ローカルAIエージェントの可能性

―AIエージェント時代に加速する新たな選択肢―

2025年11月22日

自己紹介

安東 竜平

・2001年3月30日生まれ

・中学～大学まで駅伝

・2024/2/8 に Link AI を起業

・事業内容：

AIエージェント

AIアバターの開発

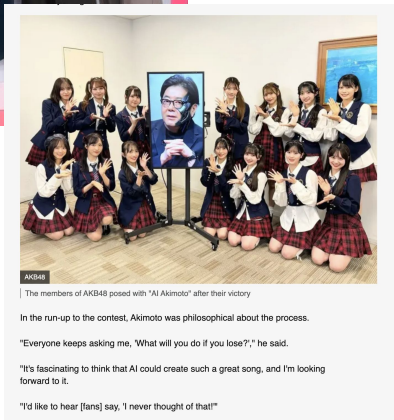
AI駆動開発研修など

・LinuC Open Network（Linuxコミュニティ）の運営

実績

【秋元康×AI秋元康 ～AKB48新曲対決～】（日本テレビ特番）

音楽プロデューサー秋元康さんの思考 /過去作/見た目/声を学習した AI秋元康を開発いたしました。



イギリスの BCCに掲載された記事 ▸
https://www.bbc.com/news/articles/c1kwvlyrjyxo.amp

音楽プロデューサー秋元康さんの
思考や過去作を学習した
AI秋元康の開発

思考だけでなく秋元康さんの
見た目、声も AIで再現

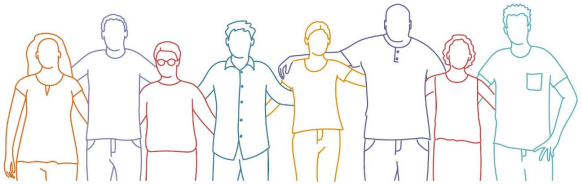
https://www.itmedia.co.jp/aiplus/articles/2509/16/news091.html

立場

LinuC Open Networkとは

IT技術を活用する人の、成長と活躍を後押しする

「アウトプット共創型 エンジニアコミュニティ」



どんな人が参加している？

Linux等のオープンテクノロジーを活用するIT技術者として、
成長と活躍を目指す方々が参加しています。

より良いIT社会の実現へ

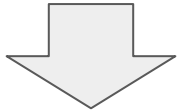
Link AI

講演の目的

今後の「技術者としての成長」を加速

サマリー

なぜ「ローカル AIエージェント」がテーマ？



ローカル AIエージェント が実装できる
エンジニアは希少価値が高い

※価値を高める＝成長としている

サマリー

①なぜローカル AIエージェントが大事か理解する（メイン）
②何をしたらいいか考える

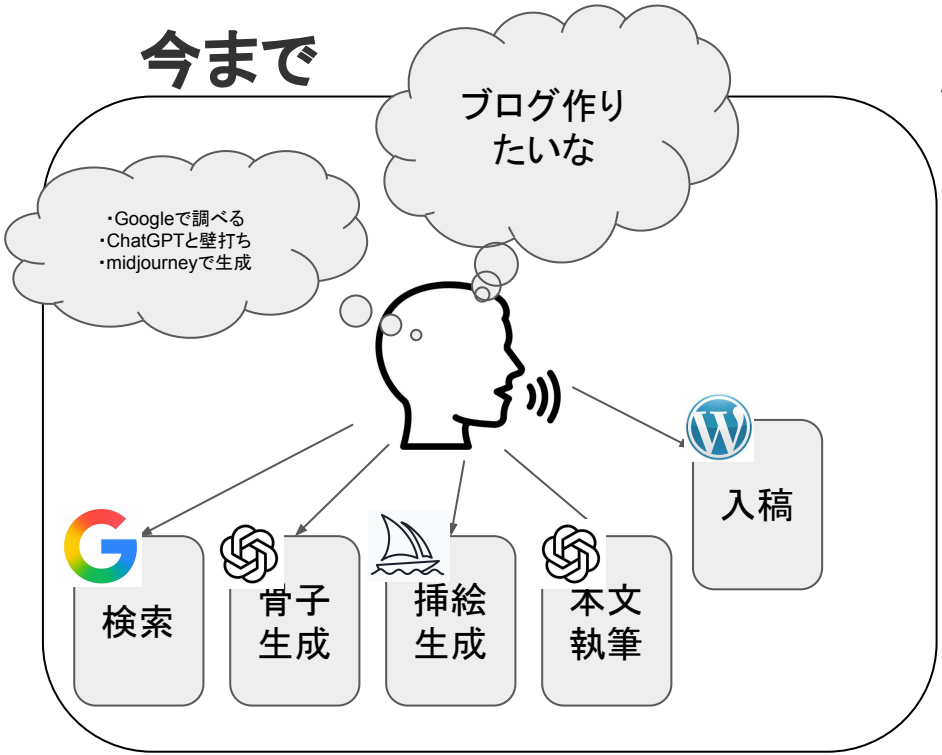
Link AI

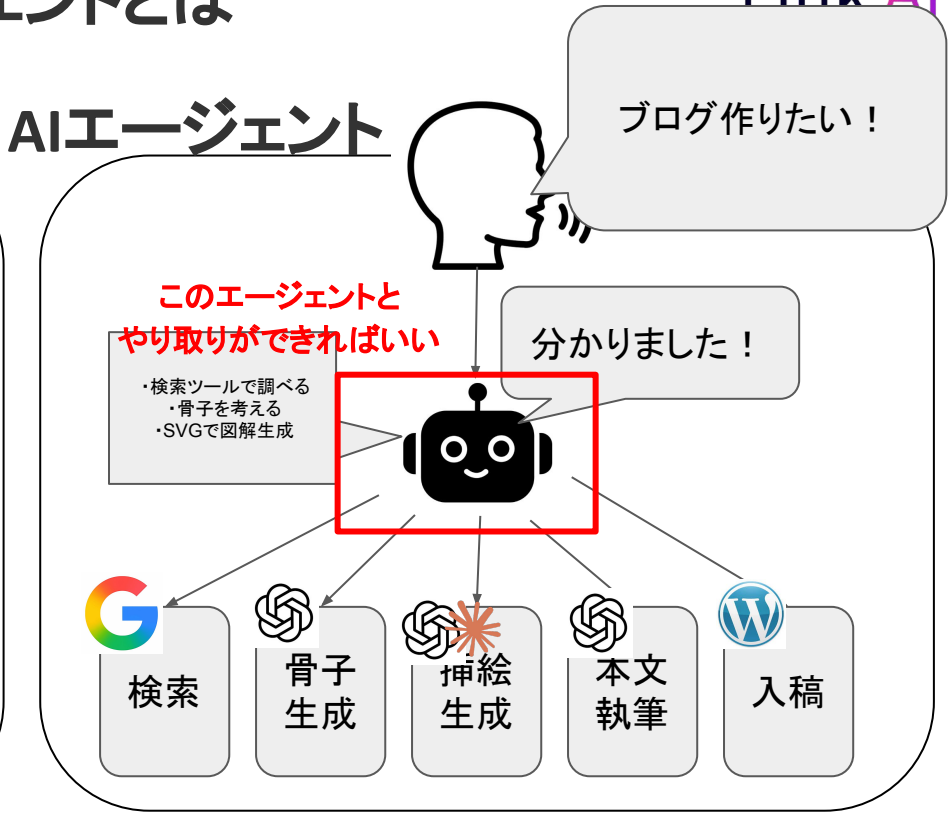
そもそもAIエージェントとは？

自律的に考え複数のツールを使いながらタスクを遂行するシステム

https://manus.im/share/cQ3x59rgaiG76pAaee6vXJ?replay=1

AIエージェントとは





ツール選択に人が介入

ツール選択に人は介入なし

Link AI

なぜAIエージェントを実装できると価値が高いのか

Link AI

まず前提・・・

技術者（エンジニア）の役割は

・人の仕事をより楽にしていく

・できることを拡張していく

※いろんな哲学はあると思いますが

Link AI

システムがないときは人間大変

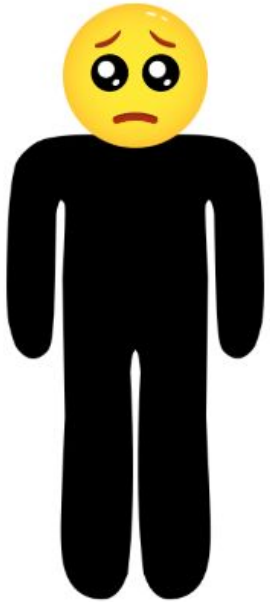
商談データ

支出データ

顧客データ

売上データ

情報整理　思考　タスク実行

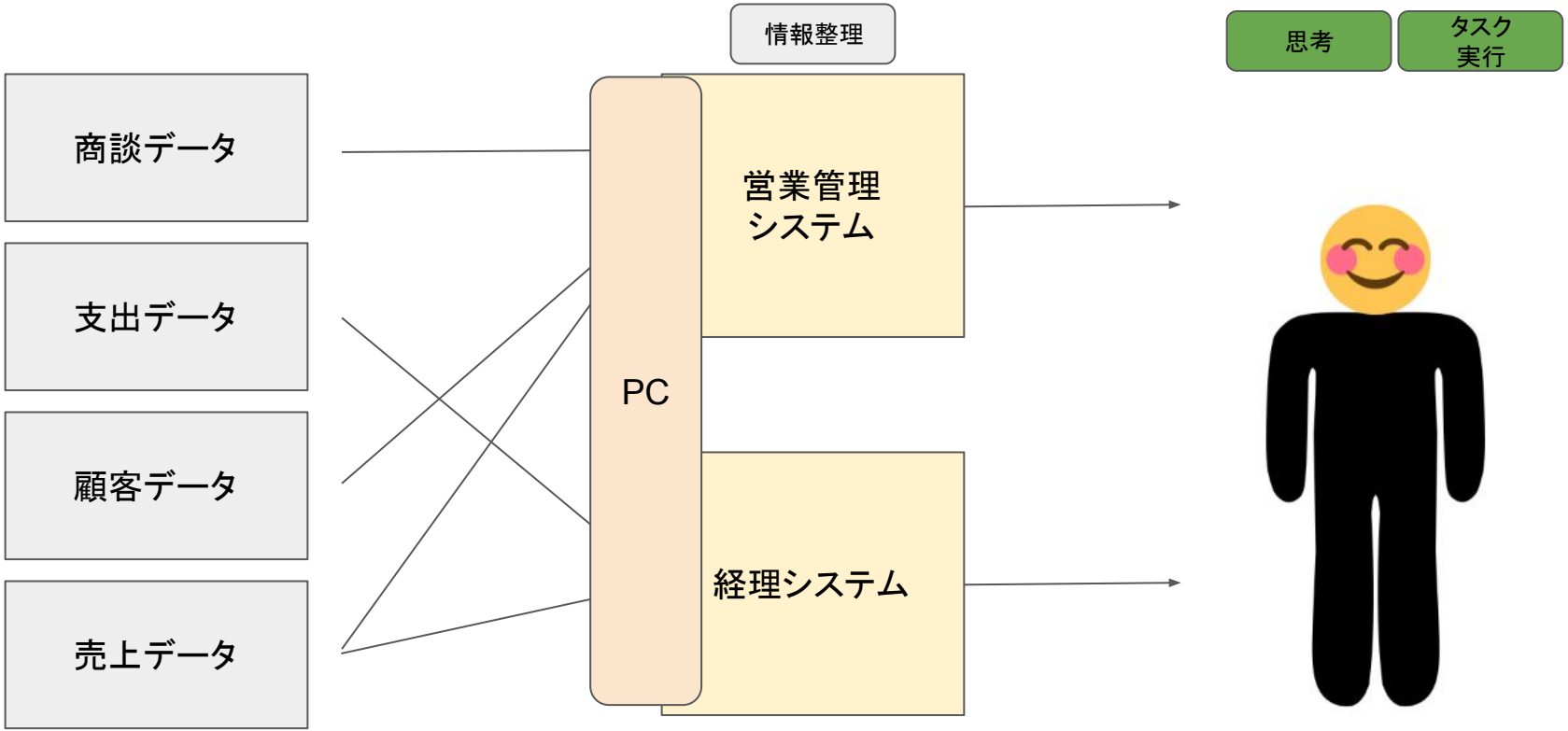


人間側の情報処理が大変
パフォーマンス発揮できない

前提

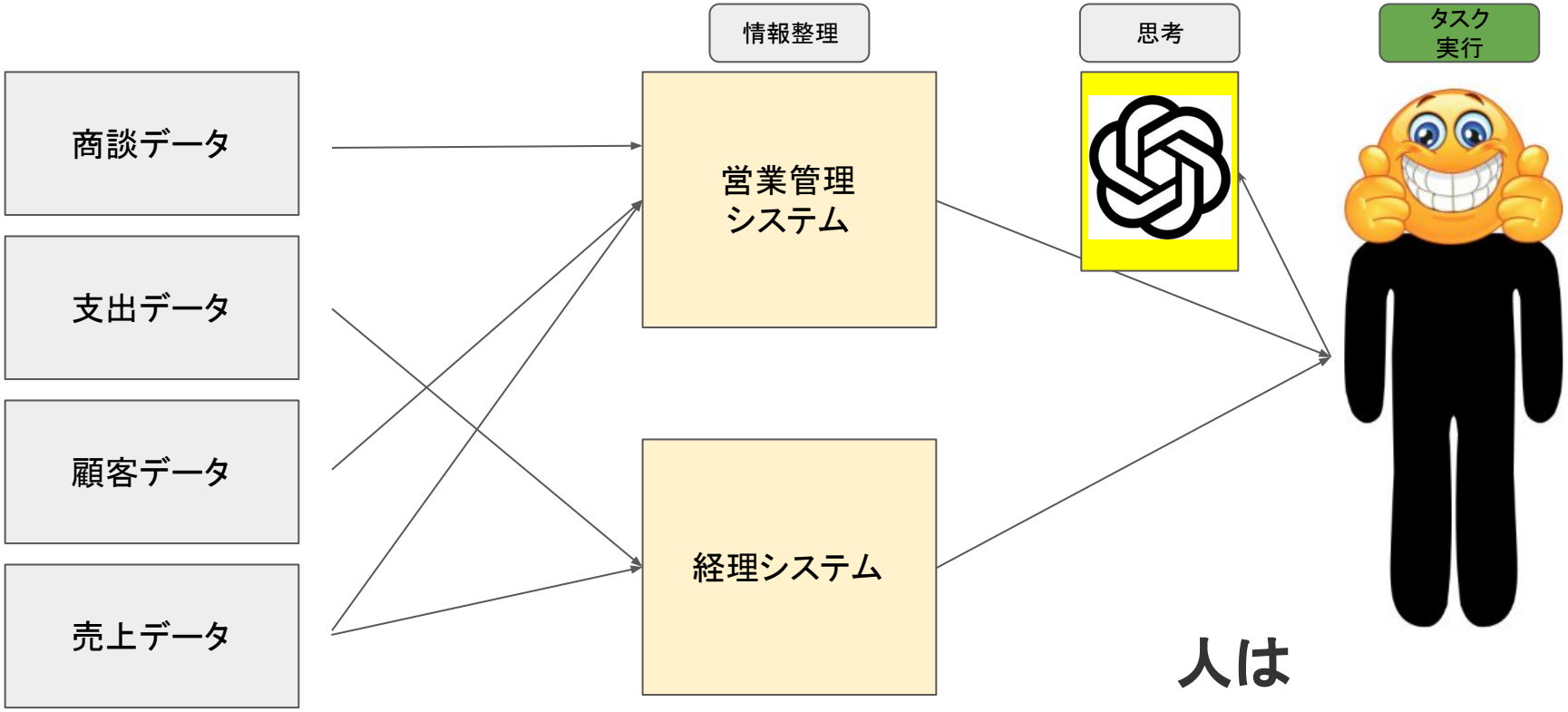
システムがあると人間楽ちん
だけどシステムが複雑になるという課題

Link AI



AIによってデータの整理だけでなく「思考」までできるようになった

Link AI

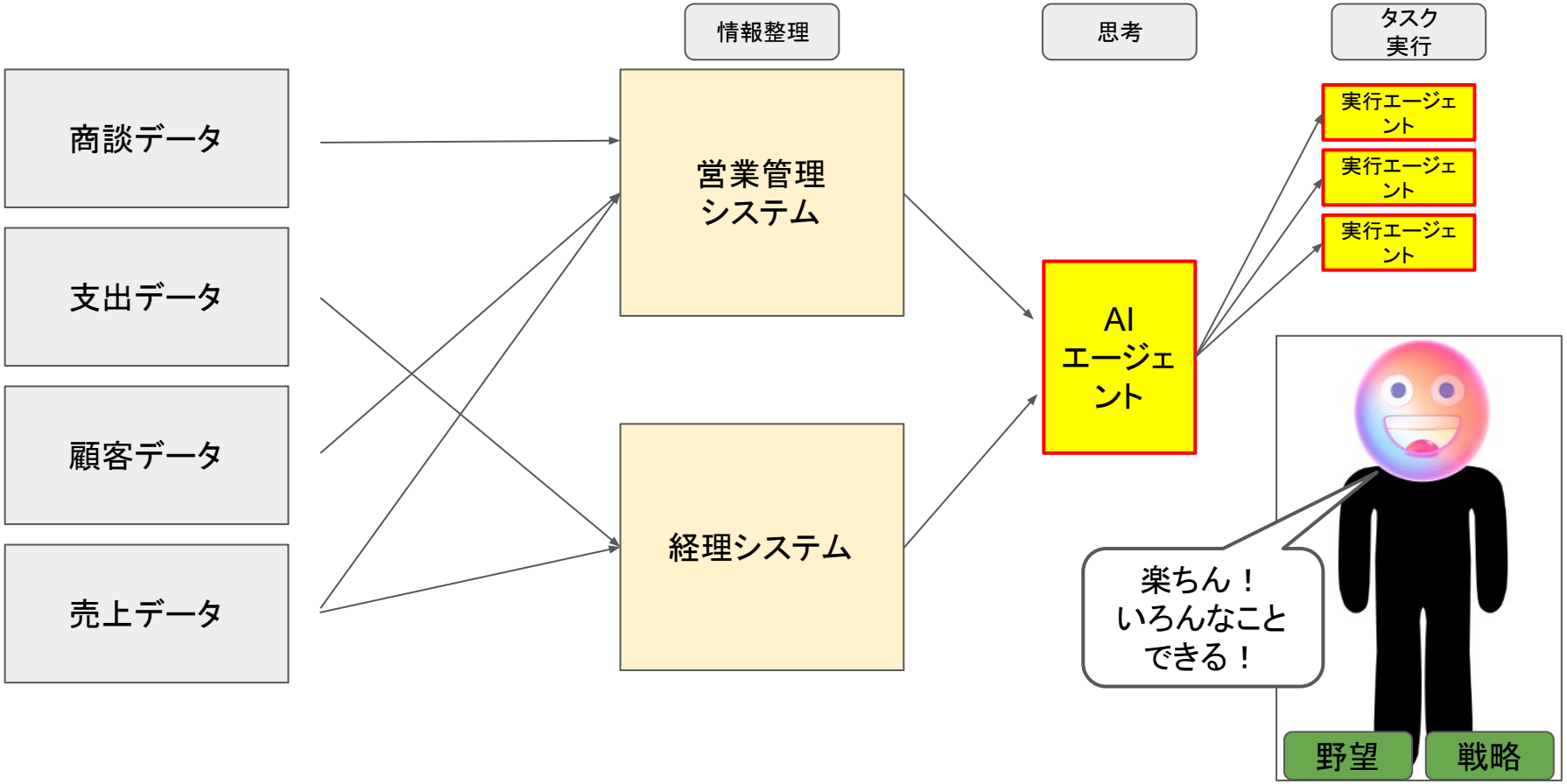


人は
「実行」に集中できる

前提

エージェントによって「実行」まで任せられるように

Link AI



Link AI

整理

なぜシステム開発するか？
→仕事を楽にする /できること増やす

AIエージェントの方が楽になる
かつ多くの人が使える

AIエージェントの開発が技術者の仕事になる
(将来はロボット？)

背景：AIエージェントの現在地

自律性アップ

Devin

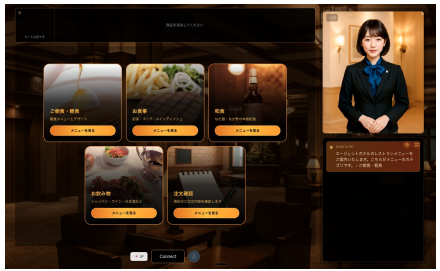
manus

https://manus.im/app/cQ3x59rgaiG76pAaee6vXJ

https://notebooklm.google.com/notebook/ba1bd70b-40ff-4cde-b508-e7db608a7918

リアルタイム性アップ

AIウィトレス



AI秘書

AIはどんどん「常時稼働」で便利になっていく

AIエージェント導入する際の課題

ランニングコストの高騰

セキュリティ課題は言わずもがな

AI実装をする際の問題点

（エージェント開発するなら、という話）

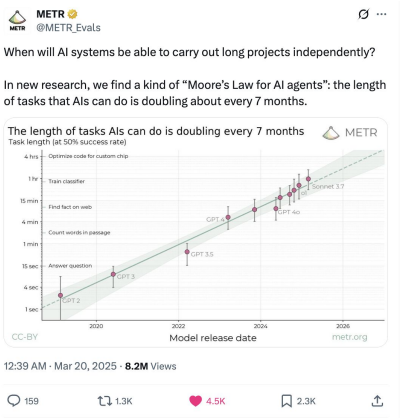
モデルの進化
（例：gpt-3.5→o3 Pro）

✖

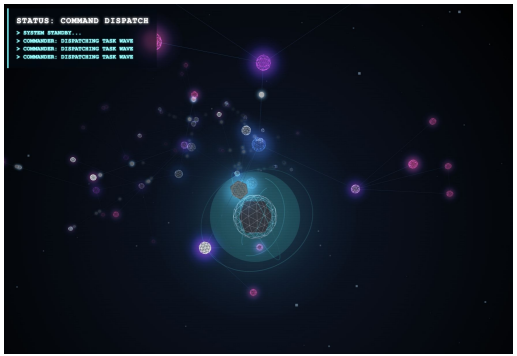
ツールの進化
（例：MCP/A2A）

✖

複数並列実行
（例：Antigravity、カムイ）



さらなるコスト高騰



エージェント化で便利になる一方
活用できる企業や用途が限定的に

※ManusやGensparkなどクラウドのエージェントは入れられないところも多い
→セキュリティ面はもちろんだがコスト面もそう

ソリューション

解決策

＝

ローカルAIエージェント

ローカルAIエージェントのメリット

運用コスト低価
(電気代で動く)

セキュリティ向上

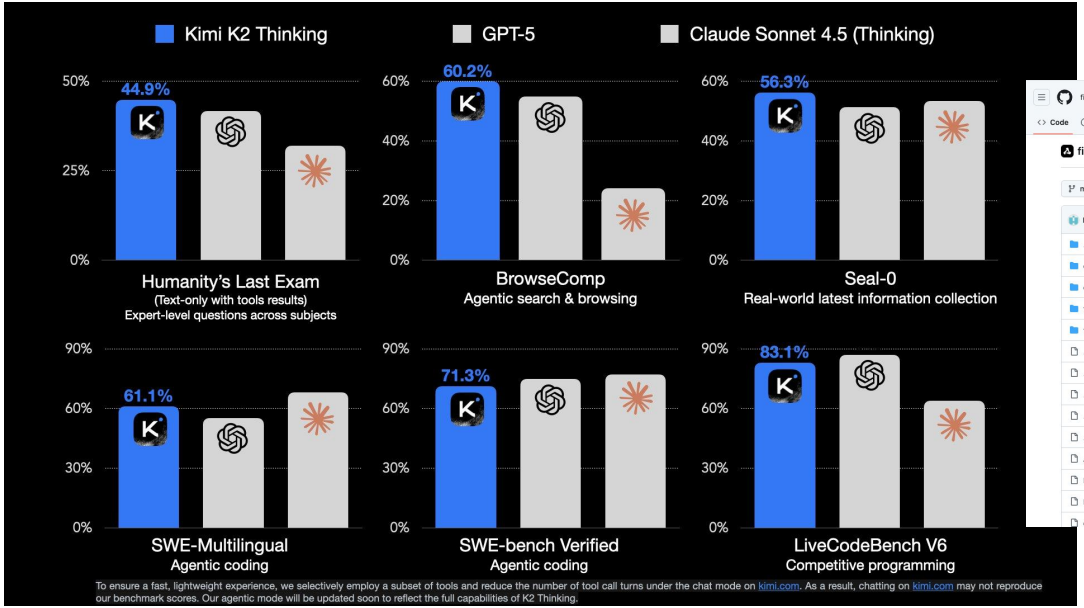
AIエージェントの
活用企業/用途が広がる

懸念：ローカル LLMは精度低い
→モデル精度は時間とともに向上していく

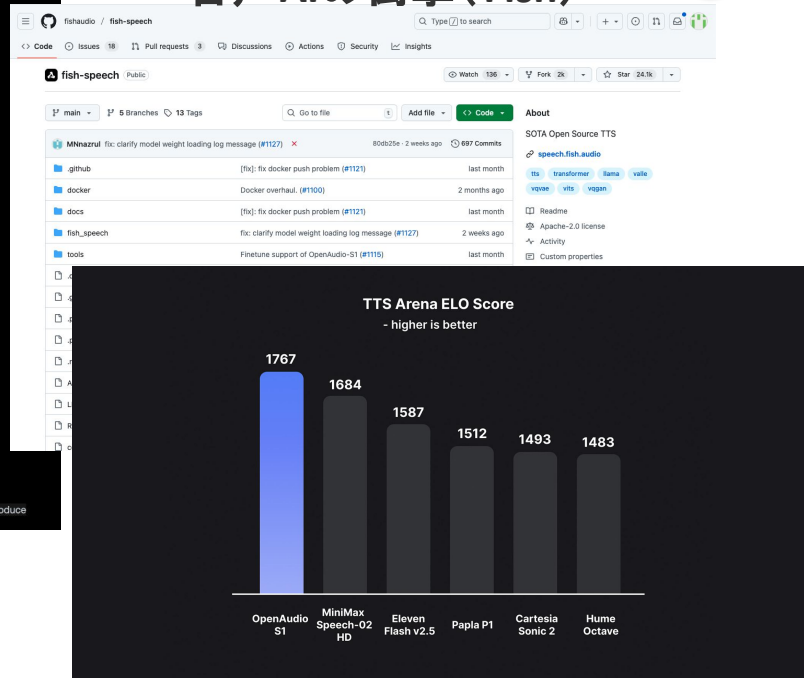
AI導入の論点は「コスト」のみになっていく？

ソリューション

モデル精度は時間とともに向上していく： Kimi K2の衝撃



テキストだけではない。
音声AIの衝撃（Fish）

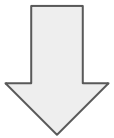


ソリューション

観点	Claude 3.7 Sonnet (API)	Tenstorrent QuietBox (ローカル)	補足
課金の仕組み	使ったトークン量に応じて変動 ($18/100万token)	電気代のみ (稼働時間ベース)	ローカルは"何トークン使っても"コストは基本一定
1日50万tokenの場合 (月1,500万token)	約 ¥41,000 / 月	約 ¥29,000 / 月 (電気代のみ)	利用規模が大きくなるとAPIは高騰
利用量が増えると？	比例してコスト増加	ほぼ一定	"使えば使うほど高くなる"のがAPI、"ずっと使っても変わらない"
スケーラビリティ	△ コスト青天井	⭕ 常時稼働・社内展開に強い	"全社導入"に進んだときの差が
初期導入のしやすさ	⭕ 即導入可能	△ 初期構築あり (本体 + セットアップ)	APIはPoC向き、ローカルは運用最適化向き
長期運用コスト	❌ 利用量に応じて膨張	✅ 予測可能・定額運用	"持ち続けるAI"ならローカルが有利
向いている用途	小規模PoC 少人数の試験導入	常時稼働型エージェント 社内展開・情報統合型	拡張性が求められるほどローカルの価値が上がる

何から始めたらいいか

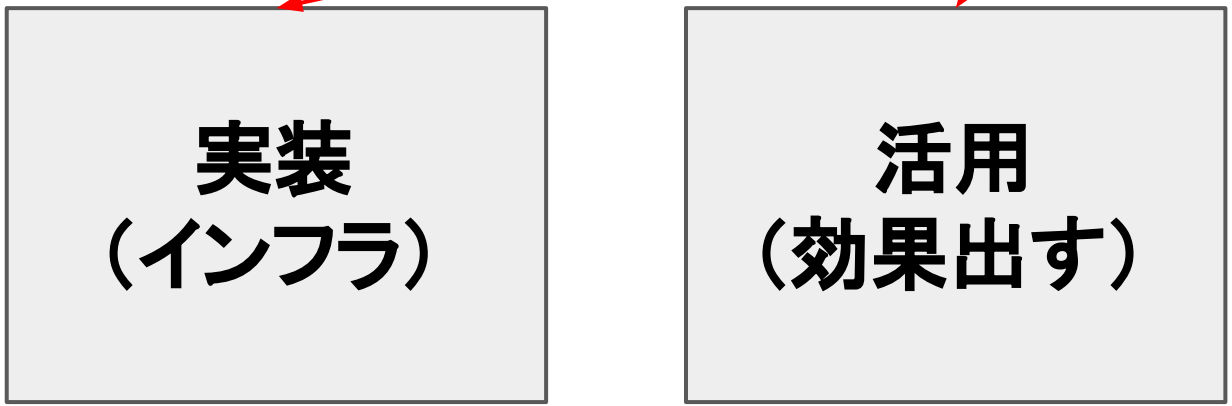
ローカル AIエージェントの実装が価値高い理由



普通のクラウド AIエージェントの実装より難しい

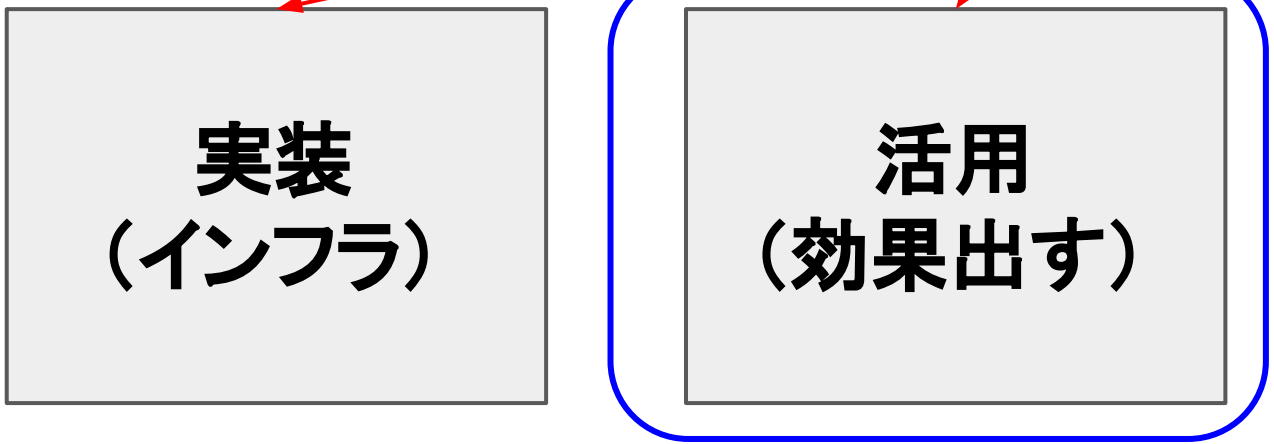
ローカルAIエージェント始め方

普通のクラウド AIエージェントの実装より難しい



ローカルAIエージェント始め方

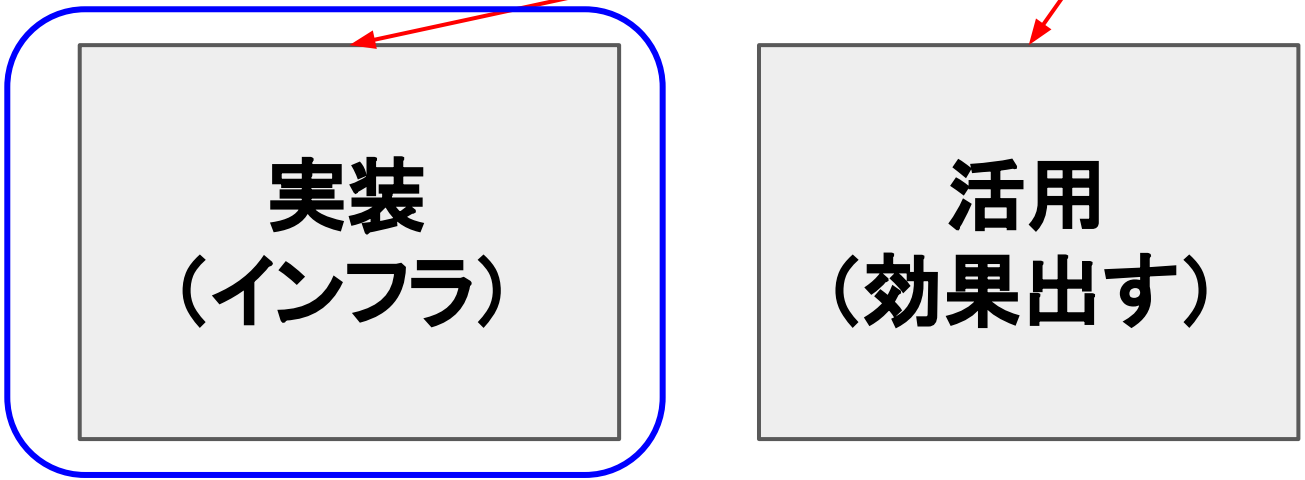
普通のクラウド AIエージェントの実装より難しい



クラウドの AI
エージェントは
これが
論点になる
（コンサル /研
修/受託開発）

ローカルAIエージェント始め方

普通のクラウド AIエージェントの実装より難しい



ローカル AIエージェントだと、ここも重要になるので大変

ローカルAIエージェント始め方

実装
（インフラ）

〇〇の用途で使うには、どんなスペックの
サーバーが必要なの？？
・オンプレミスって何？？
・GPUも必要？LPUとか色々あるけど何？
・初期費用がかなりかかるけどクラウドの
エージェントじゃダメなの？

ビジネスサイドに
説明が大変
（基礎知識が必要）

そもそも基本的な
コンピュータの知識
ないと理解できない

ローカルAIエージェントだと、ここも重要になるので大変

ローカルAIエージェント始め方

〇〇の用途で使うには、どんなスペックの
サーバーが必要なの？？
・オンプレミスって何？？
・GPUも必要？LPUとか色々あるけど何？
・初期費用がかなりかかるけどクラウドの
エージェントじゃダメなの？

実装
（インフラ）

ビジネスサイドに
説明が大変
（基礎知識が必要）

そもそも基本的な
コンピュータの知識
ないと理解できない

コンピュータの知識/ローカルLLMの知識/ビジネスインパクトなど活用の知識
これらの総合格闘技が必要なのでローカルLLM実装は難しい
▶これができる人材は「希少価値が高い！」

垂直統合スキル：ローカルLLM実装の希少価値

従来のコンサル

従来のAI研究者

従来のインフラ担当

希少価値！

3. ビジネス・戦略（成果）

- API vs ローカルのROI試算
- セキュリティ（データ秘匿性）設計
- 業務適合性とレイテンシ判断

Vertical Integration
（垂直統合スキル）

2. ソフトウェア・モデル実装（技）

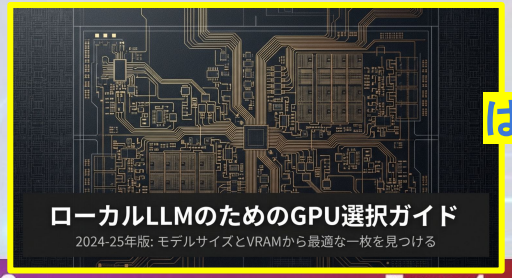
- モデルの目利き（Llama 3, Mistral）
- 量子化技術（4bit化で軽量化）
- 推論エンジン最適化（vLLM, llama.cpp）

1. インフラ・ハードウェア（土台）

- GPU選定 / VRAM計算（70億パラ=14GB?）
- メモリ帯域幅の理解
- Linux環境構築・ドライバ依存解決

LLMローカル推論：目的別 GPU 選択マップ

前提条件 (Q4量子化, 4K-8K ctx, 単一GPU)
⚠ 注意: 長コンテキスト(128K+)はVRAM 2-4倍消費



詳細はこちら

モデル規模 (パラメータ数 & 必要VRAM目安)

Zone 1
入門〜標準クラス
(7B-14B) / VRAM ~8GB

- RTX 3060/4060系 (〜12GB)
- RTX 4070系 (〜16GB)

手軽に試す
価格：安価

Zone 2
実用〜ハイエンドクラス
(32B前後) / VRAM 15-20GB

- RTX 4090 (24GB)
  32Bの鉄板・現実的ライン
  価格：〜40万円
- Tenstorrent Blackhole p100a (28GB)
  ⚠VRAMコスパ◎, 玄人向け(ソフト難)
  価格：〜15万円
- RTX 5090 (32GB)
  32B余裕・将来性◎
  価格：〜40万円+

Zone 3
ガチ勢・ワークステーション
(70B dense) / VRAM 33-40GB+

- RTX Pro 5000 (48/72GB)
  自宅で70B常用
  価格：高価
- RTX Pro 6000 (96GB)
  オンプレ開発・100B級MoE
  価格：超高価 (〜150万円)

⚠電源・冷却シビア

Zone 4
データセンター・超巨大モデル
(MoE 200B〜, マルチGPU前提)

- NVIDIA H100 (80GB)
- AMD Instinct MI300X (192GB)
- Intel Gaudi 3 (128GB)

業務用・クラスタ必須
価格：ASK (数百万円/基〜)

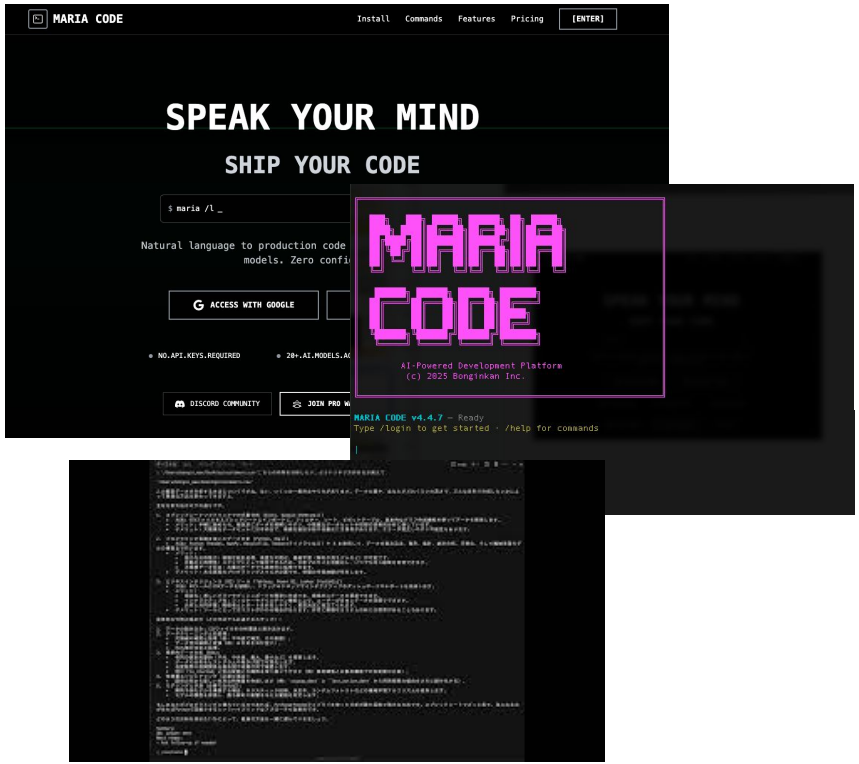
32Bの壁 (ここからガチ勢)

結論：「どのハードを買うべき？」

1. 「32BをPCで快適に」 → RTX 4090が最低ライン。余裕なら5090。p100aは挑戦者向け。
2. 「70Bもローカルで」 → RTX Pro 48GB以上が必要。導入ハードル高。

※価格・仕様は2025年時点の目安であり変動します。

ローカルLLMを使えるインターフェースとなる AIサービス



Dify

OpenAI Agents SDK

ただしセットアップや
ハードウェアとの相性など
考慮することが多い

実際にローカルLLM活用を進めていく現実的な方法

ChatGPT等を活用して進める

ユースケースを決める

必要なローカルLLMをリサーチ

必要なハードウェアをリサーチ

AIに聞きながら実装

or

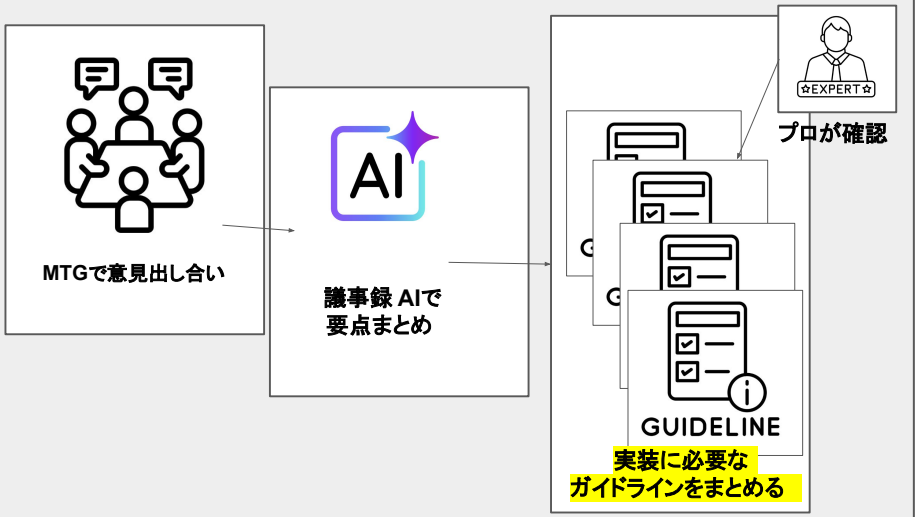
LIONに入る

LinuC Open Network

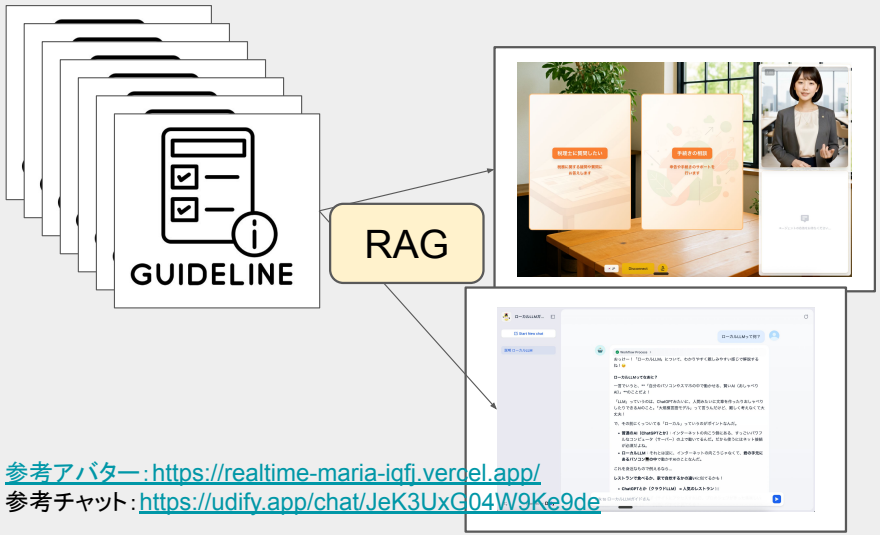
LiONでの取り組み

この希少価値の高いローカル LLM人材を育成するために LiONでは「ローカルLLM構築ガイドさん開発プロジェクト」を始動

必要知識 /事例/活用法を体系化

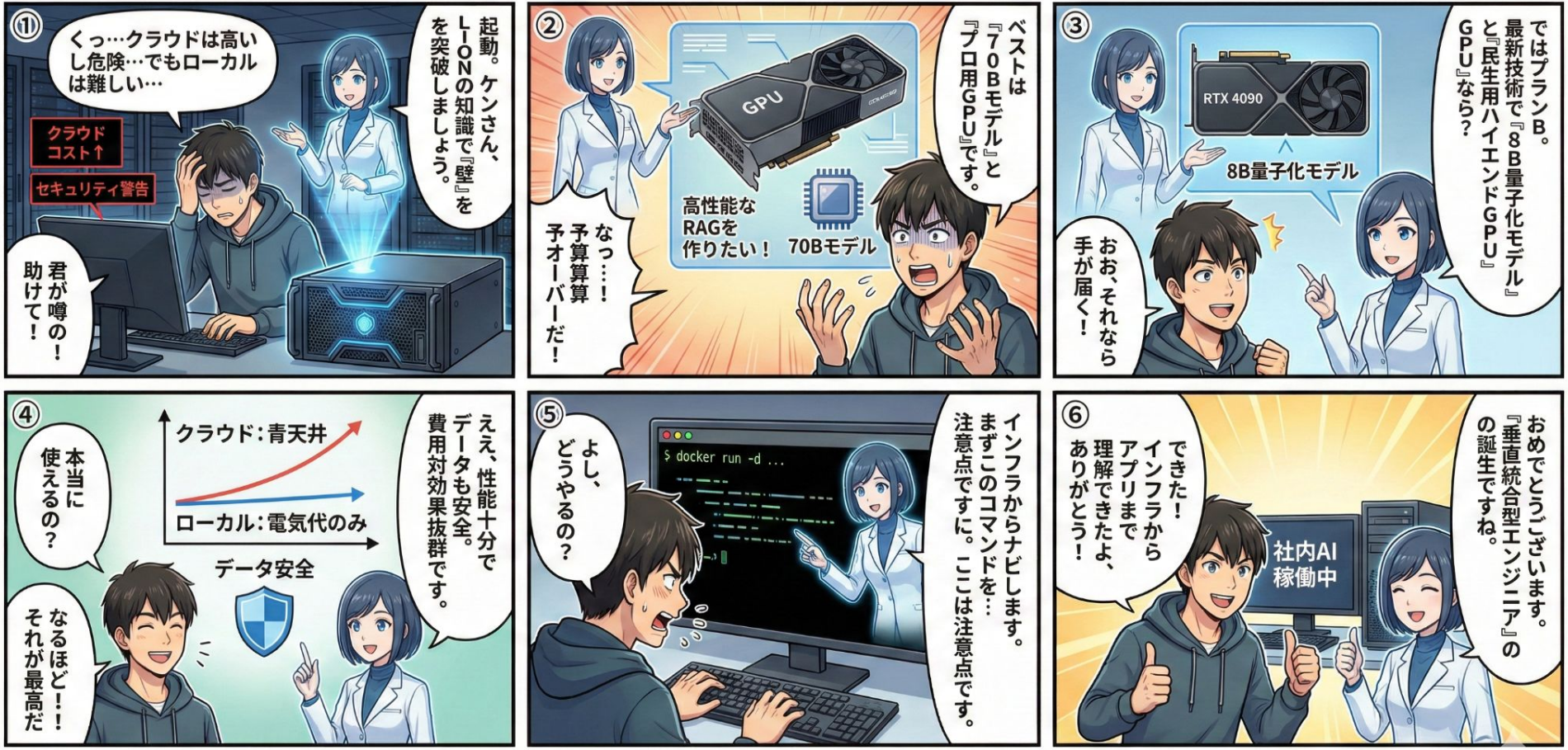


体系化したガイドラインで「 AIガイドさん」を開発



参考アバター：https://realtime-maria-iqfj.vercel.app/
参考チャット：https://udify.app/chat/JeK3UxG04W9Ke9de

頼れる相棒！ローカルLLMガイドさん



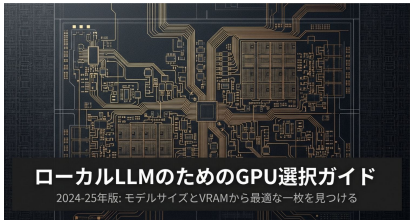
知識を力に変え、AIを使いこなす側へ。
LIONのローカルLLMガイドが支援します。

一緒に次世代の
「人材育成AIガイドさん」を開発しましょう！

この実績/経験はいろんなところで活用できるはずです！

LiONでの取り組み

入ってくれたら、
NotebookLMで一瞬で作ったローカル LLMガイド



私が想いを込めて手作業で作った今日の資料

も手に入ります

今後

従来の仕事のやり方が大きく変わるため、今まで学んできた常識を捨てよう

開発者イベントに
開発者以外の招待者
開発≠エンジニアがやるもの

AI前提の
ワークフロー構築
の必要性

ローカル AIエージェント
ぜひお試し下さい。