Tech News Daily

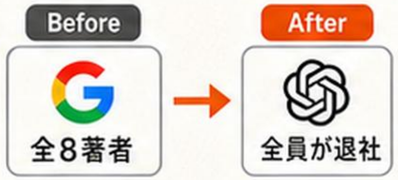
2026-07-05

MORNING DISPATCH / Vibe Coder Bootcamp Tech News

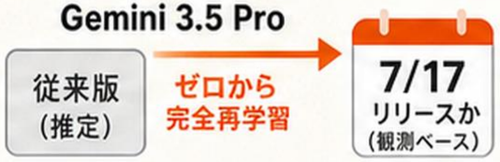
今朝のホットな話題

2026-07-05 — Vibe Coder Bootcamp Tech News

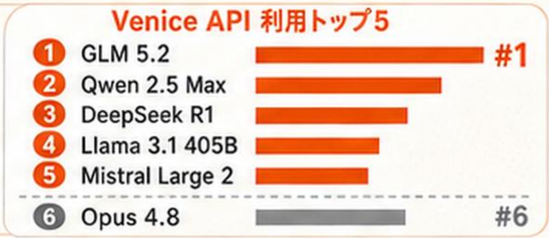
1. 「Attention Is All You Need」全8著者が Google を去った — Noam Shazeer の OpenAI 移籍で“最後の一人”も退社



2. Gemini 3.5 Pro、“ゼロから完全再学習”版を 7/17 にリリースか（観測ベース）



3. Venice API の利用トップ5をオープンソースモデルが独占 — GLM 5.2 が #1、Opus 4.8 は #6



5トピックを整理。

TEKION Group

「Attention Is All You Need」全8著者が Google を去った ― Noam Shazeer の OpenAI 移籍で"最後の一人"も退社

何が起きた？

2017年に現代AIの起点となった論文「Attention Is All You Need」（Transformer を提唱）の著者8名全員が、今週までに Google を退社した。直近の引き金は Noam Shazeer の OpenAI 移籍（6/18に本人発表）と、John Jumper の Google DeepMind 離脱・Anthropic 移籍。

主な変更点

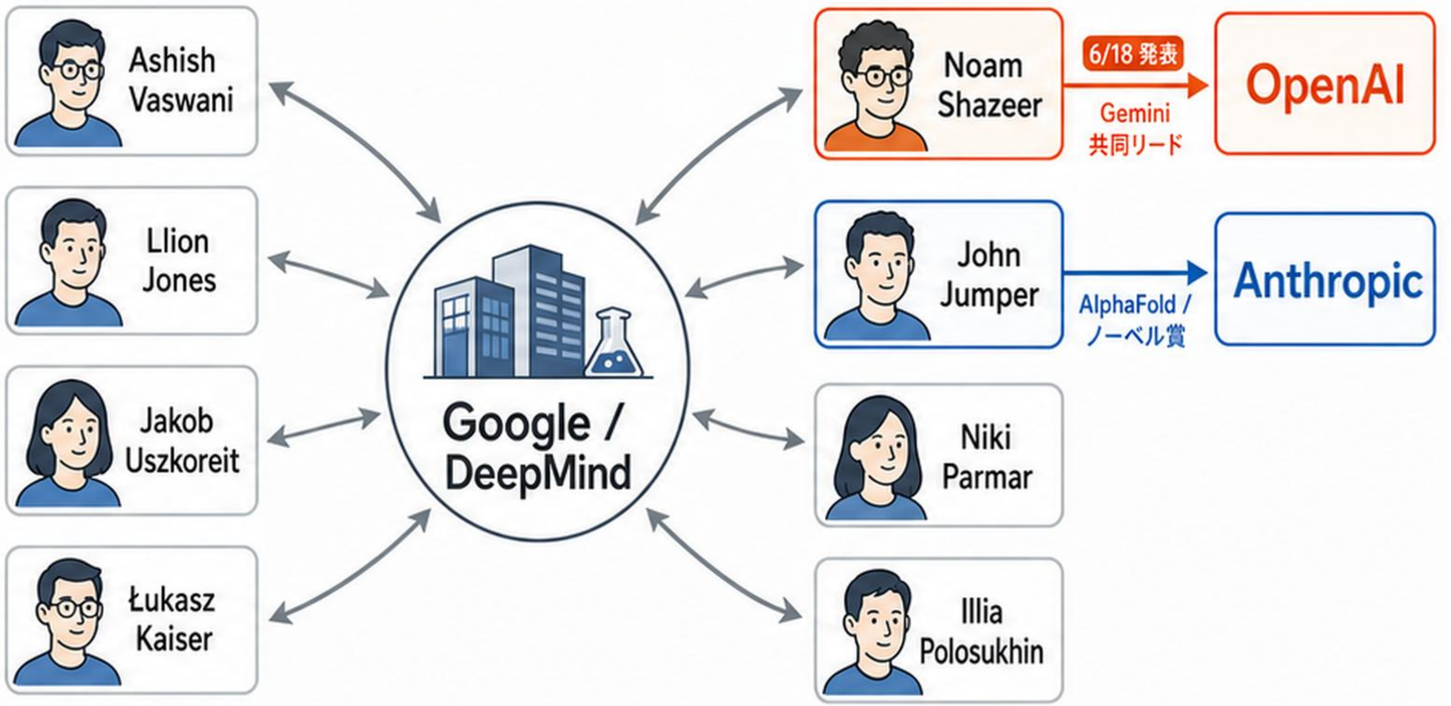
- Transformer 論文の全8著者が、今週の Shazeer・Jumper 退社をもって全員 Google を離れた
- Noam Shazeer: Transformer 共著者・VP Engineering・Gemini 共同リード。6/18 に OpenAI 移籍を発表（役割詳細は未開示）
- Google は2024年に Shazeer とチームを約2,200億円で買い戻した経緯があり、わずか2年で再流出
- John Jumper: DeepMind VP／engineering fellow、AlphaFold でノーベル賞。Anthropic へ移籍
- 頭脳が OpenAI / Anthropic の2極に集中していく構図が可視化

なぜ重要？

現代AIの基礎設計図を書いた研究者たちが、その発祥企業から一人残らず去った象徴的なマイルストーン。モデルの優劣が人材フローと連動する業界力学を読む題材として、モデル選択をする開発者にも無関係ではない。

Xでの反応：「Gemini 3.5 Pro の遅延の理由が分かった」／「Transformer の父が Google を去った」

Transformer 全著者退社 → OpenAI / Anthropic への集中



8	2017	2024	約2,200億円	2年	6/18
著者	論文発表年	買い戻し	買い戻し金額	で再流出	Shazeer 移籍発表

Gemini 3.5 Pro、"ゼロから完全再学習"版を 7/17 にリリースか（観測ベース）

何が起きた？

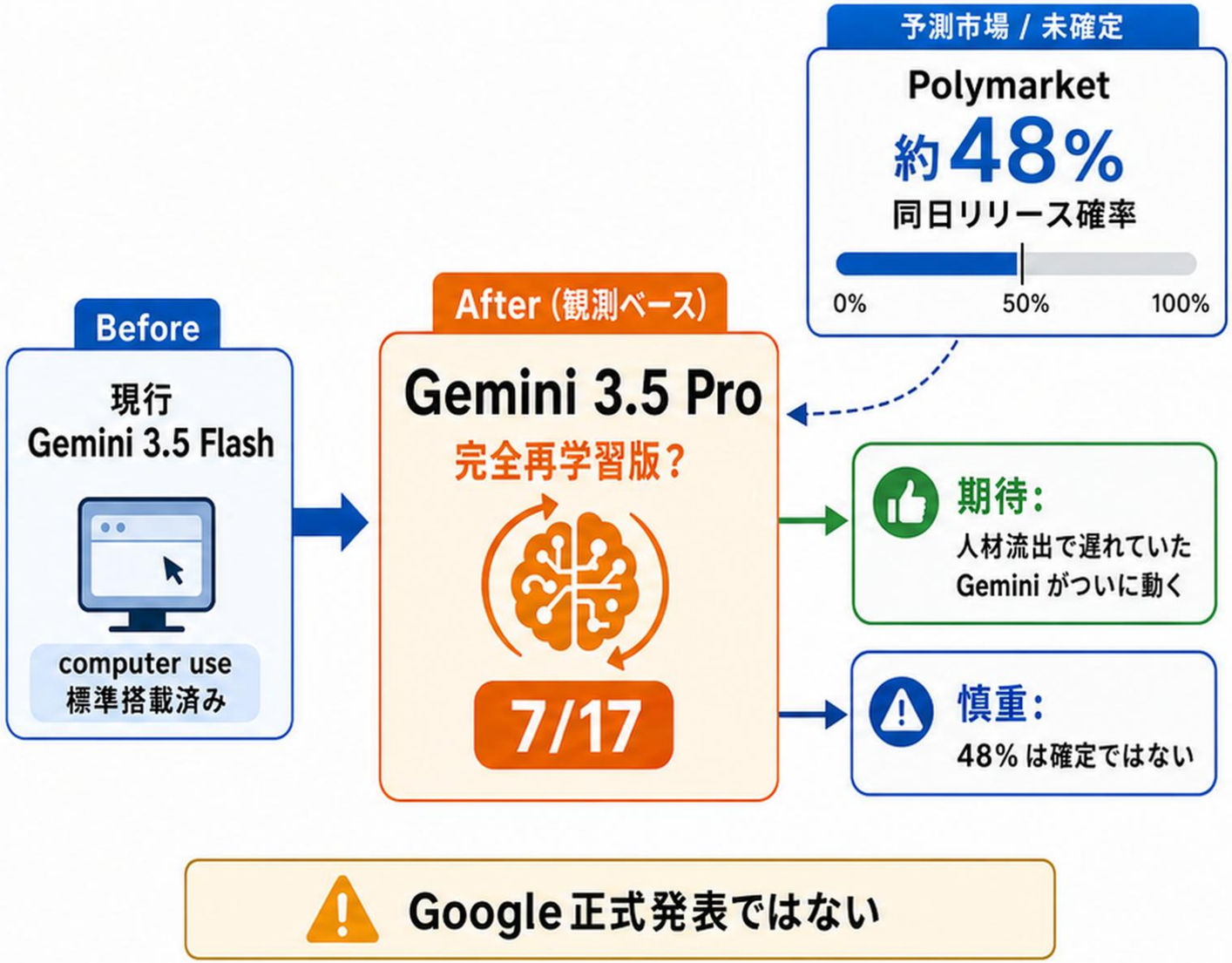
Google が Gemini 3.5 Proを「スクラッチから完全に再学習した」版として 7/17 にリリースするという観測が浮上。予測市場 Polymarket では同日リリース確率を約48%と提示。Google からの正式発表ではない。

主な変更点

- Gemini 3.5 Pro を「完全に再学習（retrained from scratch）」した版として 7/17 リリースの観測
- Polymarket が同日リリースに約48%の確率を提示（確定情報ではなく予測市場ベース）
- 現行 Gemini 3.5 Flash には既に computer use が標準搭載済みで、Pro 側の刷新が次の焦点
- 人材流出でモデル開発が遅れているという見方と対で、Google の巻き返しタイミングとして注目

なぜ重要？

フロンティア競争で Google が「作り直し」に踏み切ったとすれば、Gemini 系の性能ジャンプがありうる。使うモデルの選択肢に影響するため、日付と「再学習」という設計判断は要ウォッチ。
ただし現時点は予測市場の観測であり、確報が出るまでは話半分で扱うべき。



Tech News Daily

2026-07-05

🔍 Venice API の利用トップ5をオープンソースモデルが独占 ― GLM 5.2 が #1、Opus 4.8 は #6

X / Erik Voorhees likes ❤

🔍 何が起きた？

Venice.ai 創業者の Erik Voorhees が、Venice API 上の利用モデル上位5つが全てオープンソースで、GLM 5.2 が #1、クローズドの Opus 4.8 は #6 だと報告。オープンウェイトが実運用トラフィックで商用フロンティアを上回る、コスト最適化側の実データ。

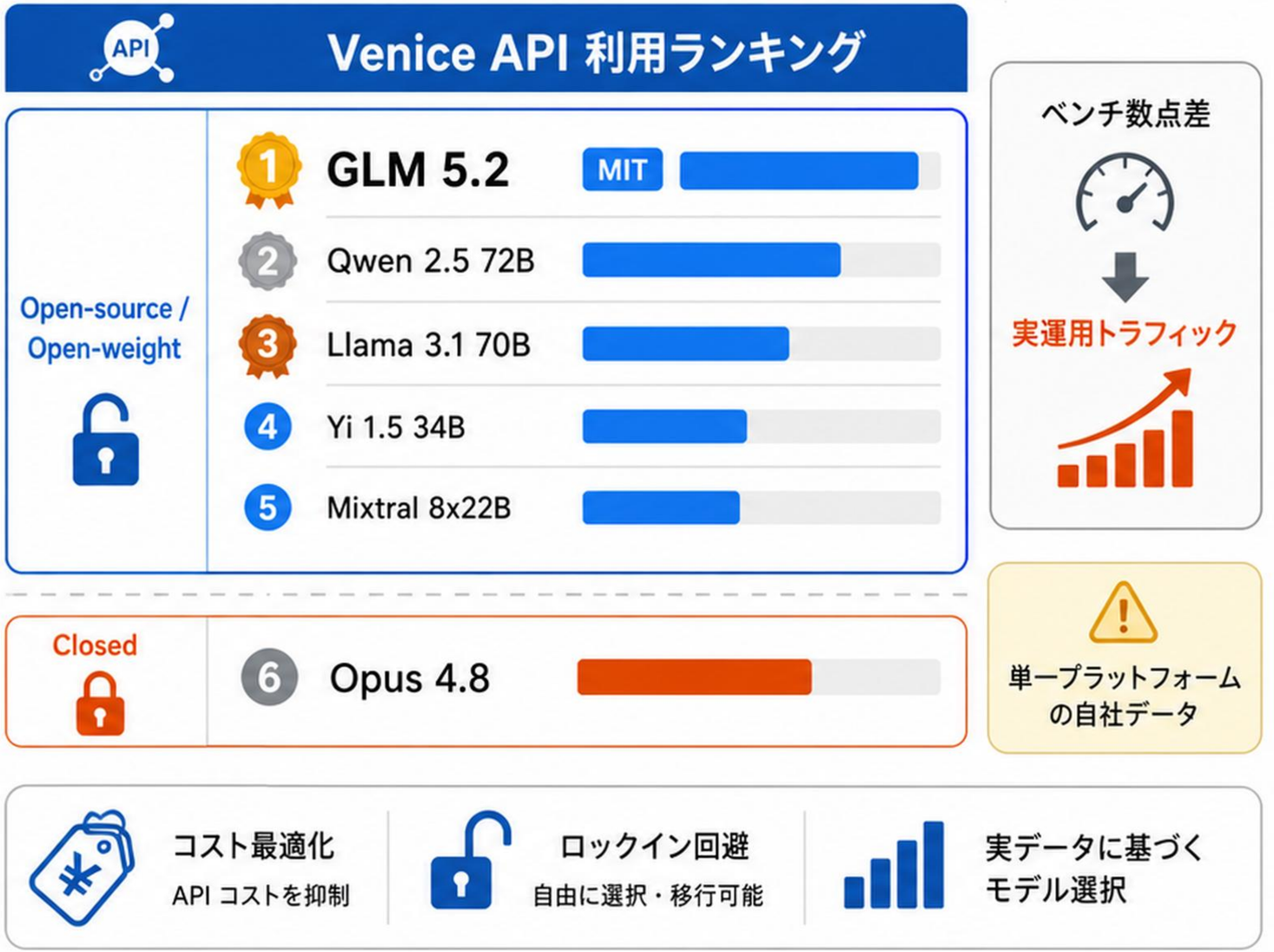
📌 主なポイント

- Venice API の利用モデル上位5つが全てオープンソース、GLM 5.2 が #1
- クローズドの Opus 4.8 は #6 に位置
- GLM 5.2（MIT ライセンス）の実運用採用が伸びている追加データ点
- ※ Venice という単一プラットフォームの自社データである点は留意

💡 なぜ重要？

「ベンチで数点差」だったオープンウェイトが、実際の API トラフィックでクローズドを押しのけて上位を独占する運用側の証拠。ロックイン回避とコスト最適化を重視するなら、モデル選択の実データとして有用。

Xでの反応: オープンウェイト推進派は「コスパで選べば当然」と歓迎。一方で「Venice のユーザー層が価格重視に偏っている」という限定付きで読む声も。



x.com/ErikVoorhees/...

TEKION Group

AEON vLLM Ultimate v0.24.0 を単体 DGX Spark で実測 — 公称 536 tok/s は再現できずも Gemma-4-26B が 578 tok/s@C32 で実用最速

何が起きた？

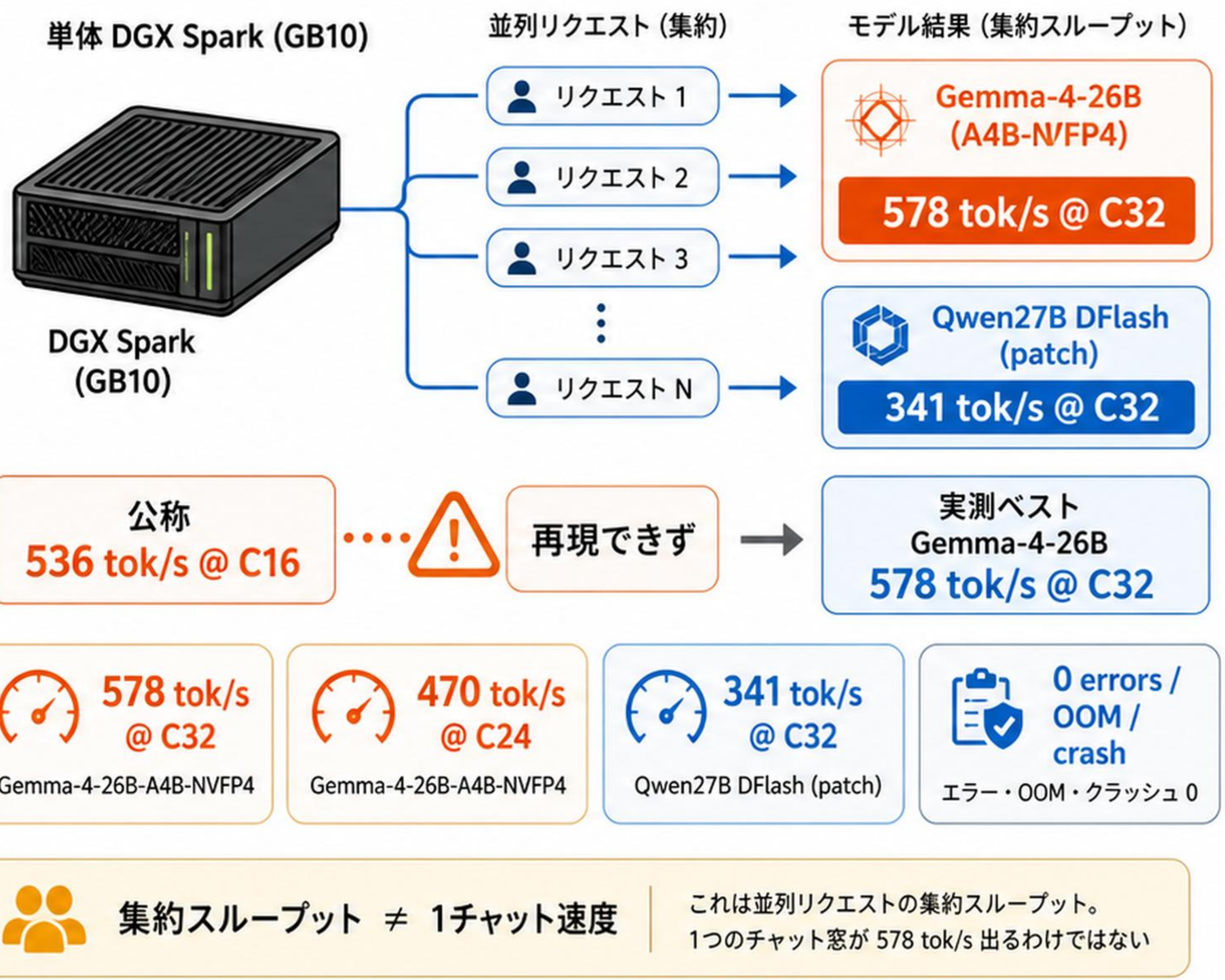
前日配信の AEON vLLM Ultimate v0.24.0（リリース告知）に対する第三者の実測追試。単体の DGX Spark (GB10) で実際に回し、公称の 536 tok/s @ C16 は再現できなかったが、Gemma-4-26B なら1台で 578 tok/s @ C32 まで安定して出せると結論づけた。

主な検証結果

- Gemma-4-26B-A4B-NVFP4（262K コンテキスト有効 / FP8 KV）：C24 で 470 tok/s、C32 で 578 tok/s、エラー・OOM・クラッシュ 0
- これは並列リクエストの集約スループット。1つのチャット窓が 578 tok/s 出るわけではない
- Qwen27B DFlash も config を自力 patch して 262K で安定動作（C32 で 341 tok/s）

なぜ重要？

公称の 536 tok/s @ C16 は再現できず。ただし1台の Spark で 578 tok/s @ C32 は達成。高スループットの実用的勝者は Gemma 26B-A4B、公称ヘッドラインは構成依存。ローカルエージェント・バッチ・マルチセッション API 向けの数字。



Tech News Daily 2026-07-05

Anthropic / X

Likes 6,435 Bookmarks 14,627

A Field Guide to Fable: Finding Your Unknowns — Fable と働くコツは「自分の unknowns を見つけること」（Thariq / Anthropic）

要点

Anthropic の Thariq による Claude Fable 5 協働ガイド。核となる比喩は「地図は領土ではない」。プロンプト・スキル・コンテキスト（地図）と、コードベースや現実（領土）の差分が unknowns（未知）。Fable は作業品質が「自分が unknowns を明確化できる能力」でボトルネックになる初めてのモデルだと述べる。

具体的な手法 / 使いどころ

- unknowns を4象限で分解：Known Knowns / Known Unknowns / Unknown Knowns（見れば分かる）/ Unknown Unknowns（考慮すらしていない）
- 実装前：Blind Spot Pass（盲点を洗い出させる）／プレスト＆プロトタイプ（HTMLで複数デザイン案）／インタビュー（1問ずつ、アーキテクチャが変わる質問優先）／リファレンスはソースコードが最良
- 実装計画は「変わりやすい決定（データモデル・型・UXフロー）」を先頭、機械的リファクタは末尾に
- 実装中：implementation-notes.md に逸脱を記録／実装後：説明資料で buy-in、マージ前にクイズで自分を試す
- Fable ローンチ動画は全編 Claude Code で編集。Whisper・Remotion・カラーグレーディングの unknowns を Claude に教わりながら埋めた

なぜ刺さるか / 学び

モデルが賢くなるほど、勝負は「指示の精度」ではなく「自分が何を分かっていないかを明らかにする力」に移る。blindspot pass・1問ずつインタビュー・マージ前クイズは日々のワークフローに落とせる実践プロンプト。

unknowns の4象限

Known Knowns 自分も知っていて、領土にも存在する	**Known Unknowns** 自分はは知ないが、領土には存在する
Unknown Knowns（見れば分かる） 自分は気づいていないが、見れば分かる	**Unknown Unknowns（考慮すらしていない）** 自分も気づいておらず、領土に存在する未知

> “プロンプト術の次はこれ
> unknowns の4象限が刺さる”

実装前 → 実装中 → 実装後

実装前：Blind Spot Pass
盲点を洗い出す
プレスト & プロトタイプ
1問ずつインタビュー
リファレンスはソースコード

実装中：implementation-notes.md
逸脱・判断・決定を記録
変わりやすい決定を優先

実装後：quiz / buy-in
説明資料で buy-in
マージ前にクイズで自分を試す

Likes 6,435

Bookmarks 14,627

x.com/...

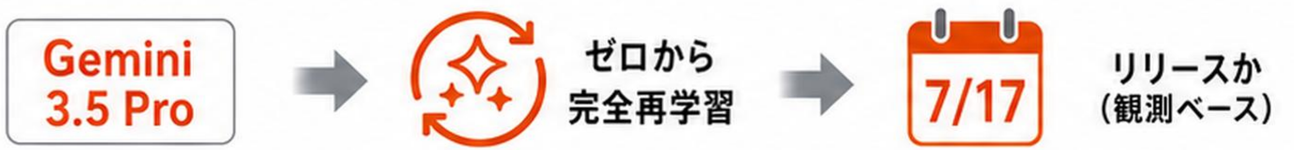
TEKION Group

今日のまとめ

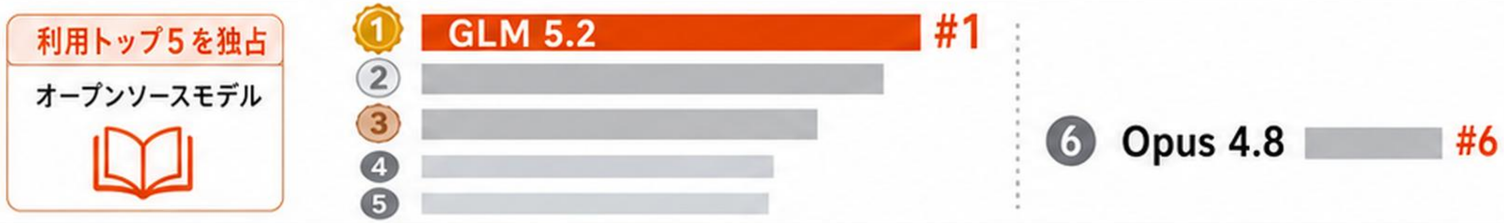
本日のトピック一覧

1 『Attention Is All You Need』全8著者が Google を去った — Noam Shazeer の OpenAI 移籍で“最後の一人”も退社

2 Gemini 3.5 Pro、“ゼロから”完全再学習”版を 7/17 にリリースか（観測ベース）



3 Venice API の利用トップ5をオープンソースモデルが独占 — GLM 5.2 が #1、Opus 4.8 は #6



4 AEON vLLM Ultimate v0.24.0 を単体 DGX Spark で実測 — 公称 536 tok/s は再現できずも Gemma-4-26B が 578 tok/s@C32 で実用最速

5 A Field Guide to Fable: Finding Your Unknowns — Fable と働くコツは『自分の unknowns を見つけること』（Thariq / Anthropic）

OpenAI | Anthropic | Google | Venice API | AEON vLLM