2026-08-04

1. OpenAI、未発表モデルの内部版が数学の未解決問題10件で新結果
2. Qwen3.8-Max 発表
3. 「No Vibes Allowed」

8トピック

OpenAI、未発表モデルの内部版が数学の未解決問題10件で新結果

トークン代は約$2,000、ただし検証はこれから

OpenAI が「次期メジャーモデルの内部版」を使い、10年以上未解決だった数学・理論計算機科学の問題に対して**10件の新しい結果**を出したと発表した。

- 対象領域は球充填・符号理論・群論・量子計算複雑性・格子暗号・極値組合せ論。具体的には非 sofic…
- 二次的な集計によれば 10件のうち4件は「数学者がそうだと信じていたことへの**反例**」
- OpenAI は議論の中身はすべて**モデル由来**とし、原稿の執筆と Lean での形式化は**人間が行った**と説明

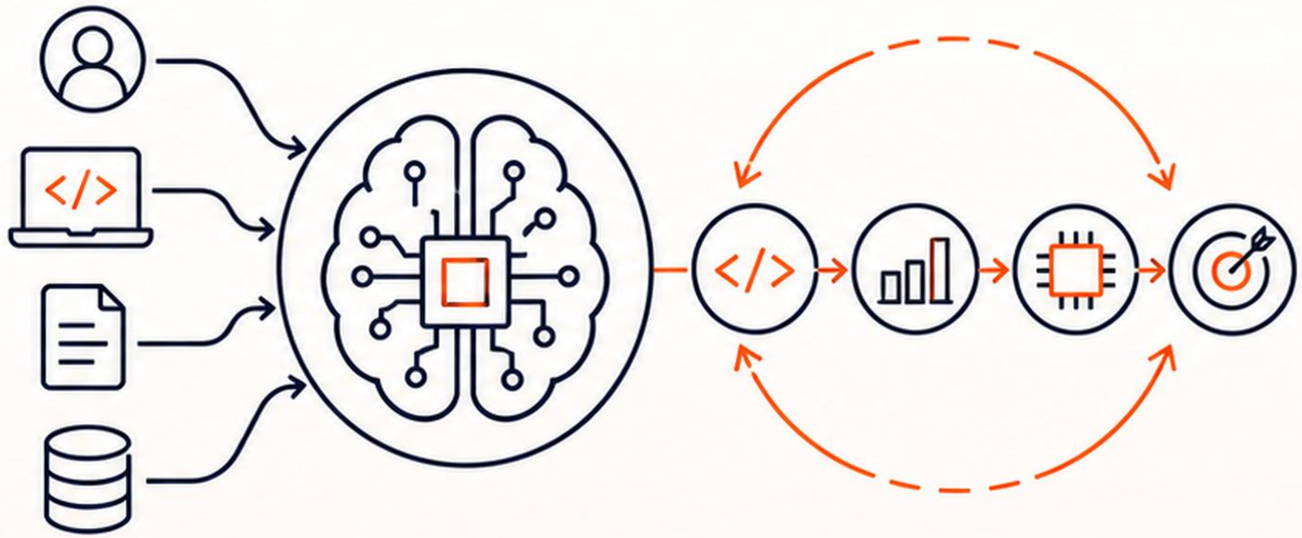
openai.com

Qwen3.8-Max 発表

2.4Tパラメータ、来週フラッグシップの重みを公開

Alibaba が最上位モデル Qwen3.8-Max（2.4Tパラメータ）を発表した。

- **2.4T パラメータ。** コーディングと "cowork"（数百職種の実務遂行）に照準
- **自律コーディング10日超:** 全プロジェクトトレースを qwen-code-dev-bot/oh-my-cli で公開…
- **長期ホライズン:** チップ設計最適化で500ターン超、EC戦略で365日相当の閉ループ適応学習を駆動したと主張

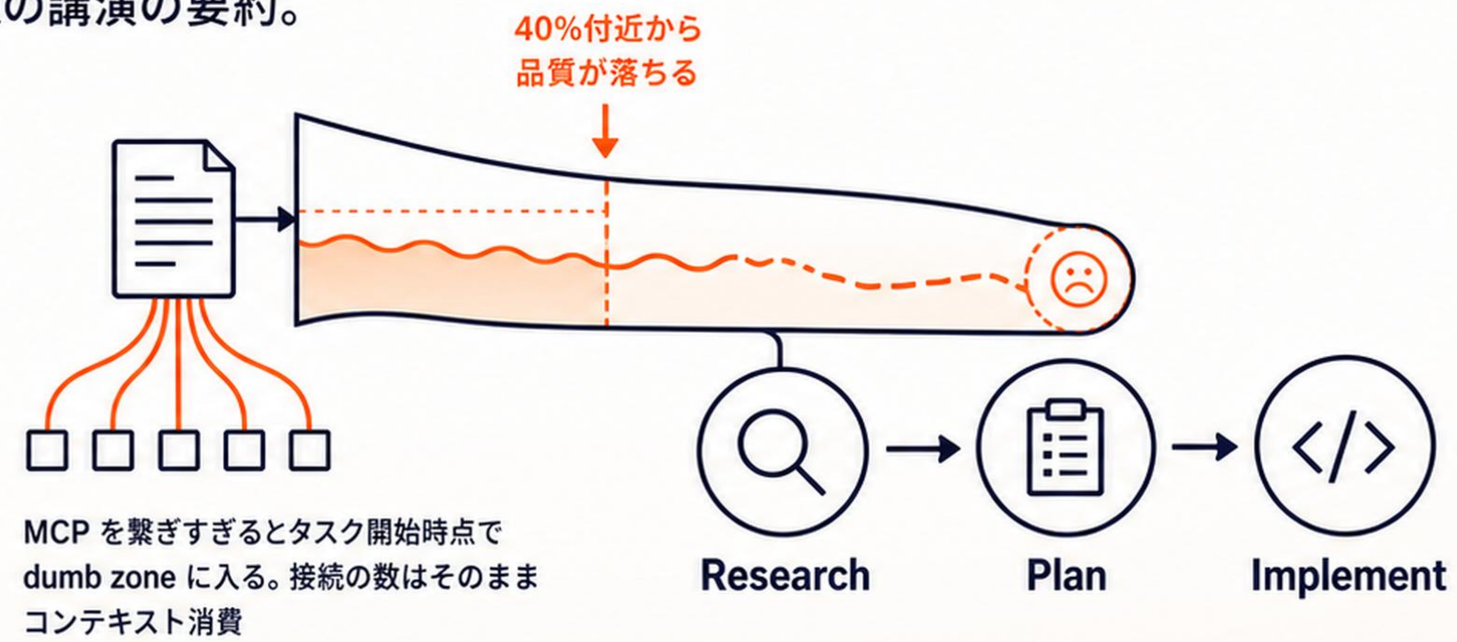


「No Vibes Allowed」

エージェントの品質はコンテキストの40%から落ちる。だから Research → Plan → Implement

"context engineering" という語を作った本人の講演の要約。

- **品質の劣化はコンテキストウィンドウの40%付近から始まる**（使い切る手前ですでに落ちている）
- **MCP を繋ぎすぎるとタスク開始時点で dumb zone に入る。**接続の数はそのままコンテキスト消費
- **Research:** エージェントはコードベースを読むだけ。何も書かせない



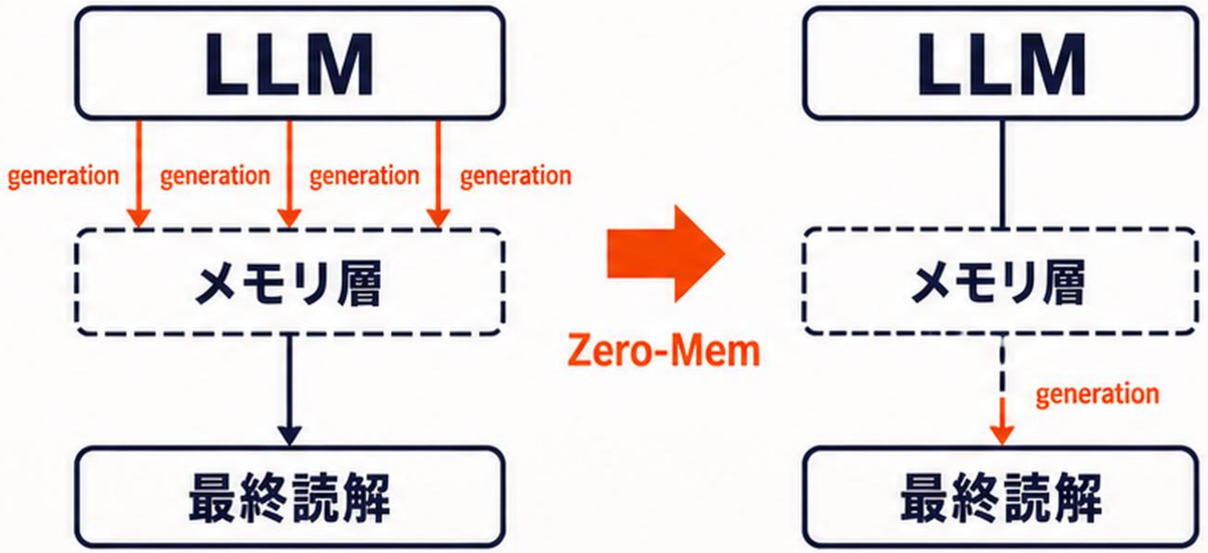
x.com

Zero-Mem

エージェントのメモリから LLM 呼び出しを全部外す

「エージェントのメモリに LLM は本当に必要か」を検証した論文が話題になっている。

- 問題は3か所のモデル呼び出し: ①やりとりの要約 ②レコードの書き込み ③検索結果の再ランキング
- 要約は情報の損失を伴う。捨てられるのは往々にして後から必要になる証拠
- Zero-Mem はメモリ層から generation を排除。生成を使うのは最終読解のみ

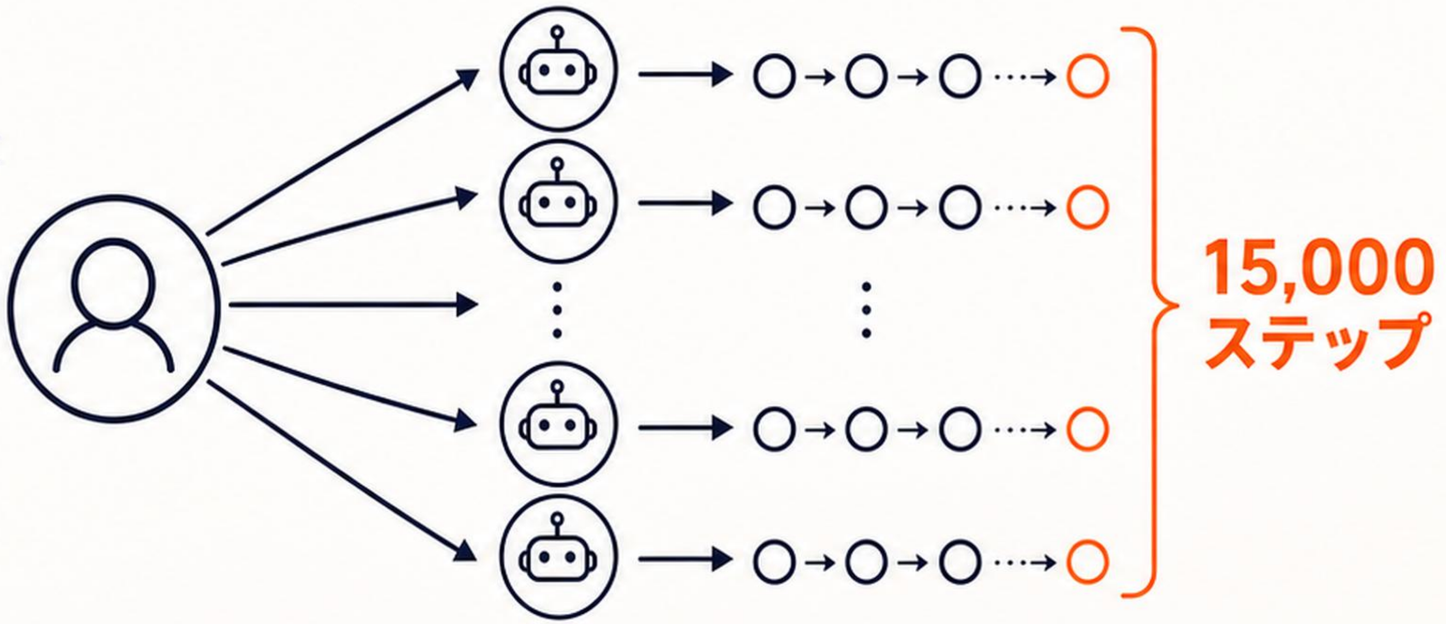


15,000ステップ・150サブエージェントの swarm

「作った側に自分の仕事を採点させない」

単一のエージェントループは数百ステップで力尽きる、という前提に対する実行記録。

- **質問ではなく「仕事」を書く。**証拠の連鎖と**サブエージェント間の意見の相違**を明示的に出力要求する…
- **リードエージェントに採用させる。**リードが150体分のサブエージェント仕様を書いた
- **15,000ステップ**を走り切り、途中で力尽きなかった

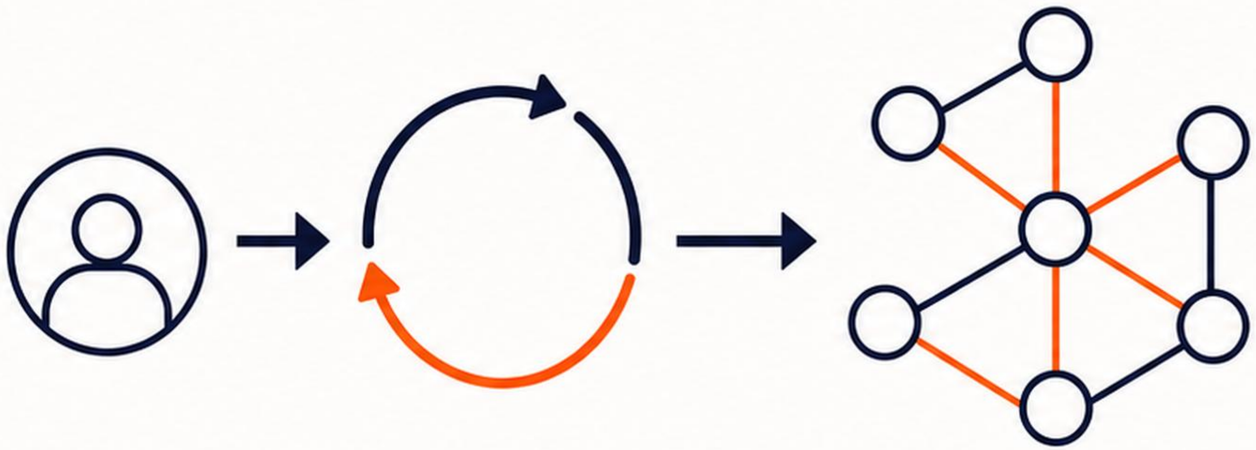


Google が2時間の Graph Engineering 無料コースを公開

単体エージェントの先を教える

Google が **Graph Engineering の２時間無料コース**を公開した。

- 17:44 → 最初のエージェントを作る
- 39:30 → **ループ工学 (Loop engineering)**
- 1:12:38 → **グラフ工学 (Graph engineering)**



GPT-Live が「話しながら聞ける」ように

音声スタックをクライアントからモデルまで作り直し

OpenAI が GPT-Live の音声アーキテクチャを刷新した。

- 音声が**専用の fast path** を通る。
  深い推論とツール使用は非同期で並走
- そのため推論やツール実行が
  会話を**中断しない**
- 音声セッションの起動を
  6回のネットワーク往復から1回に削減

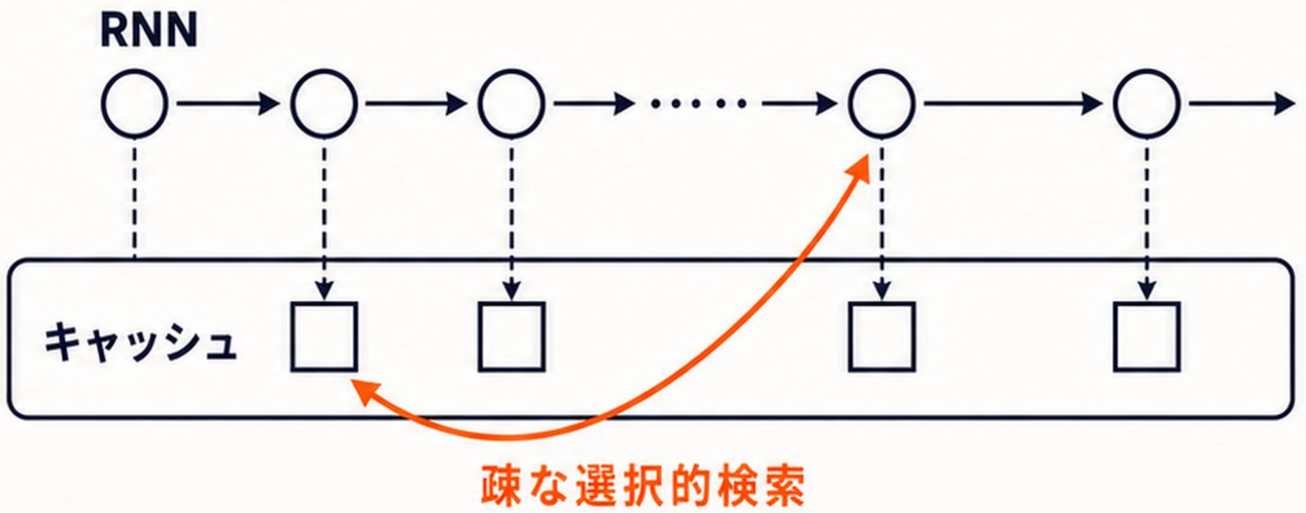
openai.com

Google の Memory Caching

RNN が Transformer 精度に並び、線形計算量を維持したという主張

Attention の $O(L^2)$ を回避しつつ長文の精度を保つ、というアーキテクチャ研究の話題。

- 従来の二択: Transformer = 高精度だが $O(L^2)$ / RNN = 線形で速いが固定サイズの記憶ゆえ長文で忘れる
- MC は隠れ状態のチェックポイントをキャッシュし、記憶容量が系列長に応じて伸びるようにする
- 疎な選択的検索 (sparse selective retrieval) で必要なときだけ過去を参照。全トークン間の総当たり Attention を避ける



本日のトピック一覧

1. OpenAI、未発表モデルの内部版が数学の未解決問題10件で新結果
2. Qwen3.8-Max 発表
3. 「No Vibes Allowed」
4. Zero-Mem
5. 15,000ステップ・150サブエージェントの swarm
6. Google が2時間の Graph Engineering 無料コースを公開
7. GPT-Live が「話しながら聞ける」ように
8. Google の Memory Caching

2026-08-04

TEKION Group