LAT: 35.6895° N, LONG: 139.6917° E

Y-AXIS: 9999.9999

CONFIDENTIAL
EXPIRATION: 2026-11-25

X-AXIS: 0000.0001

Qwen4-Max-Preview 徹底調査レポート

巨大スペックの裏に潜む実態と、本番適用の「罠」

Model_ID: qwen4-max-preview
Alias: qwen4.0-max-preview
Status: Active (Preview)

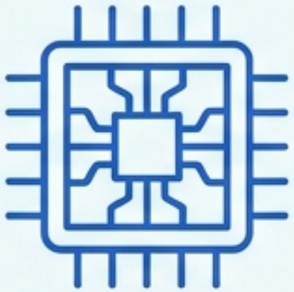
SYSTEM-REF: QWN-M-P-2024-11

GRID: 0.001MM

Gemini Notebook

Executive Summary: 評価者のための羅針盤

驚異のスペック



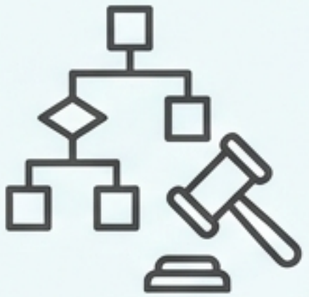
- 1Mトークンコンテキスト
- 393,216トークンの最大出力予算
- マルチモーダル（画像/動画）ネイティブ対応

致命的な制約



- 2026年11月25日で提供強制終了（Retire）
- 安定版への移行保証なし
- 自社アプリバックエンドからのAPI呼び出しは規約違反

評価者の結論



- 現時点でのプロダクト本番組み込みは不可
- 社内用の自律型Agent評価や研究用途としては、今すぐ試すべき極めて興味深い対象

Gemini Notebook

わずか14日間の生存期間：安定版への自動移行はない

11月11日 10:00 (UTC+8)

11月25日 24:00 (UTC+8)



Preview公開開始

Preview提供終了
(Retire)

※一般的な「同一IDのまま安定版へ昇格」するPreviewモデルではない。

公式ドキュメントに「PreviewモデルをRetireする」と明記。
評価期間終了後の動作保証は一切存在しない。

Gemini Notebook

確認済みの技術スペック (Technical Specifications)

Context Window
1,000,000 tokens

Max Completion Budget
393,216 tokens

Multimodal Input
Text / Image (最大2,048枚)
/ Video (最大2時間・2GB)

Reasoning
デフォルトON
(reasoning_effort=xhigh
が既定値)

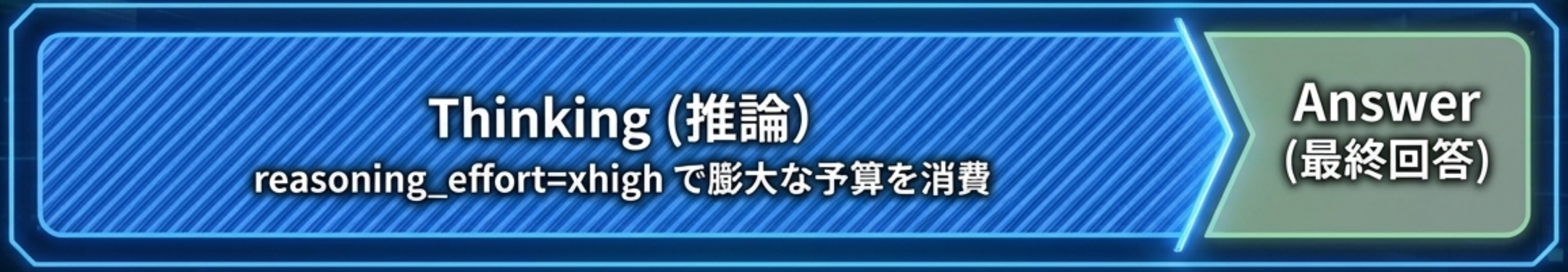
Ecosystem
Function Calling /
Built-in tools / Web Search

Audio & Generation
音声入力やネイティブ画像
生成は公式対応の確認なし

Gemini Notebook

Output Budgetの罠：「最大出力＝長い回答」ではない

総予算: 393,216 tokens (max_tokens)



残された枠が圧迫される

- ・推論（Thinking）と最終回答（Answer）は、同一の予算枠を共有する。
- ・膨大な推論プロセスが予算の大半を食いつぶし、最終的な回答本文が意図せず短くなるリスクが存在。

Gemini Notebook

ブラックボックス：確認済みの事実と未確認の推測

Confirmed (事実)

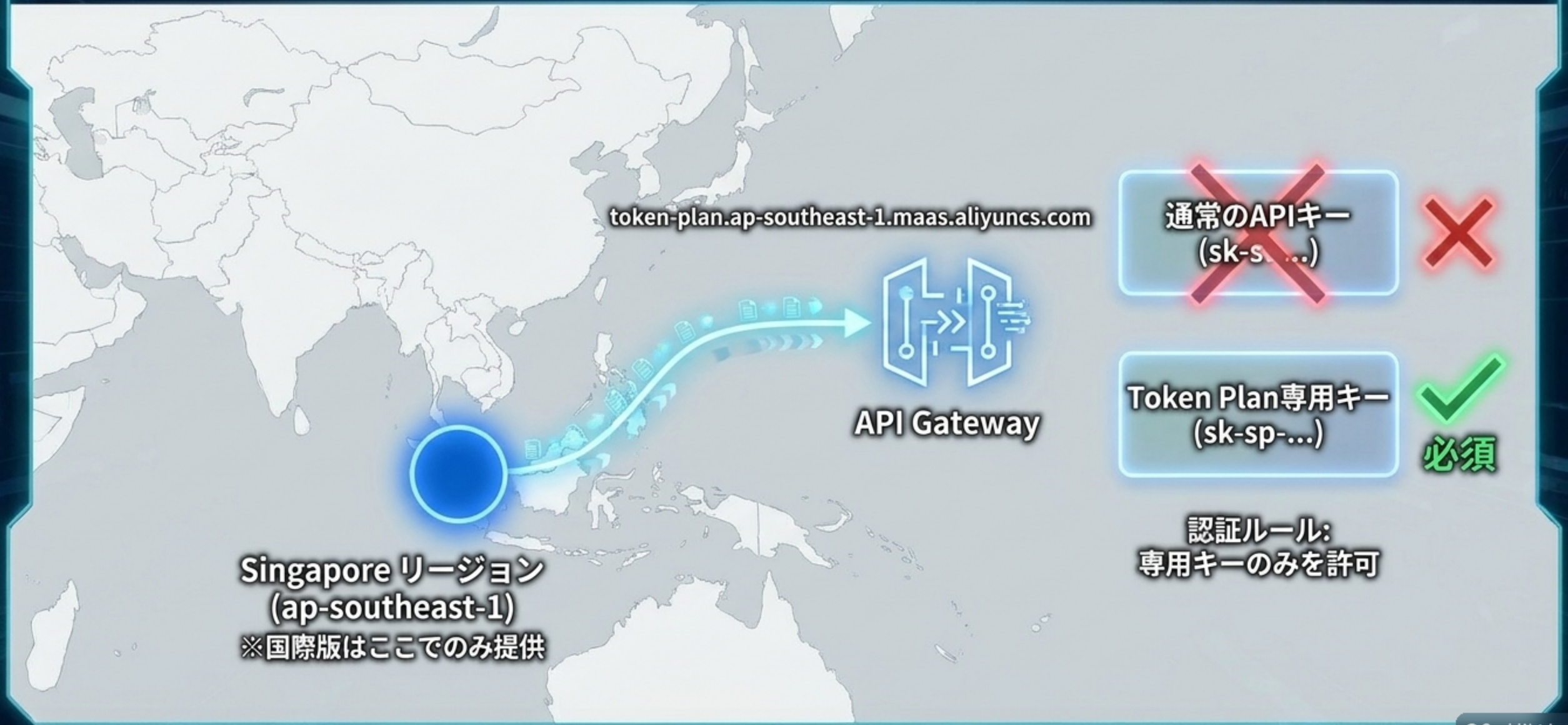
- Model ID: qwen4-max-preview
- Alias: qwen4.0-max-preview
- 提供元: Alibaba Cloud Model Studio

Unconfirmed / Unknown (未公表・未確認)

- パラメータ数: 前世代は2.4T MoEだが、Qwen4に同じ数字を流用する根拠はなし
- アーキテクチャ: 技術論文、モデルカード、オープン重みは非公開
- 正式版の発売日: 完全未定

Gemini Notebook

隔離された提供環境：アクセス網と認証



Gemini Notebook



許可 (Interactive Use ONLY)



Claude Code、Cursor、Qwen Code等の「対話的なCoding / Agentツール」内でのユーザー操作による利用。

禁止 (Automated & Backend FORBIDDEN)



- 自社SaaSやアプリのバックエンドからのAPI呼び出し
- Python等による独自スクリプトからの直接実行
- 非対話的な自動バッチ処理

結論：「APIエンドポイントが存在する＝自由にプロダクトへ組み込める」わけではない。

Gemini Notebook

Preview特有の経済学：動的なCredits消費



- **従量課金の不在**: Token Planに「100万トークンあたり〇ドル」という公開価格は存在しない。
- **動的変動**: リクエストごとのCredits消費量は一定ではない。

Preview特別レート:
オフピーク時(22:00–08:00 UTC+8)は消費量が通常の8%に割引。

Gemini Notebook

公式データなし：第三者初期測定（StealthMark）による参考値

Performance Reality: ベンチマークの現在地



分析：独自指標のため総合性能を決定づけるものではない。ただし「公開直後から既存モデルを圧倒している」わけではない現実を冷静に認識すべきである。

Gemini Notebook

競合製品ランドスケープ（Competitive Matrix）

モデル名	Context	Max Output	価格形態	本番バックエンド利用
Qwen4-Max-Preview	1M	393,216	Token Planのみ	不可
Qwen3.8-Max	1M	128K	$2(in)/$6(out)	可
DeepSeek V4.1	1M	384K	$0.30(in)/$1.20(out)	可
GPT-6 Astra	1.05M	128K	$10(in)/$50(out)	可
Gemini 4 Argon	約1M	Decode対応	11月末まで$2/$10	可

Gemini Notebook

データ保護とプライバシー基準



学習への利用

原則として、プロンプトと回答はモデルの学習（Training / Fine-tuning）には使用されない。

保持期間

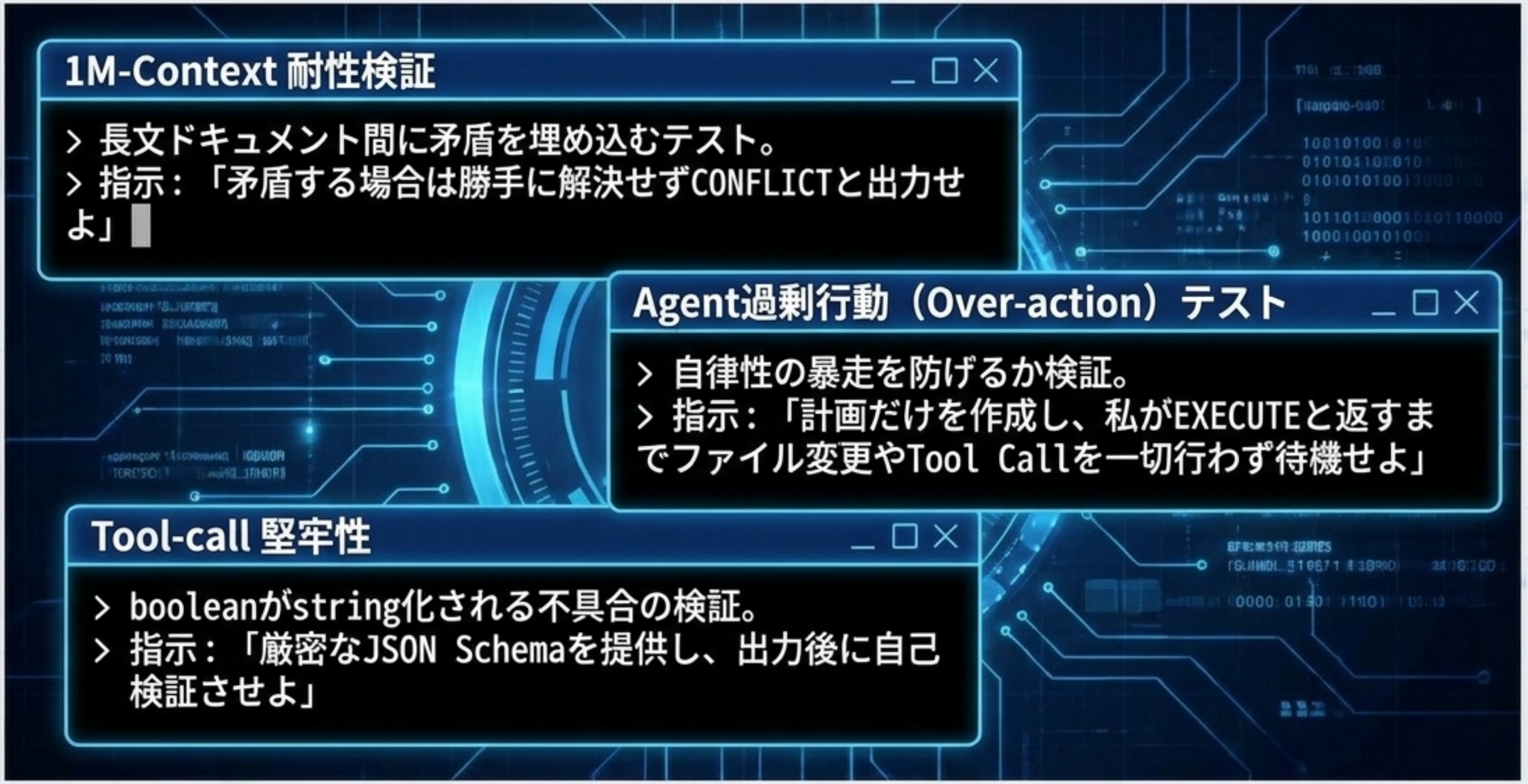
通常、データは最大30日間保持される。

? ZDR（Zero Data Retention）の不確実性

エンタープライズ向けの完全非保持機能（ZDR）に、Qwen4-Max-Previewが対象として含まれているかは現時点で公式確認不可。機密データを投入する際は事前の書面確認が必須。

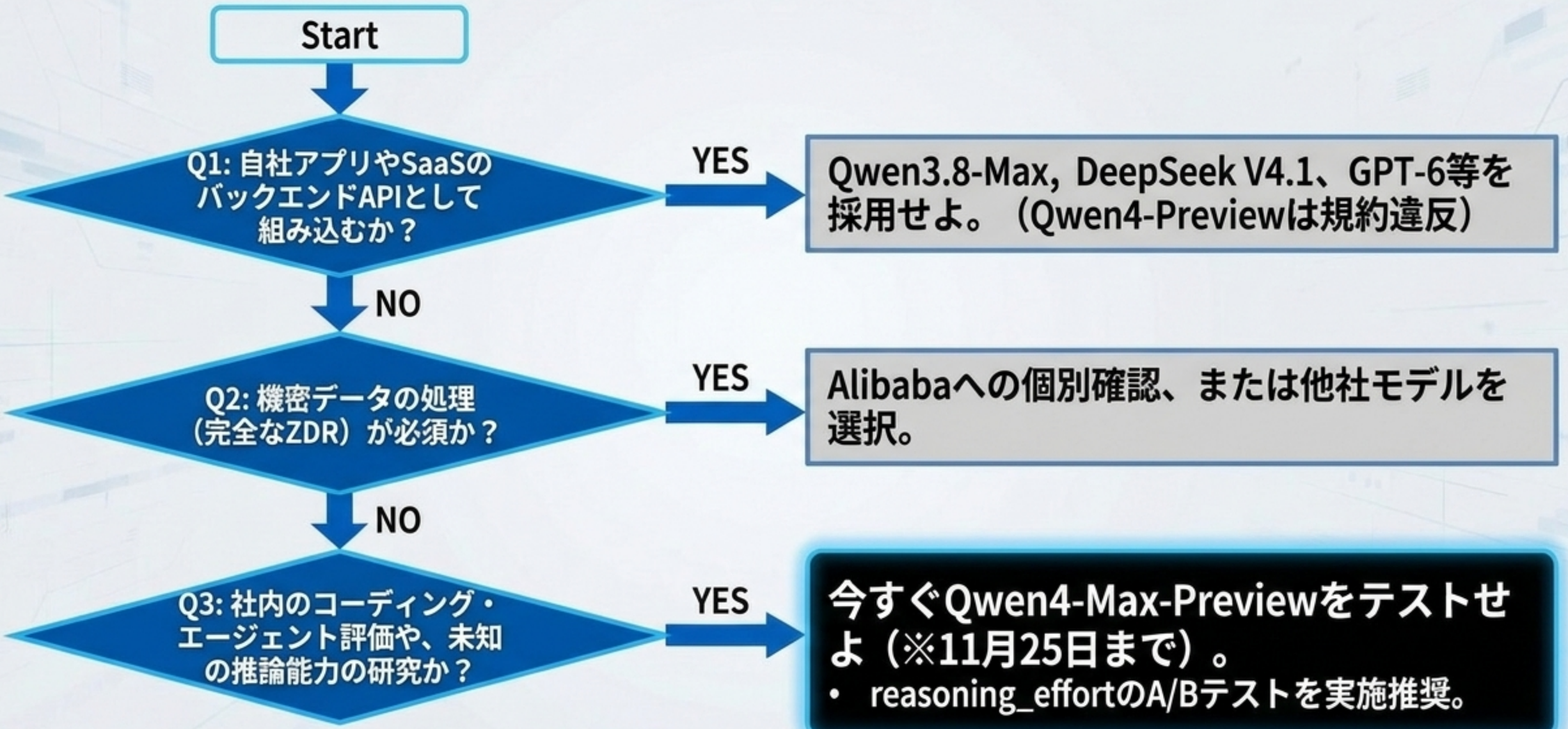
Gemini Notebook

評価者のための検証プレイブック（Prompt Strategies）



Gemini Notebook

採用判断フローチャート（The Decision Tree）



Gemini Notebook

結論とNext Steps（Final Assessment）



1. 本番環境への依存は厳禁

短い生存期間（11/25 Retire）とAPI利用規約の壁により、プロダクトのインフラにはなり得ない。



2. 規格外のスペックを「今」体験せよ

393Kの出力予算と1Mコンテキストの組み合わせがもたらすAgenticな挙動は、将来のLLMのパラダイムを予見させる。



3. 安定版（Stable）を待て

評価を済ませた後は、アーキテクチャや価格が透明化された「正式なQwen4-Max」のリリースを待つのが最良の戦略。

END OF REPORT

Gemini Notebook