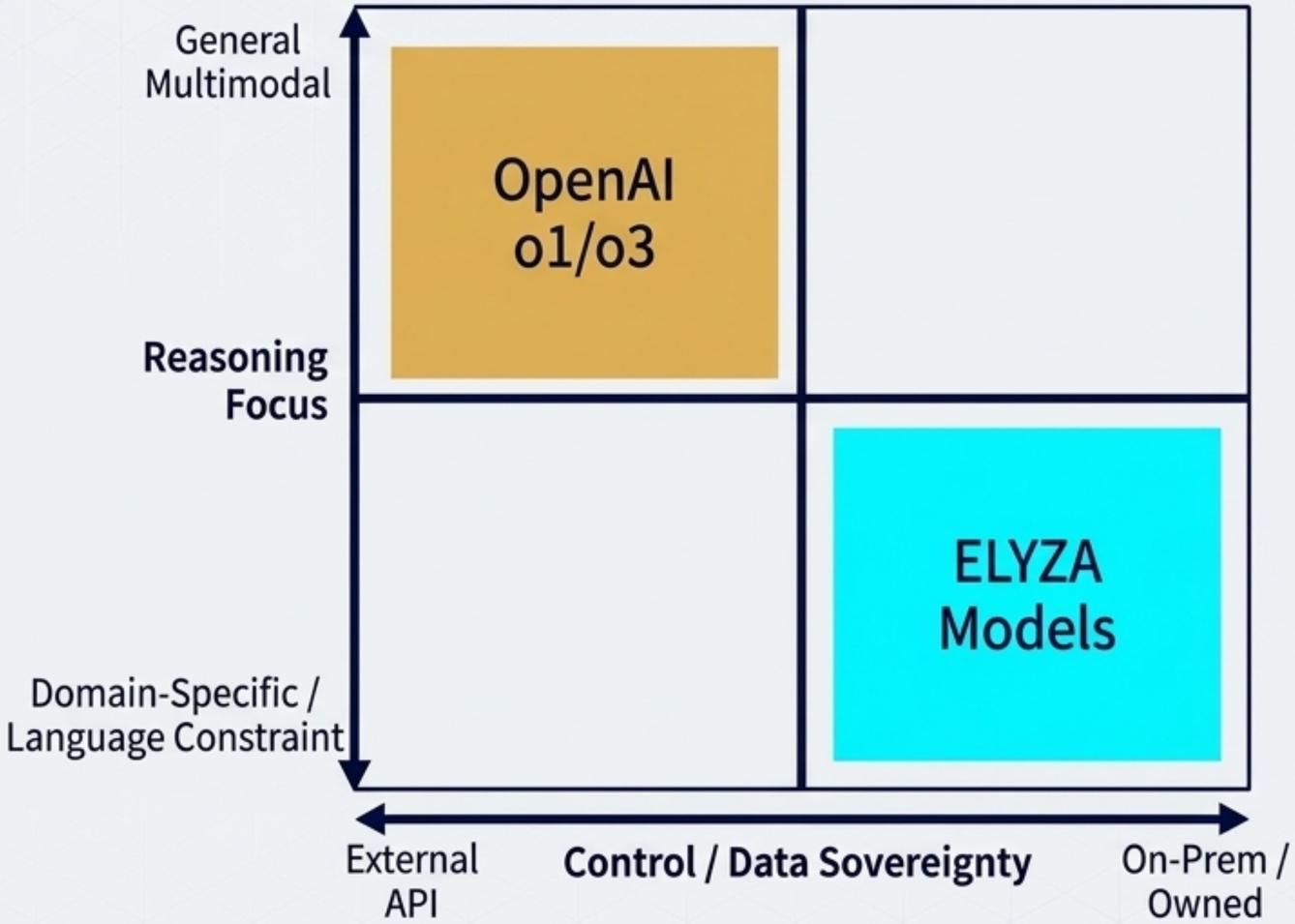
結論:

汎用的な高度推論能力においては、現在も o3 > o1 ≳ ELYZA 33B と推定するのが最も合理的である。

Gemini Notebook

真の競争軸：生殺与奪の権を握るか、外部APIに委ねるか

ELYZAモデルはo1/o3と「生の知能」で直接競争するものではない。その本質的価値はエンタープライズの「主権」にある。



1. データ主権と閉域運用

機密情報を外部APIに送れない環境（RAG、社内文書QA）での完全なオンプレミス稼働。

2. 重みの自己保有と改変

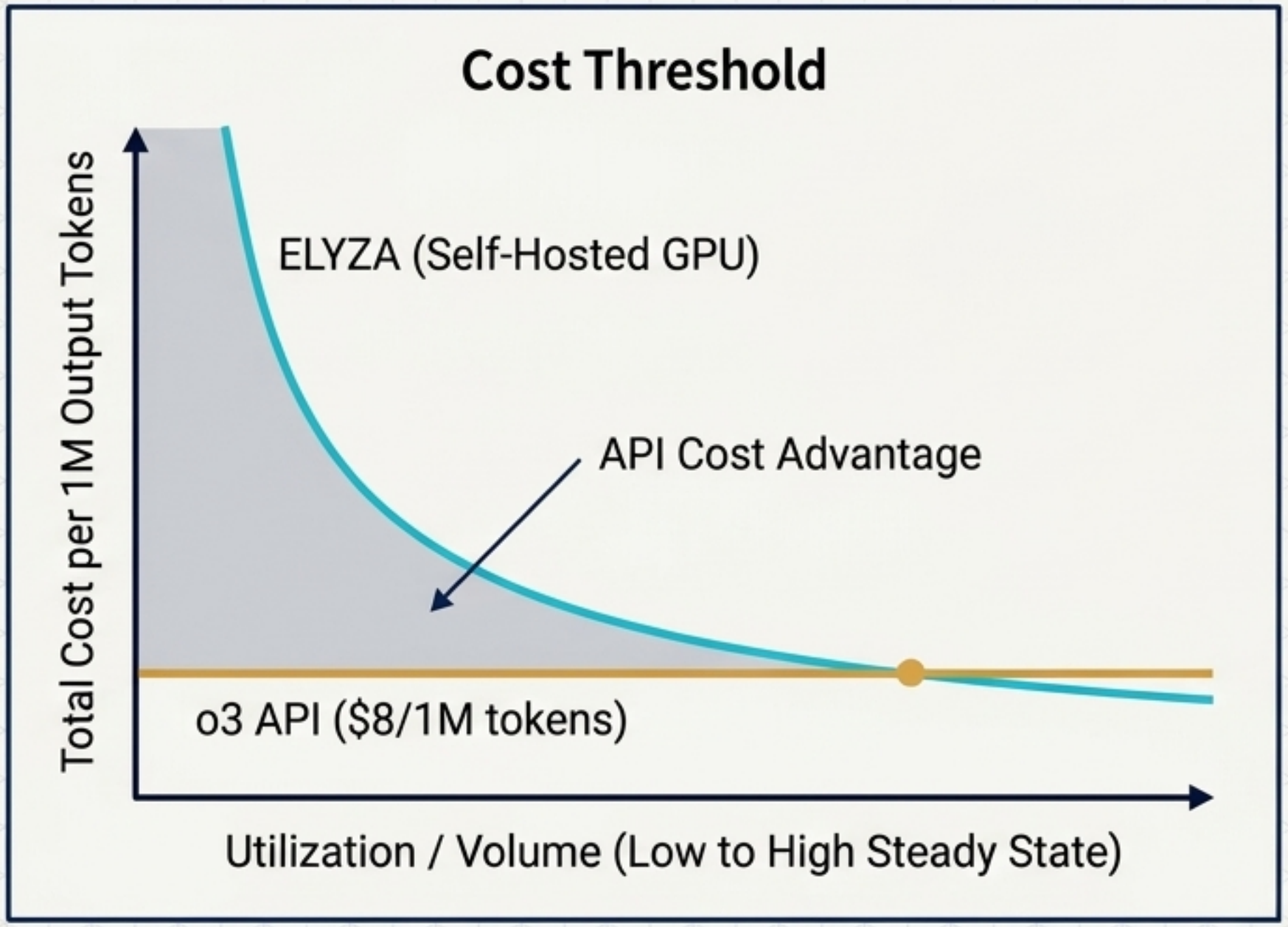
Apache 2.0による商用利用。業務に合わせた独自のファインチューニングやモデルバージョンのの固定が可能。

3. 日本語特化の制約追従

「日本語固有知識」や「複雑な形式制約」を伴うタスクにおいて、実運用上o1/o3と同等以上の使い勝手を発揮する可能性。

Gemini Notebook

コスト・ダイナミクス：オープンウェイトは「無料」ではない



APIの現実

OpenAI o3の出力単価は極めて安価（$8 / 1M tokens）。低〜中程度の稼働率では、外部APIの方が圧倒的にコスト優位。

GPUホスティングの計算式

$$\text{Cost} = \frac{(277.78 * \text{GPU台数} * \text{GPU時間単価})}{\text{スループット(tok/s)}}$$

例: 40 tok/sで1GPU稼働の場合、約6.94時間分のGPU費用が100万トークンの生成コストとなる。

戦略的判断: 高稼働率でGPUを回せる環境や、トークン単価よりも「データ通信量の制限（Rate limit）の自社制御」に価値を置く場合のみ、セルフホストの経済的合理性が成立する。

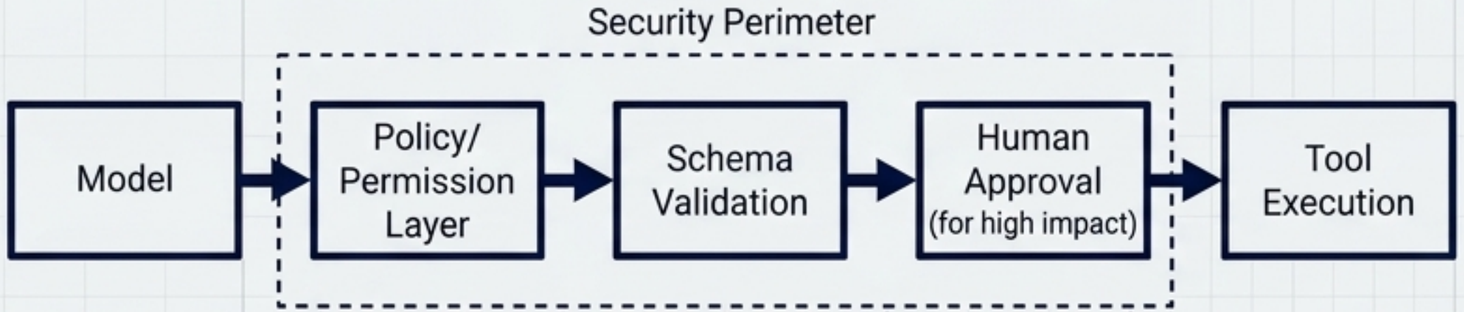
Gemini Notebook

セキュリティ、ツール呼び出し、およびガバナンス

Vulnerable Architecture



Recommended Architecture



1. ツール呼び出しが広げる攻撃面

エージェントとしてDBやShellに直接接続すれば、プロンプトインジェクションによる誤操作・権限昇格のリスクが急増する。ポリシー層とHuman-in-the-loopによる分離が必須。

2. サプライチェーン・リスク

推奨される `--trust-remote-code` は便利だが、リモートリポジトリ由来の実行コードを信頼境界に入れることを意味する。本番環境ではハッシュ固定とコードレビューが必要。

3. 推論ログ（Reasoning Trace）の取り扱い

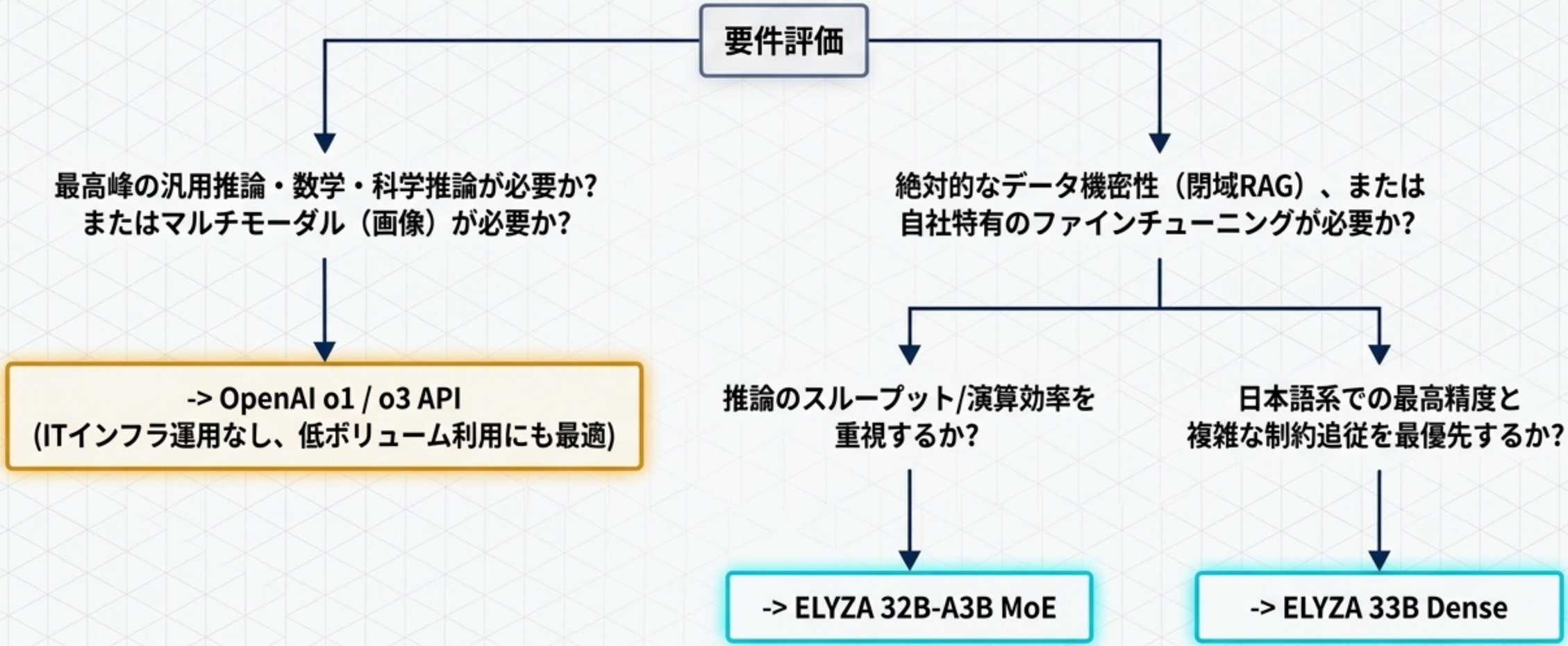
推論過程には機密情報が含まれる可能性がある。「最終回答」と「内部推論ログ」を分離し、ユーザー表示・ログ保管の権限を分けること。

4. Apache 2.0の限界

寛容なライセンスだが、「非侵害性」等の保証はない。出力内容に関する法務リスクは自社で管理する必要がある。

Gemini Notebook

エンタープライズ・アーキテクチャ 選択マトリクス



Conclusion: 日本語の独自要件と機密性が絡むプロジェクトにおいて、ELYZAは必ずPoC候補に入れるべき有力な選択肢である。

Gemini Notebook

アクションプラン：本番採用に向けた必須検証ステップ

ブラインド評価
(Blind Evaluation)
自社業務ベンチマークにはなく、公開同題を実施した公開ベンチマークに依存しない。

1

ブラインド評価
自社業務特有の日本語プロンプト50〜500件を用意し、o3とELYZAで完全同条件のブラインドテストを実施。公開ベンチマークに依存しない。

2

インフラ実機テスト
(Infrastructure Test)
想定ハードウェア（vLLM）上で、TTFT（Time‖スループット（tok/s）を実測。64Kコンテキスト長での性能劣化を確認。

レッドチーミング
(Red-Teaming)
日本語の有害コンテンツ生成、RAGポイズニング、ツール誤用のリスクに対する自社レッドチーム評価の実施。

3

レッドチーミング
(Red-Teaming)
日本語の有害コンテンツ生成、RAGポイズニング、ツール誤用のリスクに対する自社ﾚｯﾄﾞﾁｰﾑ評価の実施。

4

ガバナンス・ロック
(Governance Lock)
リモート実行コードのコミット/ハッシュ固定。推論ログの保存期間・アクセス権限のルール策定。法務部門とのポリシーすり合わせ。

Gemini Notebook