Tech News Daily

2026-07-04

MORNING DISPATCH / Vibe Coder Bootcamp Tech News

今朝のホットな話題

2026-07-04 — Vibe Coder Bootcamp Tech News

1. Kimi K2.7 Code が GitHub Copilot で一般提供 (GA) — Copilot のモデルピッカーに初のオープンウェイトモデル

2. Apple が Safari に MCP サーバをネイティブ搭載 (Safari Technology Preview 247) — ブラウザ操作の AI 連携が Chrome 独占でなくなる

3. Alibaba「SkillWeaver」— 2,209 個のツールでも溺れないエージェント framework、タスク分解＋必要ツールだけ取得でトークン使用を99%削減

7トピックを整理。

TEKION Group

Tech News Daily
2026-07-04

Kimi K2.7 Code が GitHub Copilot で一般提供（GA）— Copilot のモデルピッカーに初のオープンウェイトモデル

GitHub公式
♡ 1.2K

何が起きた？

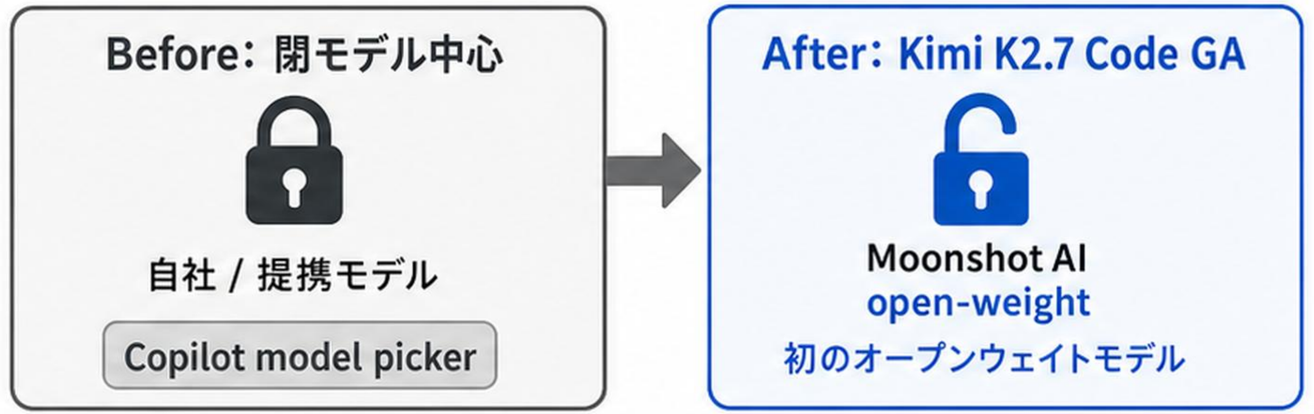
GitHub が中国 Moonshot AI のオープンウェイトモデル「Kimi K2.7 Code」を GitHub Copilot で一般提供（GA）にした。Copilot のモデルピッカーで選択できる初のオープンウェイトモデルで、閉モデル中心だった選択肢に「自由に落とせるオープンモデル」が正式に加わった。

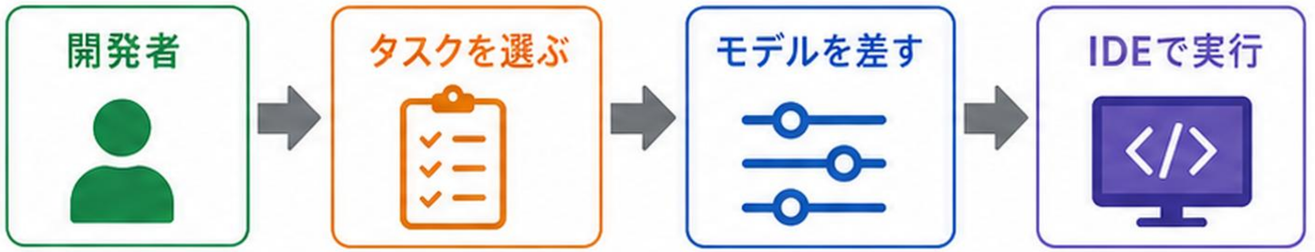
主な変更点

- Copilot のモデルピッカーで選択可能な初のオープンウェイトモデル
- Pro / Pro+ / Max プランで順次ロールアウト
- VS Code / JetBrains / Xcode / Eclipse など主要 IDE の Copilot から利用可
- Kimi K2.7 Code はリポジトリ規模のコーディングに焦点を当てた Moonshot AI の open-weight モデル（K2.7 系）

なぜ重要？

「どのモデルが総合1位か」より「どのタスクにどのモデルを差すか」というモデル選択の実務段階に入ったことを象徴。Xでは「このドアが開いたことが単一モデルの勝敗より重要」「総合1位ではなくタスク別最適の時代」という受け止めが目立った。





対応 IDE: VS Code / JetBrains / Xcode / Eclipse

- このドアが開いたことが単一モデルの勝敗より重要
- 総合1位ではなくタスク別最適の時代

github.blog / github.com

Apple が Safari に MCP サーバをネイティブ搭載（Safari Technology Preview 247）— ブラウザ操作の AI 連携が Chrome 独占でなくなる

何が起きた？

Apple の WebKit チームが Safari Technology Preview 247 に
Model Context Protocol (MCP) サーバをネイティブ搭載。
サードパーティのブリッジや拡張ではなく、Safari 本体の機能として
Web 開発・デバッグのワークフローを AI エージェントから操作できる。
Mac 上でローカル動作し、既存の Safari セッション/クッキーを引き継ぐ。

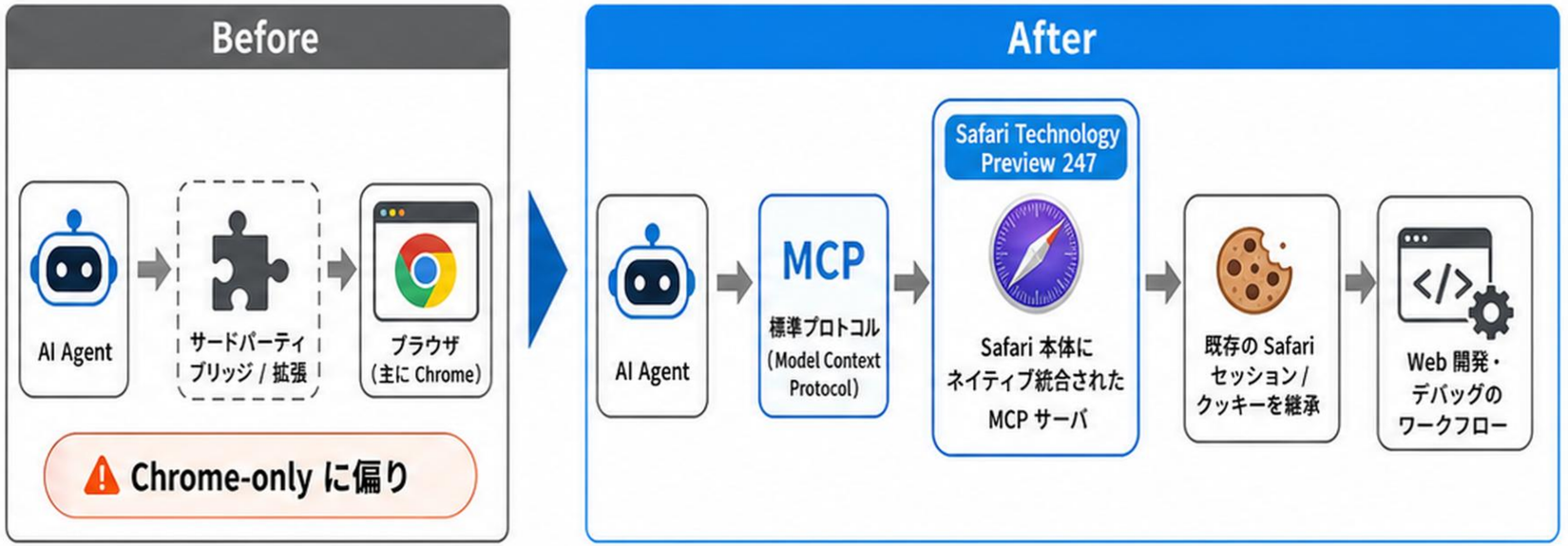
主な変更点

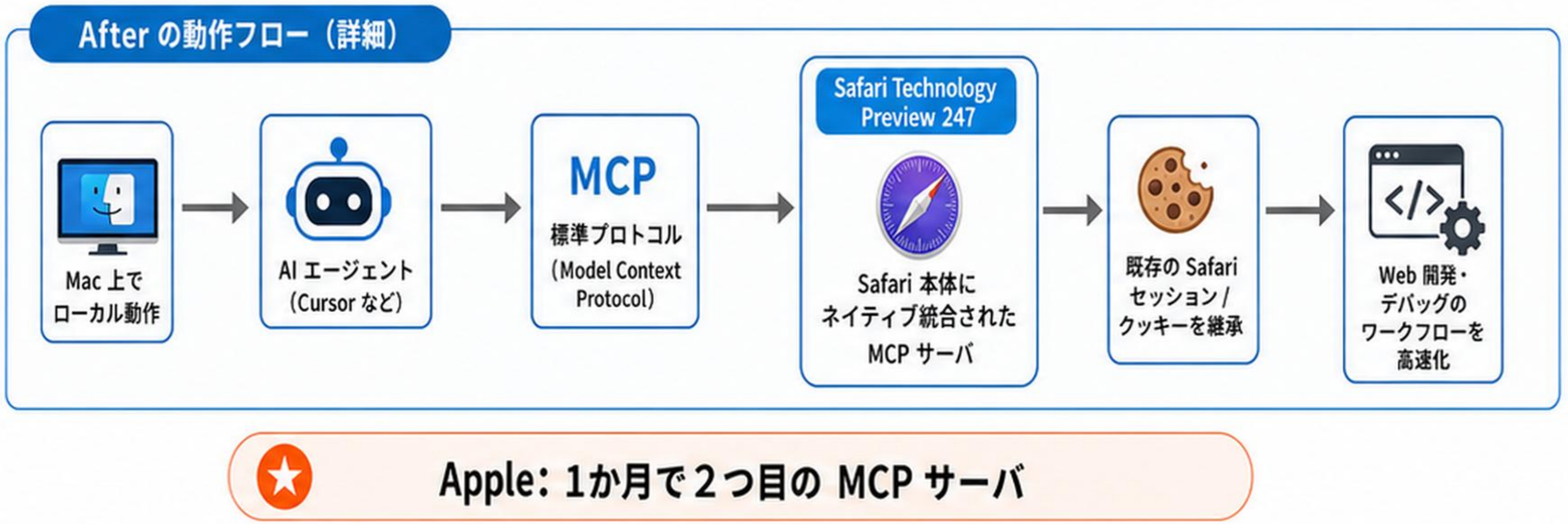
- Safari MCP サーバは Safari Technology Preview 247 で提供
- Mac 上でローカル実行し、既存の Safari セッション/クッキーを継承
- サードパーティ製ブリッジ・プラグインではなく、ブラウザ本体にネイティブ統合
- フロントエンド開発・デバッグのワークフローを AI エージェントから高速化（WebKit 公式ブログ）

なぜ重要？

Apple にとって1か月で2つ目の MCP サーバ出荷とされ、
MCP が「標準プロトコル」化しつつあることを示す。
ブラウザレベルのエージェントアクセスが Chrome 独占でなくなる。

Xでの反応: Apple が本当に MCP を載せたのは大きい、
chrome-only でないブラウザレベルのアクセスがついに来た





Apple: 1か月で2つ目の MCP サーバ

Alibaba「SkillWeaver」— 2,209 個のツールでも溺れない

エージェント framework、タスク分解＋必要ツールだけ取得でトークン使用を99%削減

Alibaba研究

♡ likes

何が起きた?

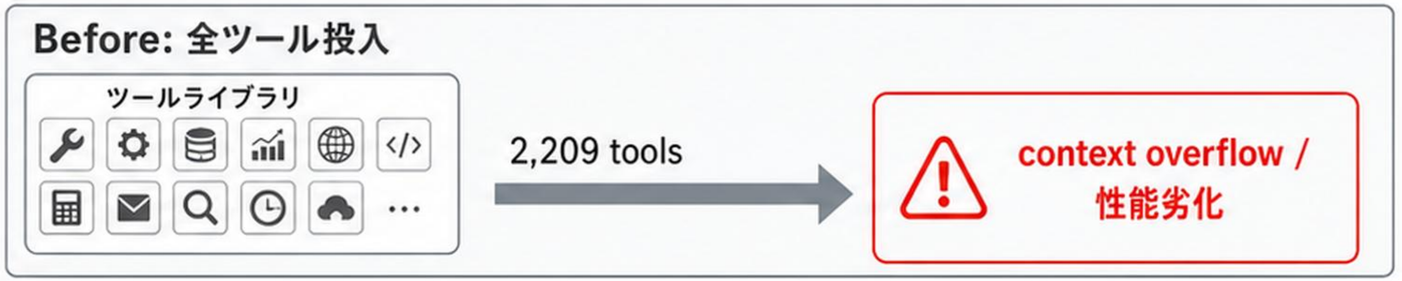
Alibaba の研究者が、大量のツールを抱えるエージェントの「コンテキスト溺れ」を解く framework「SkillWeaver」を発表。全ツールを毎回ロードせず、複雑なタスクをサブタスクに分解し、合致するツールだけを取得して実行可能なワークフローに合成する。

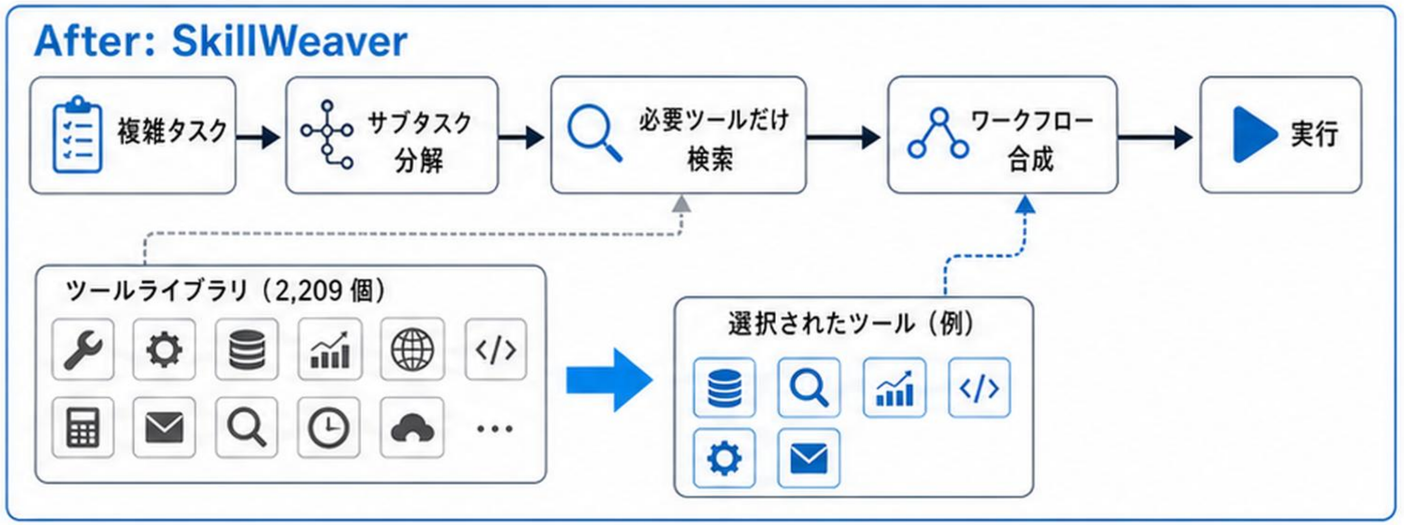
主な変更点

- 課題: エージェントに 2,209 個ものツールを渡すとコンテキストが溢れ、性能が劣化
- compositional skill routing（合成的スキルルーティング）でタスクを分解
- 必要なツールだけをライブラリから検索して合成
- 全ツールを毎回コンテキストに載せる方式に対し、トークン使用量を最大99%削減

なぜ重要?

ツールが増えるほど効く設計。大規模ツール群を持つ実務エージェントのコストと精度の両面に効く。





2,209 個
ツール規模

最大 99 %
トークン削減

source: Alibaba research

TEKION Group

Fable は 7/7 にサブスクから一旦外れるが、容量が許せば標準プランに復帰予定 — Anthropic の Thariq が説明

何が起きた？

Anthropic の Thariq が、Fable のサブスク提供について寄せられた多数の質問に回答。Fable は 2026-07-07 以降サブスクリプションから外れる。

主な変更点

- Fable は 2026-07-07 以降、サブスクリプションプランから外れる
- 容量が確保でき次第、サブスクの標準的な一部として復帰を目指す
- 当初ブログ記事どおりの公式スタンス

なぜ重要？

- 137万ビュー・9千いいねと反響が大きい
- Fable の需要と容量制約の綱引きが注目されている
- 高需要フロンティアモデルはローンチ直後、容量制約で提供形態が流動的

提供形態の流れ

137万ビュー

9千いいね

コミュニティの反応（抜粋）

“2週間で剥奪されるなら学習投資を削ぐ”

VS

“容量次第の復帰予定”

要点

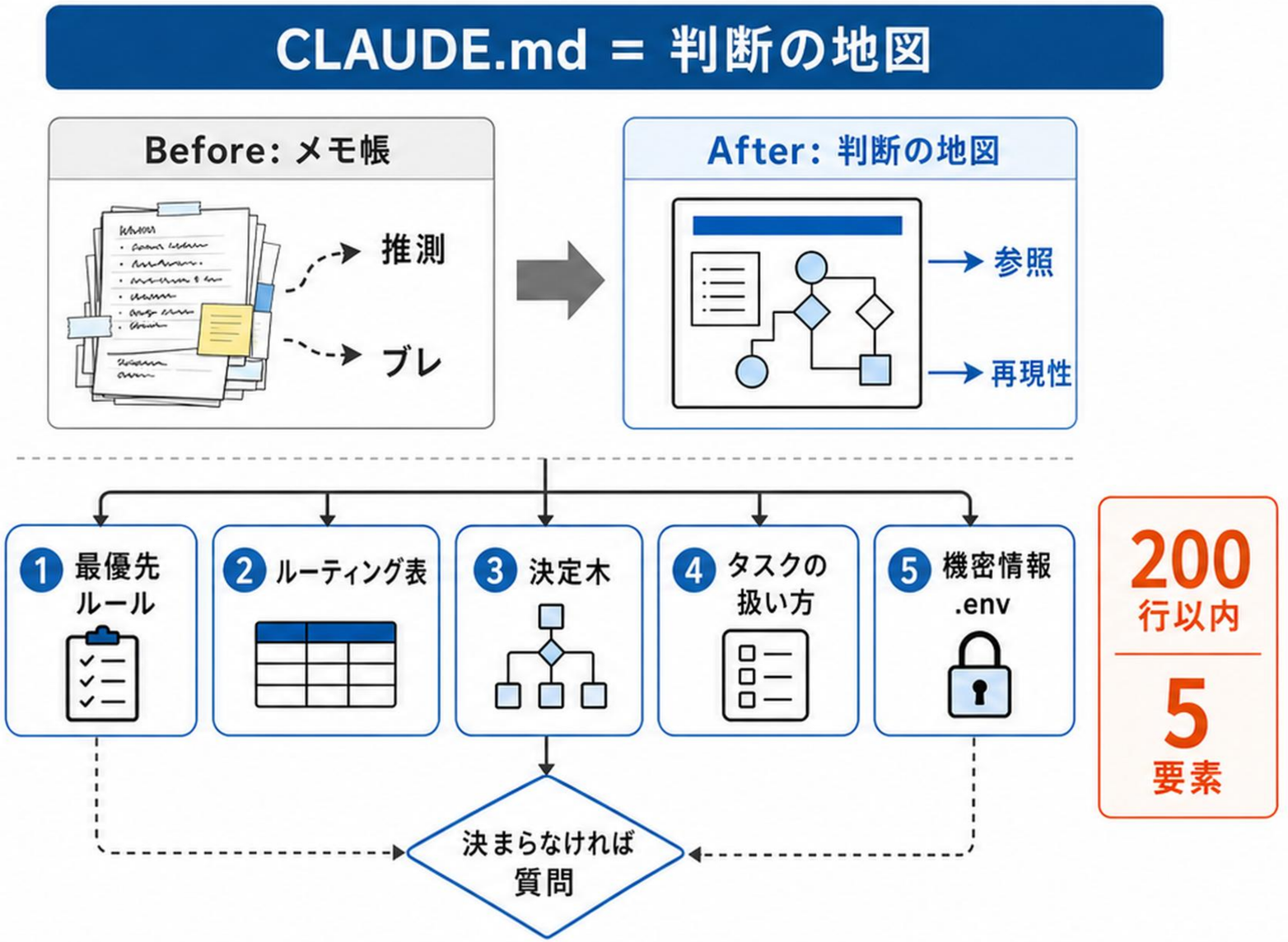
Claude Code が作業前に必ず読み込む CLAUDE.md を、思いつきを継ぎ足す「メモ帳」ではなく、AI が毎回参照する「判断の地図」として設計せよという記事。

具体的な手法 / 使いどころ

- CLAUDE.md は毎会話に効く高レバレッジ設定。推奨は**200行以内**。長すぎると重要度が埋もれ、毎回全体を読み直して推測しブレる。
- **5要素**：①最優先ルール ②キーワード→置き場のルーティング表 ③置き場の決定木 ④タスクの扱い方 ⑤機密情報の扱い（APIキーは.envへ）
- 決定木末尾に「決まらなければ質問せよ」を置く。

なぜ刺さるか / 学び

- AI が推測をやめ参照で動く。
- 同じ依頼に同じ動きで再現性が出る。
- 自信満々の的外れ成果物が構造的に減る。
- 賢い AI ほど、自由にさせるより「型」を渡したほうが性能が出る。



4bit 量子化は fp16 の「妥協」ではない — Qwen3.6-27B (NVFP4) で実測、コサイン 0.997 の near-lossless

要点

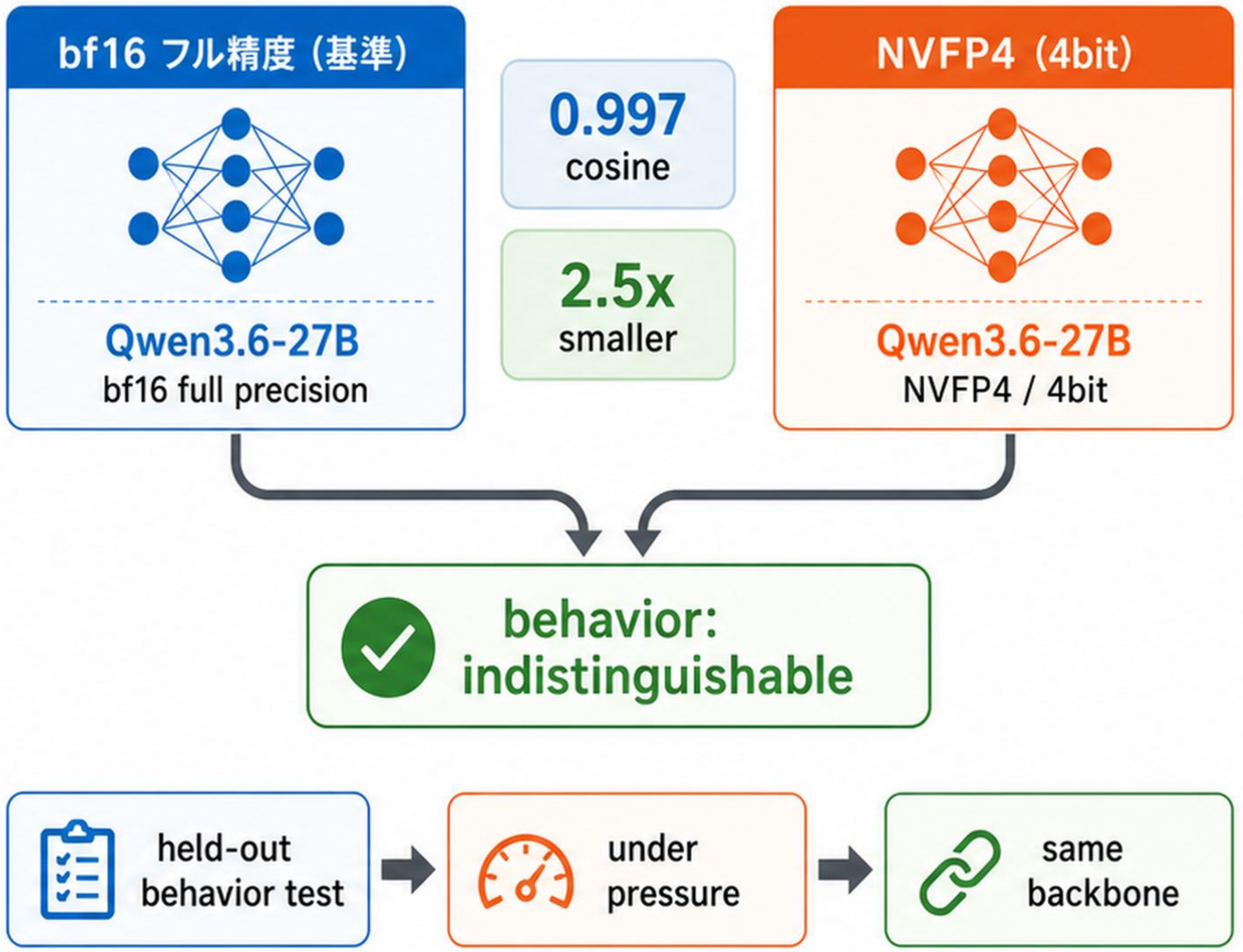
「フル精度の重みこそが本物のモデルで、4bit はやむを得ず妥協するもの」という通念を、実測で逆転させる主張。
NVIDIA の 4bit 版 Qwen3.6-27B (NVFP4) を検証。

具体的な手法 / 使いどころ

- NVFP4(4bit) の Qwen3.6-27B は元モデルにコサイン類似度 0.997 で near-lossless
- 独自の held-out 行動テストで、フル bf16 版と区別不能
- same backbone under pressure: 圧力をかけたときの振る舞いも同一
- サイズは 2.5倍 小さい

なぜ刺さるか / 学び

「量子化＝劣化版」という直感は、最新の 4bit (NVFP4) では必ずしも当てはまらない。フル精度を運ぶのは、丸め誤差のために 2.5倍の重みを担ぐようなもの。



AEON vLLM Ultimate v0.24.0 — DGX Spark 専用最適化コンテナ、単体実測で Gemma-4-26B が実用最速

X / @SpaceTimeViking・@SlimTradeyBaby
likes

何が起きた？

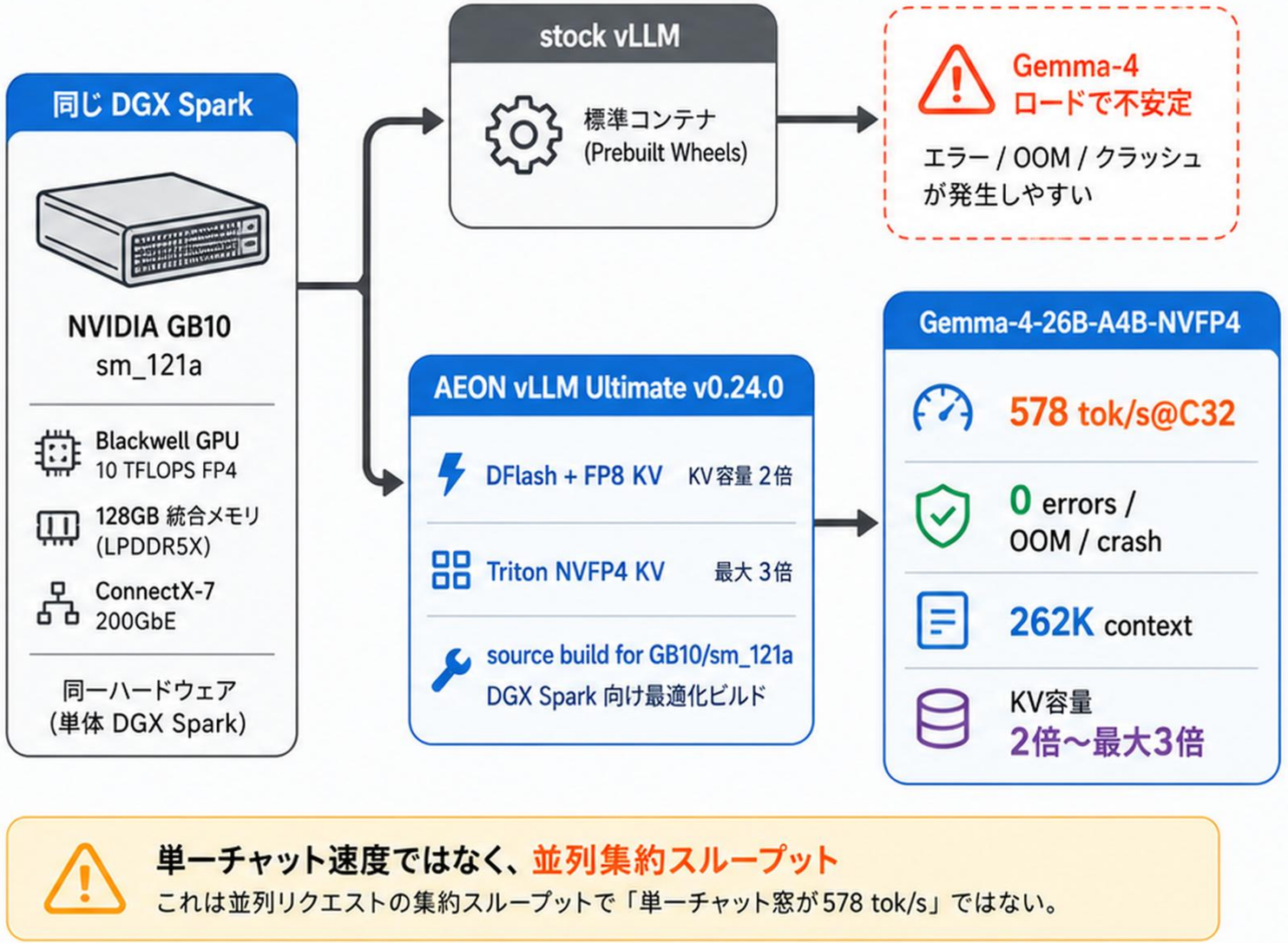
DGX Spark (GB10/sm_121a) 専用に vLLM をソースビルドした最適化コンテナ AEON vLLM Ultimate が v0.24.0 をリリース（@SpaceTimeViking）。第三者が単体 DGX Spark で実測（@SlimTradeyBaby）。

主な変更点

- DFlash+FP8 KV: KV 容量2倍
- Triton NVFP4 KV: 最大3倍
- stock が落ちる Gemma-4 ロードやツール呼び出しも動作
- 262K コンテキスト
- 公称536 tok/s@c16

なぜ重要？

実測では Gemma-4-26B-A4B-NVFP4 が 578 tok/s@C32・エラー/OOM/クラッシュ 0 で実用最速。公称536@C16 は再現せず。これは並列リクエストの集約スループットで「単一チャット窓が578 tok/s」ではない。ローカルエージェント/バッチ/API サービング向けの数字。同じハードでもランタイム側の最適化で throughput とコンテキスト長が激変する。



単一チャット速度ではなく、並列集約スループット
これは並列リクエストの集約スループットで「単一チャット窓が578 tok/s」ではない。

今日のまとめ

本日のトピック一覧

1 Kimi K2.7 Code が GitHub Copilot で一般提供（GA）

— Copilot のモデルピッカーに初のオープンウェイトモデル

2 Apple が Safari に MCP サーバをネイティブ搭載（Safari Technology Preview 247）

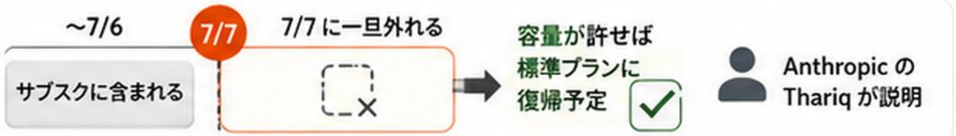
— ブラウザ操作の AI 連携が Chrome 独占でなくなる

3 Alibaba「SkillWeaver」

— 2,209 個のツールでも溺れないエージェント framework、
タスク分解＋必要ツールだけ取得でトークン使用を99%削減

4 Fable は 7/7 にサブスクから一旦外れるが、容量が許せば標準プランに復帰予定

— Anthropic の Thariq が説明



5 Claude Code の性能を引き出す CLAUDE.md 設計術

— 「メモ帳」でなく「判断の地図」にする 5 要素

6 4bit 量子化は fp16 の「妥協」ではない

— Qwen3.6-27B (NVFP4) で実測、コサイン 0.997 の near-lossless

7 AEON vLLM Ultimate v0.24.0

— DGX Spark 専用最適化コンテナ、単体実測で Gemma-4-26B が実用最速

Sources: OpenAI Anthropic Google Apple GitHub Alibaba

TEKION Group