AI Agentに必要な最後のピースとは？
～ AGIに必要なもの ～

2025.6.11

（株）テッキーズポッド

戌亥 稔

本日のトピック

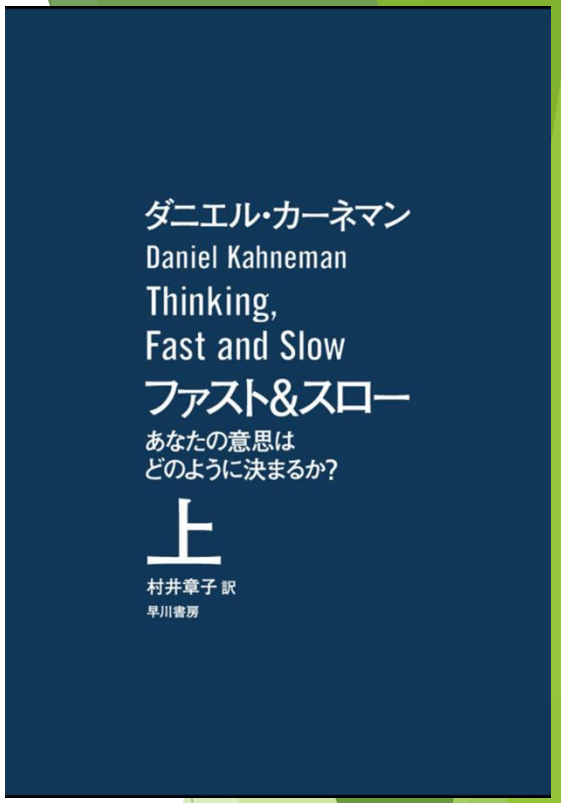
- 自己紹介（何故AIを始めたか？）
- AIモデルのアーキテクチャ
- AI Agentとは
  - ✓ AlphaEvolve、OpenAI Codex/Google Jules
  - ✓ MCP（Model Context Protocol）
- Samurai-Cloud（セキュリティ）

私とAI

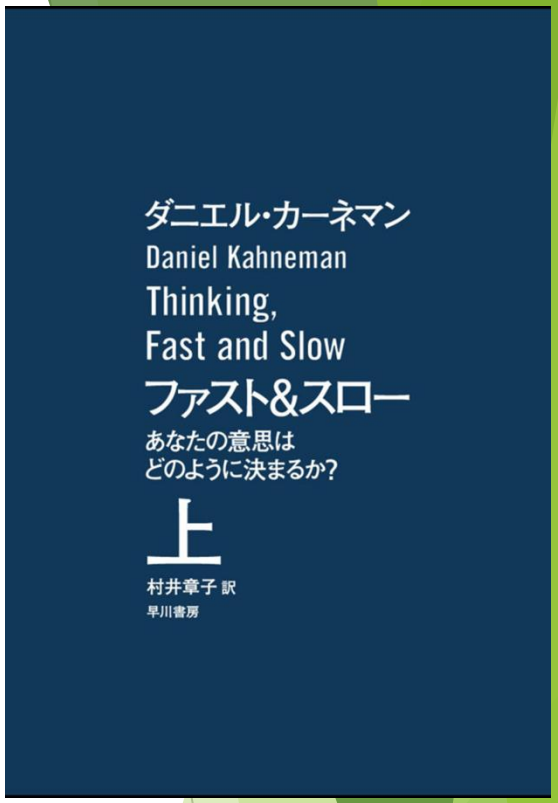
- 1992年フロリダ工科大学コンピュータサイエンスの修士取得
  - ✓ 修士論文：The Recognition of Imperfect Strings Generated by **Fuzzy Context Sensitive Grammars**
  - ✓ Publications：Inui, M. and Shoaff, W. and Fausett, L. and Schneider, M., "The Recognition of Imperfect Strings Generated by Fuzzy Context Sensitive Grammars", International Journal of Fuzzy Sets and Systems, vol. 62 (1), pages 21-29, 1994.
- 「あいまい文脈依存文法によって生成された不完全な文字列の認識」（Gemini 2.0 Flashによる翻訳）
- 2016年に画像認識のCNNのモデルを独学して2019年の新人研修にてAIの基礎を新人教育に加える
- 新経済連でABEJAという会社のCTOに「AIを知りたければ**2012年のHinton博士**の論文を読め！」と言われた
- 九州熱中屋で餃子の焼き方AI判定をプロトタイプとして作成
- 2023年量子コンピュータと量子AIについて調査を始める
- 2024年5月から本格的に生成AIの勉強を始める（キッカケはLLaMA）

脳の２つのモード？

- 2+3=?、２×3=?
- 17×24=？
  - ✓(20-3)×24=20×24-24×3=480-72=408
- ファースト＆スロー



System1=Fast/System2=Slow



- 2002年にノーベル経済学賞を受賞した認知心理学者であるダニエル・カーネマン
- System1（直感的・素早い推論）=DNN（深層学習）
- System2（論理的・じっくり考える）=Thinking（強化学習を使う場合が多い）

- システム１が困難に遭遇すると、システム２が応援に駆り出され、問題解決に役立つ緻密で的確な処理を行う。システム２が動員されるのは、システム１では答を出せないような問題が発生したときである。

AIの歴史的イベント

2025.5 Google AlphaEvolve/Sakana.ai DGM

2025 AI Agent/MCP（Model Context Protocol）

2024.12 Titans（長期メモリに関する論文）

2024 Sakana.ai AI Scientist

2017 トランスフォーマモデル：Attention is all you need （Sakana.aiの創業者Llion Jones含む8名）

2016 AlphaGO:当時の囲碁の世界チャンピオンを4勝1敗で破る。

2015年末 OpenAIがNon Profitの会社として設立

2012 Alexnet:Imagenet Classification with Deep Convolutional Neural Network

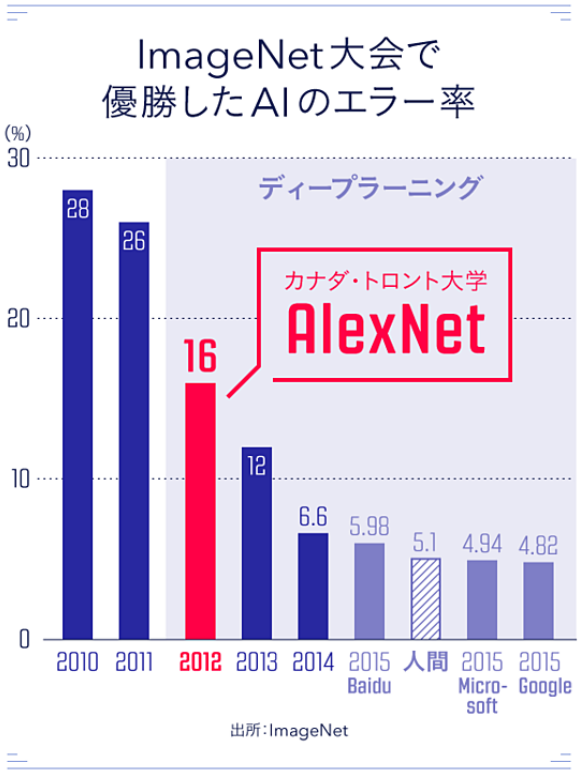
Alex Krizhevskiy, Ilya Suzkever, Geoffrey Hinton

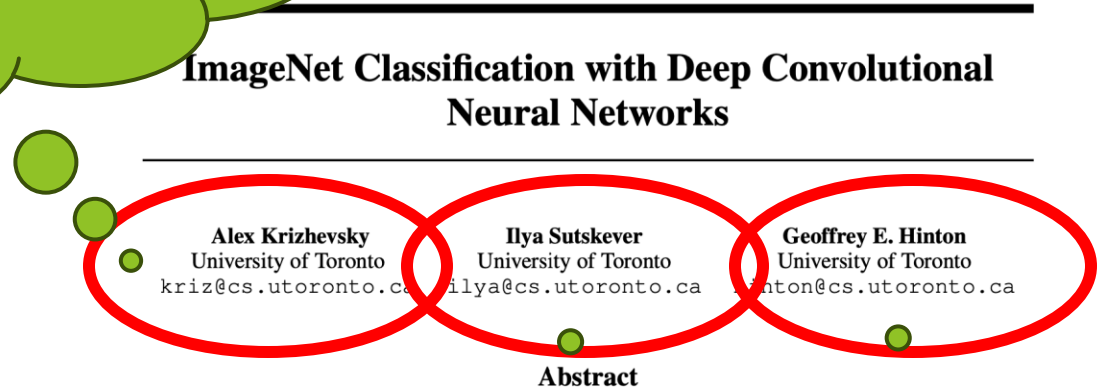
https://papers.nips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

2012年
AlexNet(2012)

AlexNetの開発者
CNNに隠れ層、GPUの計算
GoogleでAI開発

- ImageNetでの大きなジャンプアップ





ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

AlphaGOの開発
OpenAIの共同創始者

2019年にチューリング賞、2024年にノーベル物理学賞を受賞

AIの歴史的イベント

2025.5 Google AlphaEvolve/Sakana.ai DGM

2025 AI Agent/MCP（Model Context Protocol）

2024.12 Titans（長期メモリに関する論文）

2024 Sakana.ai AI Scientist

2017 トランスフォーマモデル：Attention is all you need （Sakana.aiの創業者Llion Jones含む8名）

2016 AlphaGO:当時の囲碁の世界チャンピオンを4勝1敗で破る。

2015年末 OpenAIがNon Profitの会社として設立

2012 Alexnet:Imagenet Classification with Deep Convolutional Neural Network

Alex Krizhevskiy, Ilya Suzkever, Geoffrey Hinton

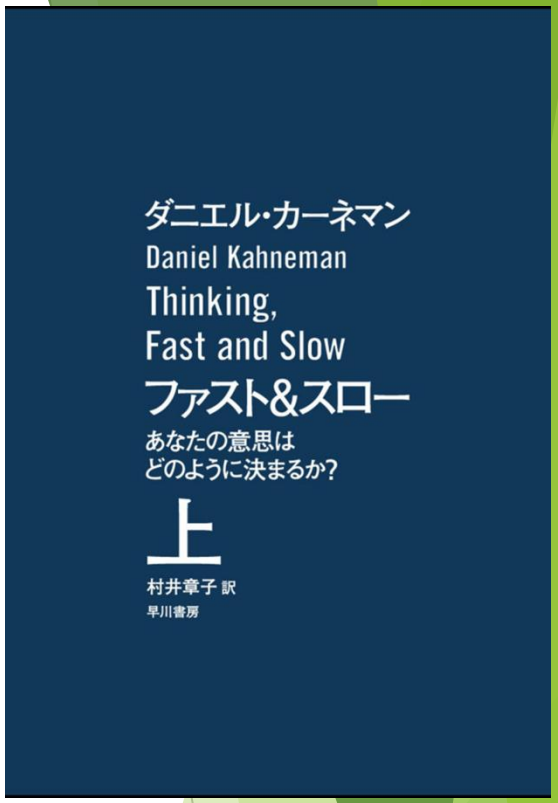
2016年
STEMモデルのブレイクスルー
Reinforcement Learning（**強化学習, Thinking**）

- 2016 年に行われた AlphaGo と李世ドル九段との対局は AlphaGoが4勝1敗で勝利
- 第 2 局で、AlphaGo が打った37番手は当時のプロ棋士の常識（**定石**）にはなかった
- 囲碁の世界ではモンテカルロ木探索（Brute Force/総当たり方式）で次の一手を推論するのは難しいというのが常識でもあった
- AlphaGo はポリシーネットワークと価値ネットワークを**強化学習**で訓練し、**モンテカルロ木探索**と組み合わせて大規模な組合せ探索空間から確率的に高いもの

組合せ探索空間が10^{170}

現在のReasoning（STEM）モデルではRLを使う

System1=Fast/System2=Slow



- 2002年にノーベル経済学賞を受賞した認知心理学者であるダニエル・カーネマン
- System1（直感的・素早い推論）=**LLM**（深層学習）
- System2（論理的・じっくり考える）=Thinking（強化学習を使う場合が多い）

- システム１が困難に遭遇すると、システム２が応援に駆り出され、問題解決に役立つ緻密で的確な処理を行う。システム２が動員されるのは、システム１では答を出せないような問題が発生したときである。

AIの歴史的イベント

2025.5 Google AlphaEvolve/Sakana.ai DGM

2025 AI Agent/MCP（Model Context Protocol）

2024.12 Titans（長期メモリに関する論文）

2024 Sakana.ai AI Scientist

2017 トランスフォーマモデル：Attention is all you need （Sakana.aiの創業者Llion Jones含む8名）

2016 AlphaGO:当時の囲碁の世界チャンピオンを4勝1敗で破る。

2015年末 OpenAIがNon Profitの会社として設立

2012 Alexnet:Imagenet Classification with Deep Convolutional Neural Network

Alex Krizhevskiy, Ilya Suzkever, Geoffrey Hinton

https://youtu.be/6tcjwdanedU?si=Ms4qpi2TBCWDccja
京大博士のAI解説

2017年
トランスフォーマモデル

"**Attention Is All You Need**" is a groundbreaking research paper published in 2017 that revolutionized the field of natural language processing (NLP)

Transformer（生成AI）：文章などを生成する

生成AIのモデルにおいて、**パラメータ数**は**Attention**の計算量と密接に関係しています。**パラメータ数**を増やすことで表現能力を向上させることができますが、計算量や過学習のリスクも考慮する必要があります。

“Context Window”=人間でいう「ショートメモリ」

Attentionとは：注視する語句の重要性と関連性をベクトル化するモデル

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Sakana.aiの創業者

“Attention”

山口が県名であると解釈した場合

- 先週、山口と広島に行きました。
- この文章を英語にしてください。

"Last week, I went to Yamaguchi and Hiroshima."

山口が友達であることを理解している場合

"Last week, I went to Hiroshima with my friend, Yamaguchi."

Context Window

私の友達の山口とは、中学生からよく遊んでいたんですよね。
…

「先週、山口と広島に行きました。」
この文章を英語にしてください。

Context Windowの役割

例：３時間の会議の議事録

３時間分のコンテキストウィンドウ

- 今日の会議の議題は
  - ✓ 今月の収益予測
  - ✓ 重要案件のリストアップ
  - ✓ ….ほにゃらら！

…

- まとめ
  - ✓ 今月の収益予測は5000万です ← いつの収益？今年度？今月？
  - ✓ 重要案件は ABC社、XYZ社 ← 会社名の意味は？

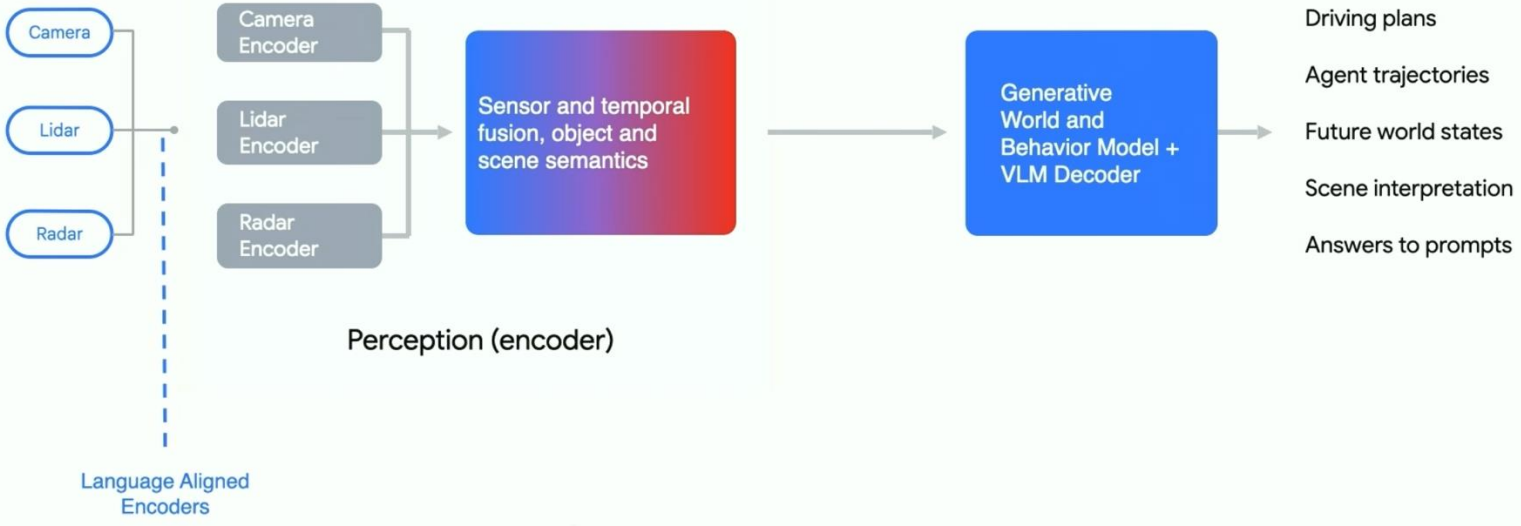
Context Windowが小さいとハルシネーションが発生する

自動運転車へのDNN、Thinking、Attention機構の応用

- Waymoの事例

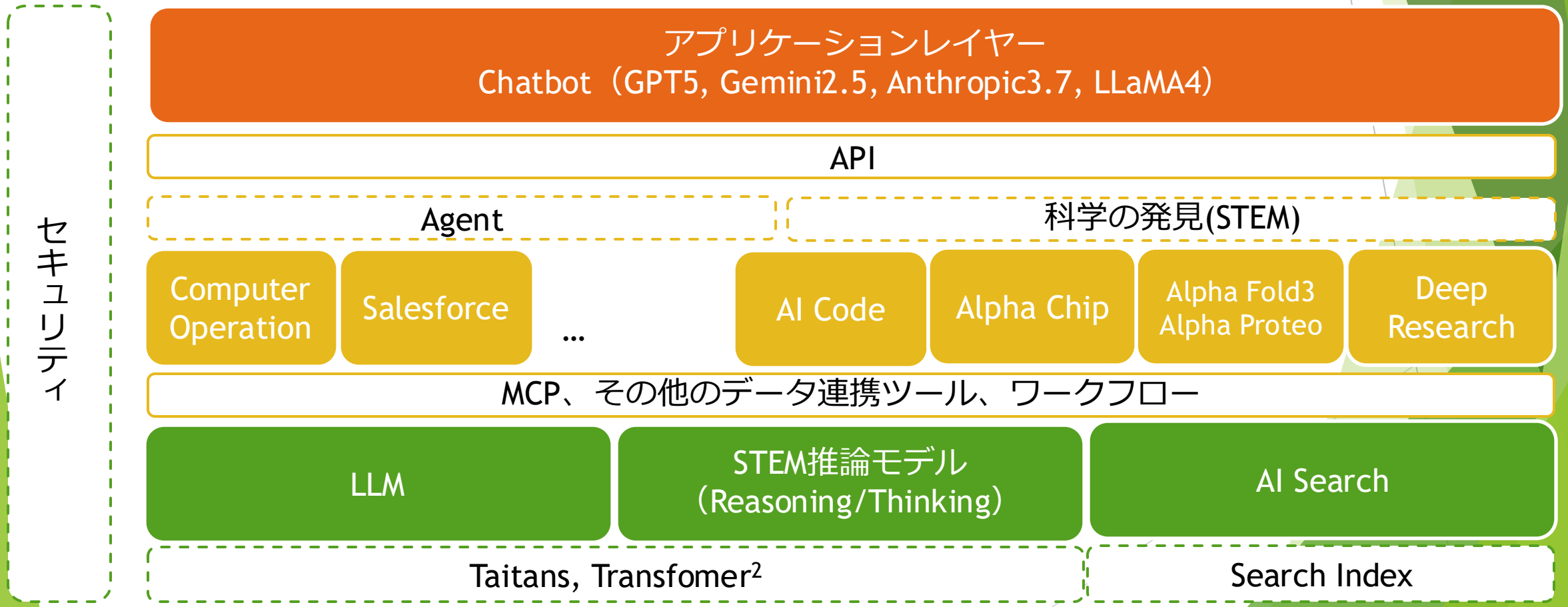
The Waymo Foundation Model: Combining AV-specific advances with general world knowledge of VLMs

Waymo Driver's superhuman sensing ability



road. So ima
our path. Thi

モデルまとめ（ドラフト）


セキュリティ
アプリケーションレイヤー
Chatbot（GPT5, Gemini2.5, Anthropic3.7, LLaMA4）
API
Agent
科学の発見(STEM)
Computer Operation
Salesforce
…
AI Code
Alpha Chip
Alpha Fold3
Alpha Proteo
Deep Research
MCP、その他のデータ連携ツール、ワークフロー
LLM
STEM推論モデル
（Reasoning/Thinking）
AI Search
Taitans, Transfomer²
Search Index

次世代AIの学習方法と分類

基本的にはパラメータの修正を伴う

事前学習 Pre-Training

- 一般的にはスケール則で学習
- 大規模な教師ありデータセットを使うのが一般的
- 学習期間１，2ヶ月
- 幅広い知識の学習
- データ品質が重要
- パラメータは多い
- Knowledge Cutoff迄の情報

事後学習 Post-Training

Fine-Tuning

- 教師あり（SFT）
- 専門知識の学習
- Instruction Tuning
- Few shot Fine-Tuning
- 言語の学習

蒸留 Distillation

- 教師モデルの出力を生徒モデルの入力として学習
- モデルの軽量化
- 知識の効率的な転移
- 教師モデルのノイズの平準化と生徒モデルの個性の活用

強化学習

- 報酬システム

ICL In Context Learning

- 推論時の学習
- 一般的にはコンテキストウィンドウを使って推論をする（中長期メモリ）

強化学習

- 報酬システム

人間で言うと小・中・高の学習方法

学士～修士課程の学習

Phd.やOn the Job Training

AIの歴史的イベント

2025.5 Google AlphaEvolve/Sakana.ai DGM

2025 AI Agent/MCP（Model Context Protocol）

2024.12 Titans（長期メモリに関する論文）

2024 Sakana.ai AI Scientist

2017 トランスフォーマモデル：Attention is all you need （Sakana.aiの創業者Llion Jones含む8名）

2016 AlphaGO:当時の囲碁の世界チャンピオンを4勝1敗で破る。

2015年末 OpenAIがNon Profitの会社として設立

2012 Alexnet:Imagenet Classification with Deep Convolutional Neural Network

Alex Krizhevskiy, Ilya Suzkever, Geoffrey Hinton

AI Agentとは

System1=DNN（直感的・素早い推論）

✚

System2=Thinking（論理的・じっくり考える）

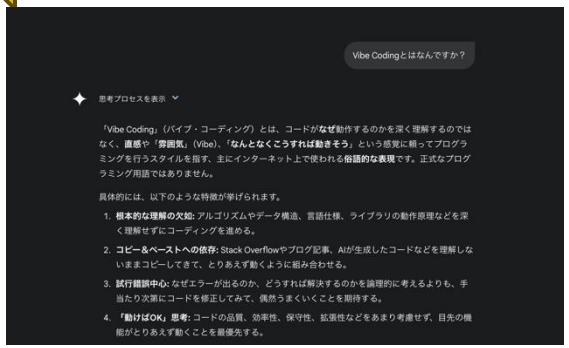
- DeepResearch
- NotebookLM（Google）
- AI Agent for SWE（Software Engineering）
- AlphaEvolve（アルゴリズムを改善する）

AgentとはAI以外のシステムと連携しつつ人間と**協力して長期タスク（業務）**をこなす自動化ソフトウェアである

DeepResearch

Chatbot
入力画面

「Vibe Codingとはなんですか？」



「Vibe CodingをAgile開発やNo Codeの観点で調査をしてレポートを作成してください」

Deep
Research
入力画面

「こういう調査方法で調べますがよろしいですか？」

Research Agent

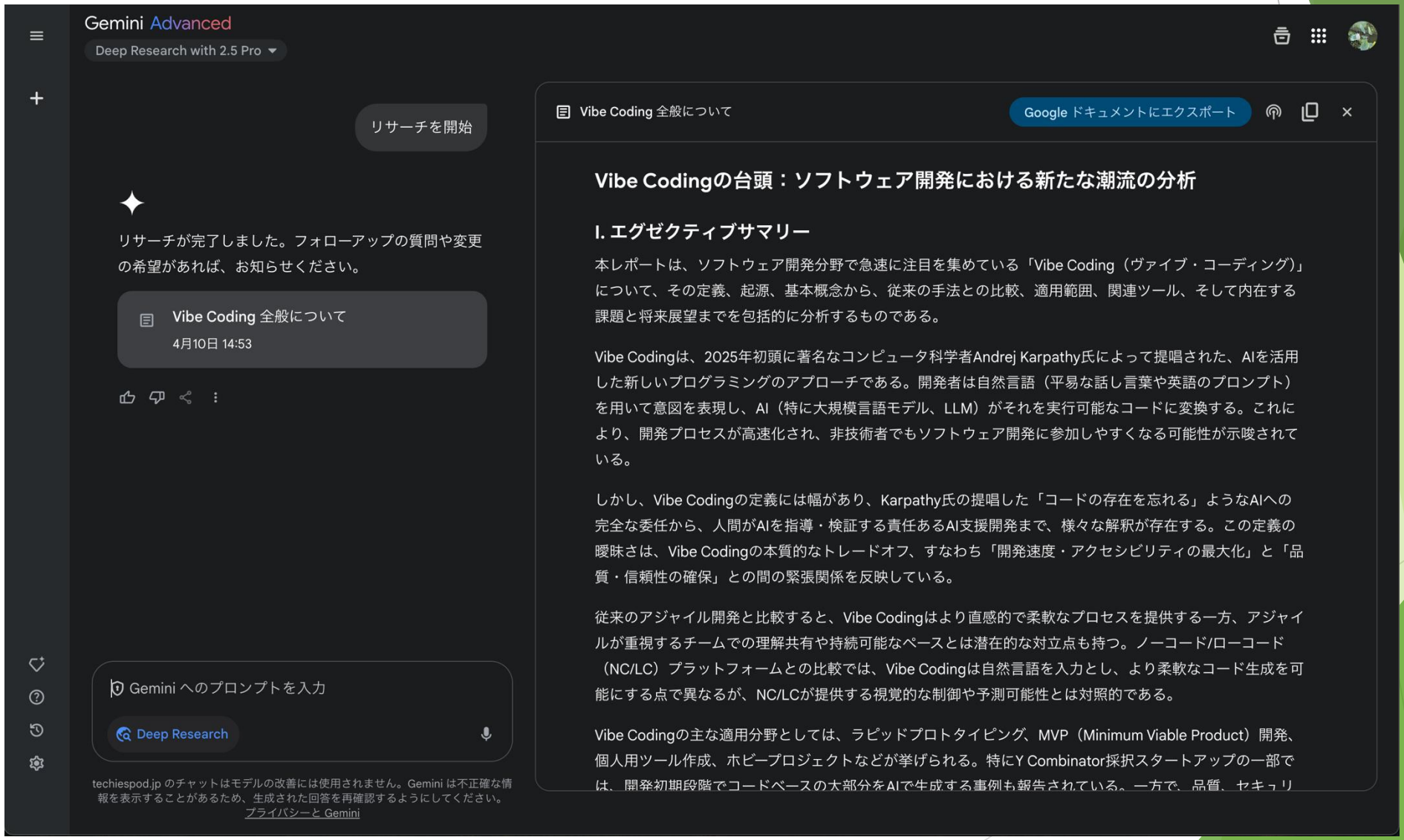
Deep
Research
入力画面

プロンプトエンジニアリング

AI Searchを使って参考文献を検索

文章化をするAI

DeepResearch



AIの歴史的イベント

2025.5 Google AlphaEvolve/Sakana.ai DGM

2025 AI Agent/MCP（Model Context Protocol）

2024.12 Titans（長期メモリに関する論文）

2024 Sakana.ai AI Scientist

2017 トランスフォーマモデル：Attention is all you need （Sakana.aiの創業者Llion Jones含む8名）

2016 AlphaGO:当時の囲碁の世界チャンピオンを4勝1敗で破る。

2015年末 OpenAIがNon Profitの会社として設立

2012 Alexnet:Imagenet Classification with Deep Convolutional Neural Network

Alex Krizhevskiy, Ilya Suzkever, Geoffrey Hinton

アルゴリズム改善や発見をするエージェント（進化的アルゴリズム）

- Google AlphaEvolve（AI Agent）
  - ✓ 56年間破られなかった4×4行列の乗算アルゴリズムを新しく刷新した（49回の計算→48回）
- データセンターオーケストレーションシステムの刷新
- Gemini学習における行列演算の新アルゴリズムの発見
- FlashAttentionカーネル実装の高速化（32.5%）
- 数学やコンピュータサイエンス問題の新しい発見

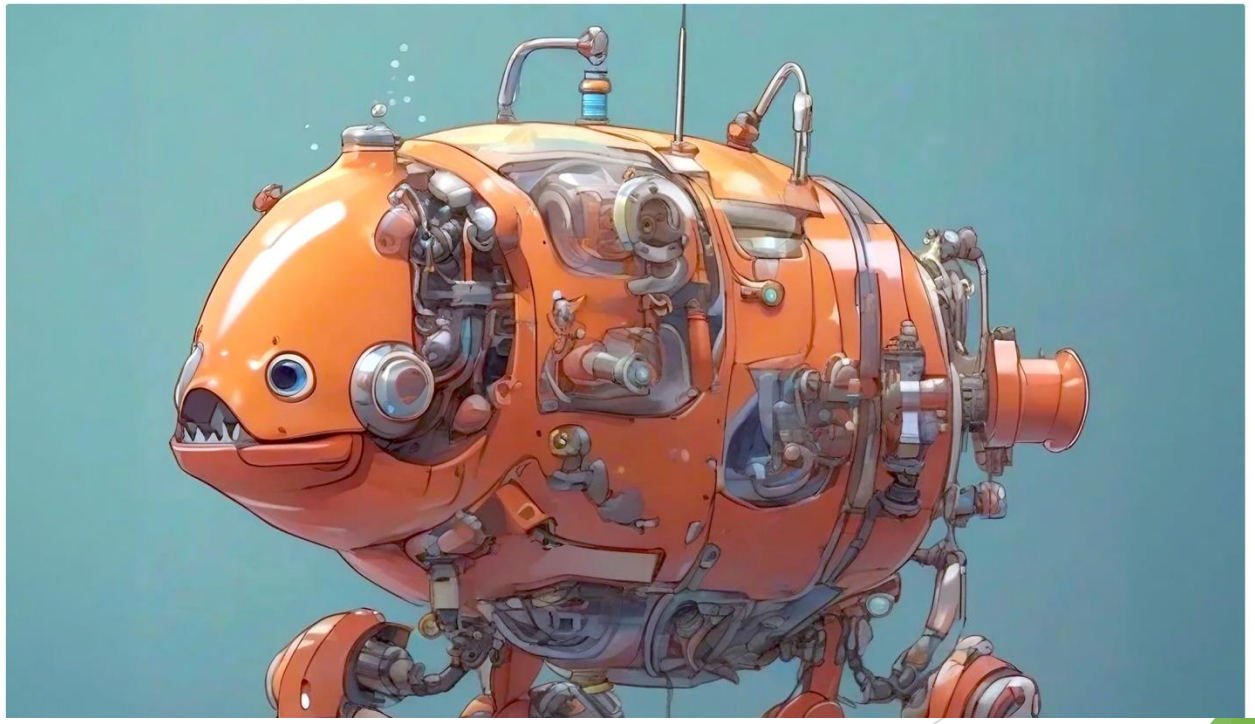
AI Agentにも自律的に改善を繰り返し、改善や新しい発見をするものも存在する

DGM（ダーウィン・ゲーデル・マシン）

sakana.ai

自らのコードを書き換え自己改善するAI：「ダーウィン・ゲーデルマシン」（DGM）の提案

May 30, 2025



AI Agentの例

- UI:チャットやアプリ
- AIが一連の流れを取りまとめる
- 様々なシステムと連携するツール

AI Agent for Sales Assistant

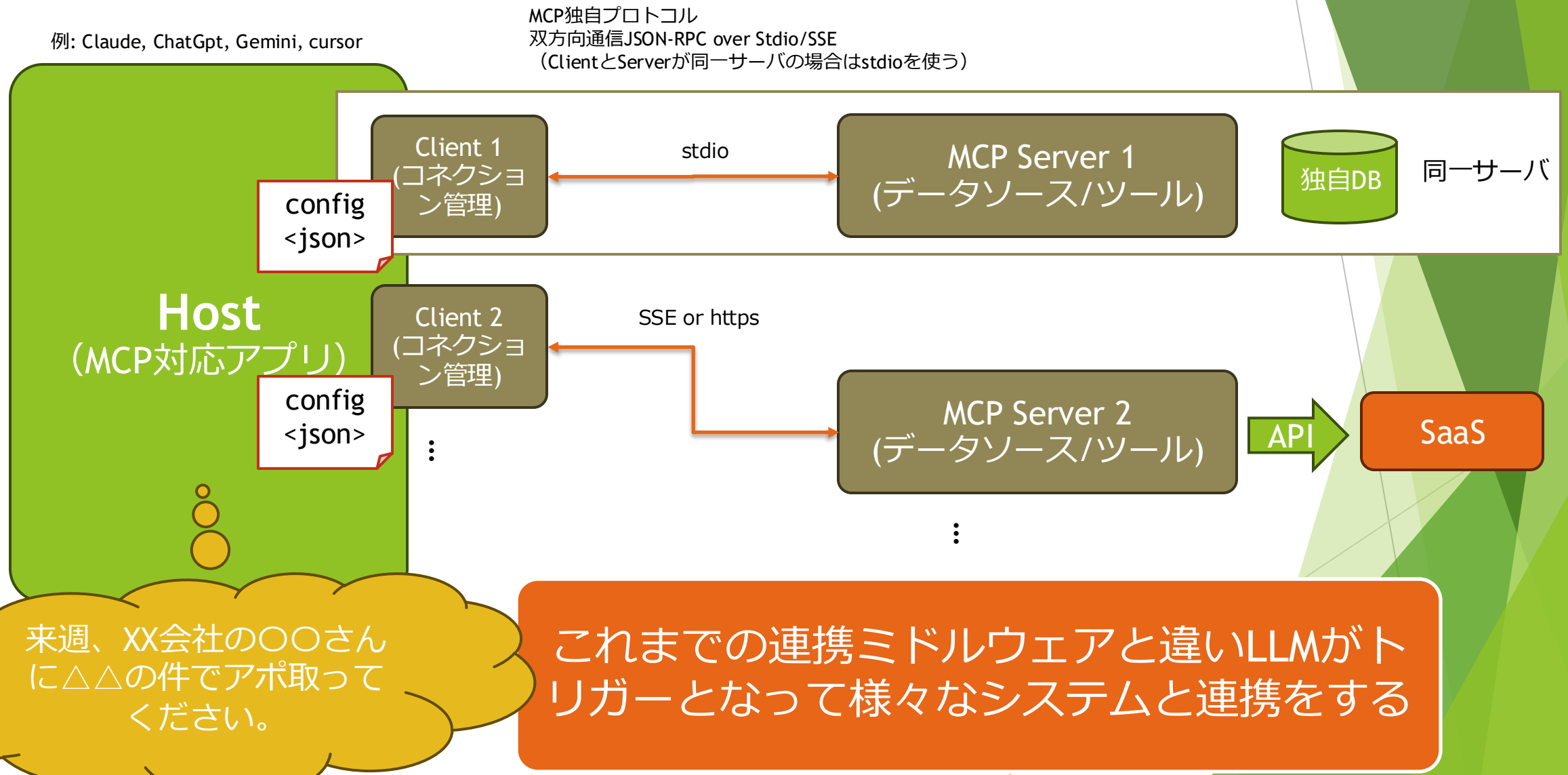
① SlackでAgentにアポの依頼をする
② CRMからアポを取る相手を探す
③ スケジュールを確認し、メールの下書きをする
④ メールの下書きとアポをとる相手の確認をSlackに返信する
⑤ 営業からOKが出たら、メールを自動で流す
⑥ 相手からの返事を待って確定する

API＆セキュリティ

様々なシステムと連携するツール

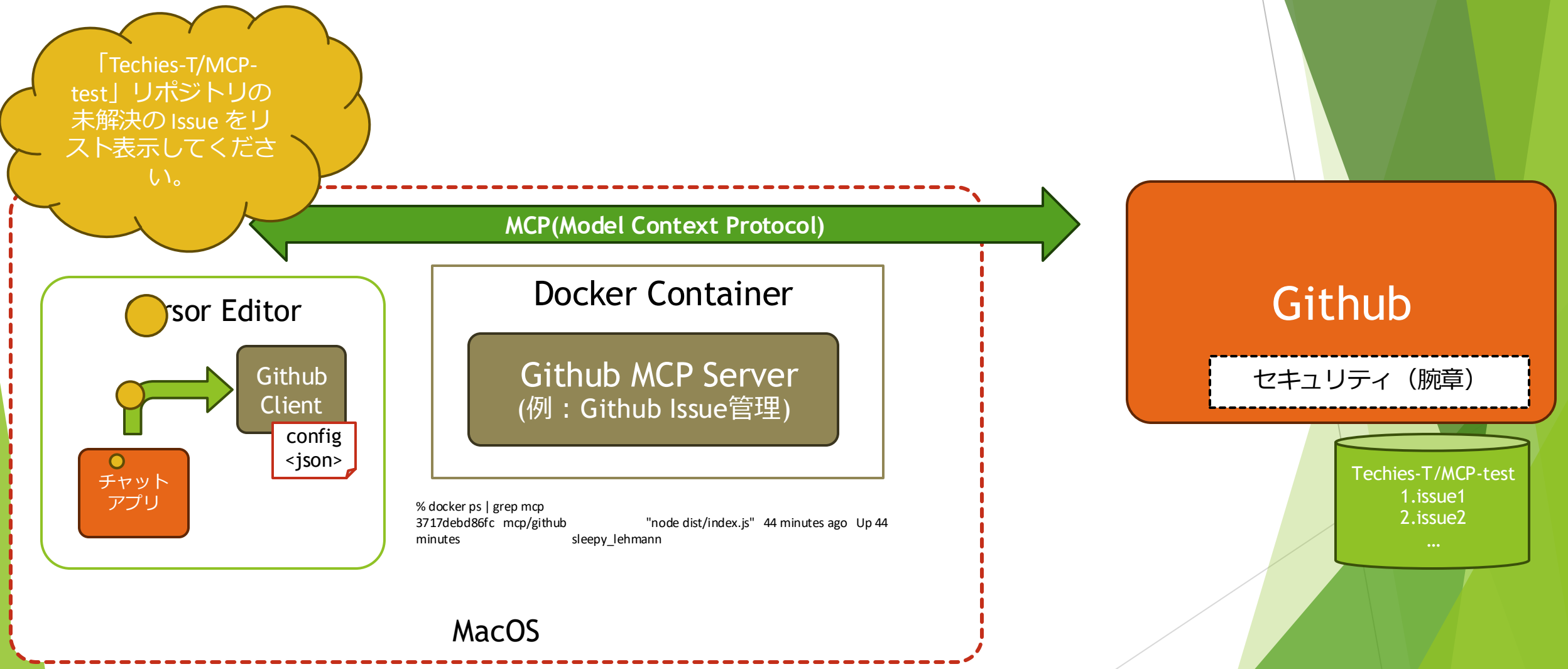
- Anthropic MCP（Model Context Protocol
- Bindit（ユニリタ）
- Zapier

MCPアーキテクチャ図



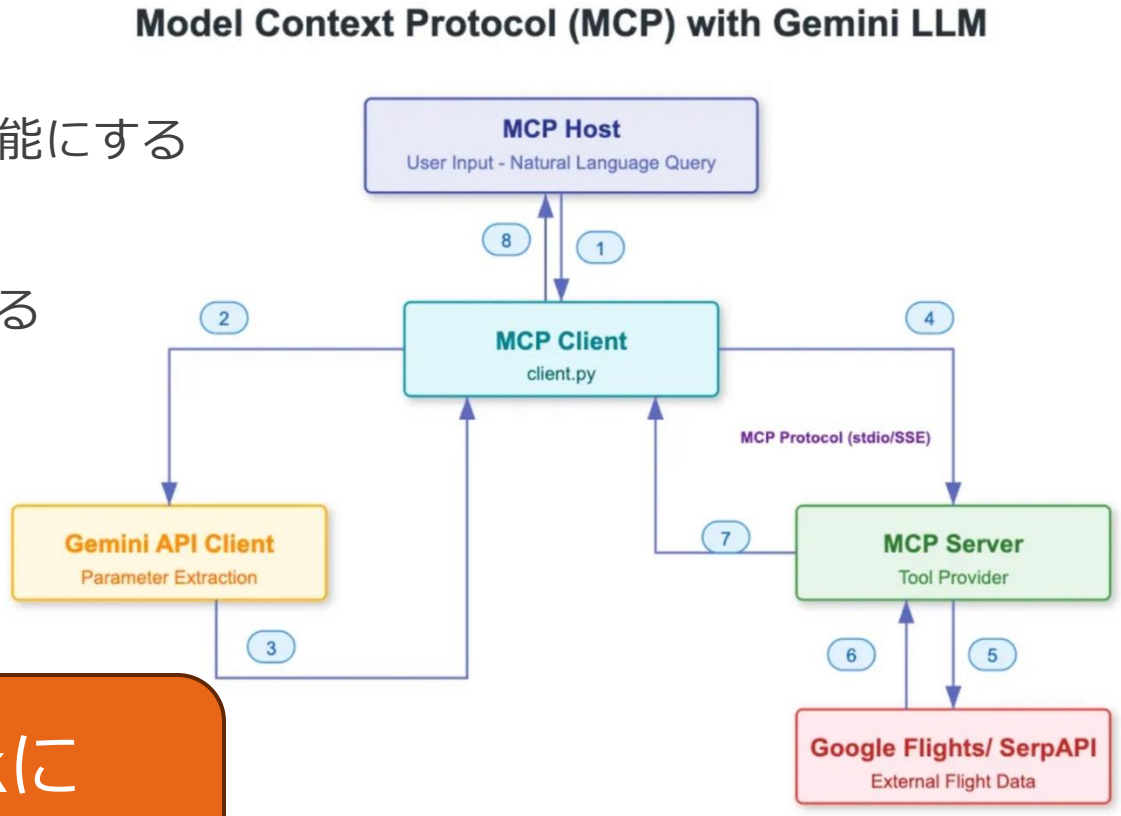
GithubをMCPで連携する

Gitbub+MCP（このPCでの環境）



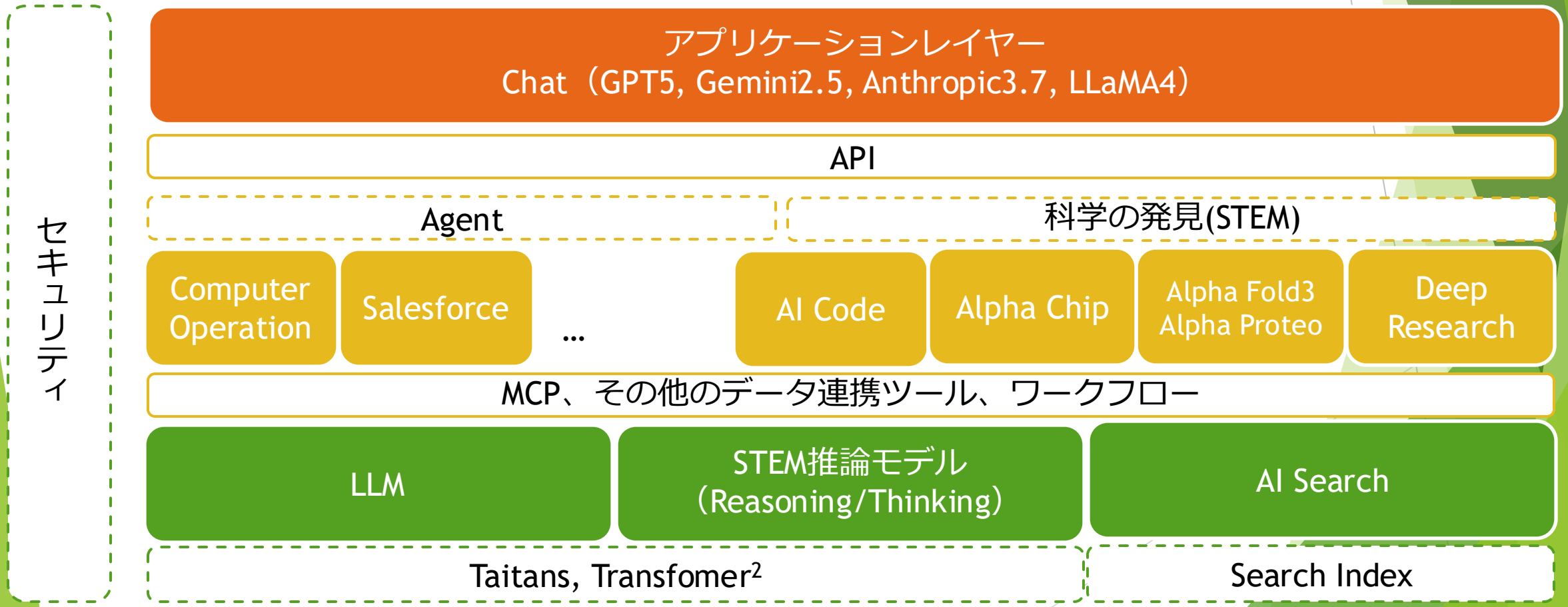
Gemini+MCP

- Agentic Systemsを作るときに利用する
- ツールをモジュール化、再利用可能,、発見可能にする
- 複数の外部リソースを使える
- 複数のツールの連携を可能にしてスケールする



Google I/O 2025でGenAIのsdkに
MCPの機能を追加したと発表

モデルまとめ（ドラフト）

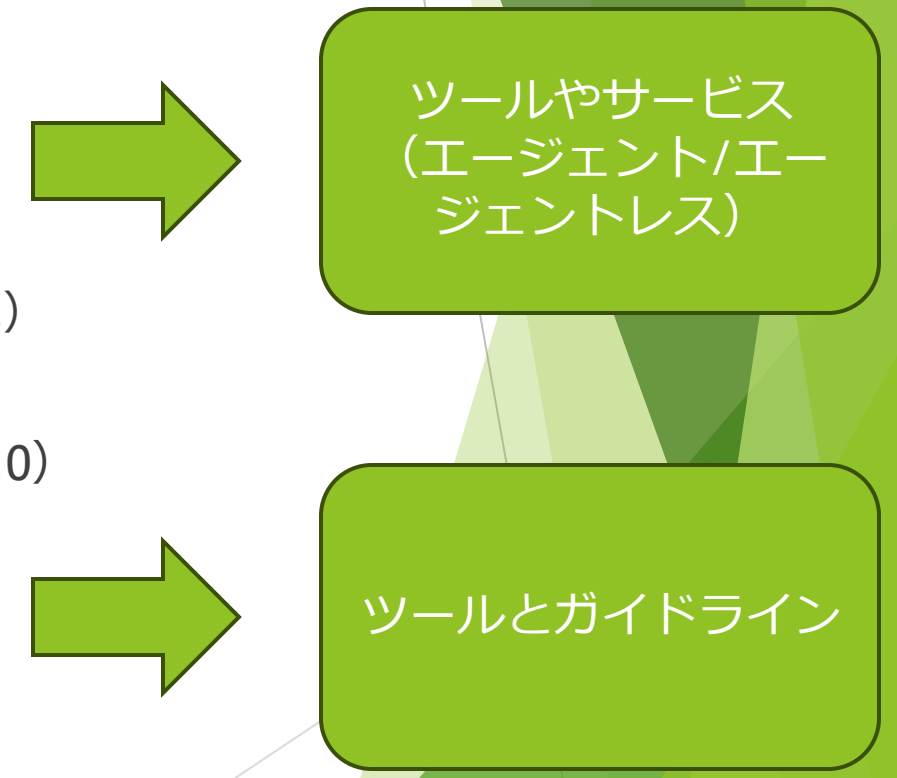

セキュリティ
アプリケーションレイヤー
Chat（GPT5, Gemini2.5, Anthropic3.7, LLaMA4）
API
Agent
科学の発見(STEM)
Computer Operation
Salesforce
…
AI Code
Alpha Chip
Alpha Fold3
Alpha Proteo
Deep Research
MCP、その他のデータ連携ツール、ワークフロー
LLM
STEM推論モデル
（Reasoning/Thinking）
AI Search
Taitans, Transfomer²
Search Index

生成AIの今後

- 基本モデルの進化
  - ✓ Thinking（STEM/Reasoning）モデルの進化
  - ✓ 推論時の計算アルゴリズムの開発（AIモデルを使って半導体チップの設計やアルゴリズムの効率化）
  - ✓ On device AI（Humanoid, AI Asistant）
- 生成AIにまつわるセキュリティ
  - ✓ AIを悪用する人からのセキュリティ（Slopsquatting）
  - ✓ AIの実行環境のセキュリティ（MCPのセキュリティ）

WGでの今後の焦点
AIの発展でセキュリティはどう変わるべきか？

- AIを悪用する人からの攻撃を排除するセキュリティ
  - ✓ Computer Operation（Computer Use） Agent
  - ✓ GoogleのProject Astera（倫理問題）
  - ✓ AI AgentのCertification/証明書
  - ✓ Slopsquatting（AIが自動生成するパッケージ名称を予測して攻撃）
- 生成AIの実行環境とセキュリティ
  - ✓ MCPでアクセスするサイトのセキュリティ（API Key, OAuth2.0）
  - ✓ アラインメント/ガードレール
  - ✓ プロンプトインジェクション
  - ✓ APIセキュリティ
  - ✓ モデルのJailBreak
  - ✓ AI Agent又はAIが開発したプログラムの脆弱性チェック



これはもはや、未来の物語ではない。
AIは日々、この瞬間も、自らを進化させている。
その驚異的な加速は、AI自身の力によるものなのだ。

To be Continue…

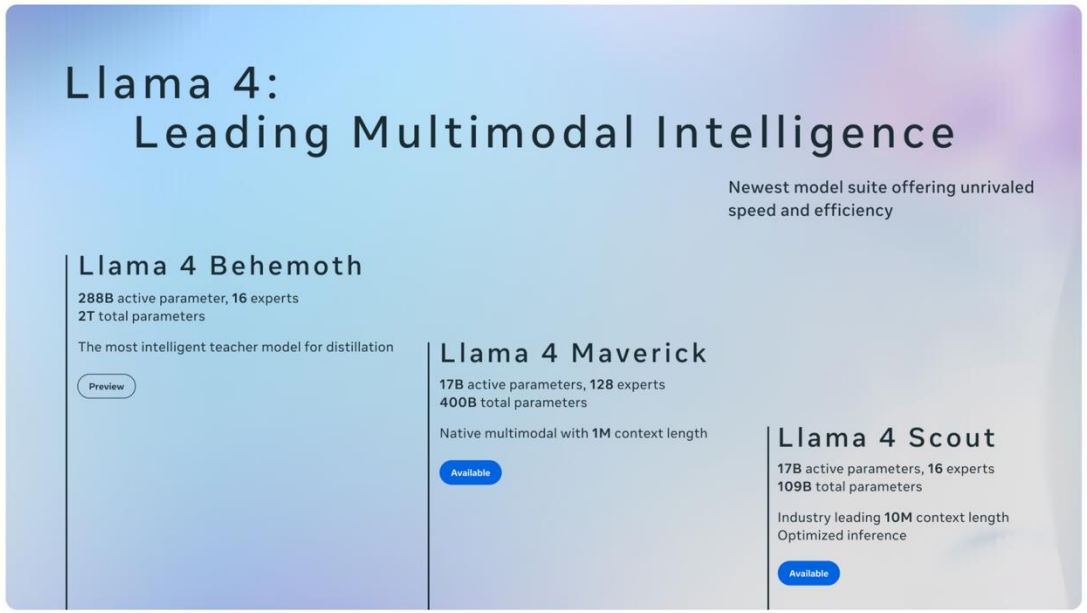
Minoru Inui, Founder & CEO Techiespod

添付資料

AI Agent for SWE（SoftWare Engineering）

- ソフトウェアの開発の自動化
- 例：Devin
- ソフトウェアの開発を自動化するための課題
  - ✓ Context Length（window）
  - ✓ 永続化メモリ
  - ✓ Slopsquatting等の脆弱性対策
  - ✓ エディタのProject Ruleなどの利用
  - ✓ MCPやbinditの様なツール

LLaMA 4



- ３つのモデルを発表
- MoE（Mixture of Experts）アーキテクチャ
- LLaMA 4 Behemoth, Maverick, Scout
  - ✓ Behemoth 2T (288B Active), 16Experts
  - ✓ Maverick 400B(17B Active), 128 Experts
  - ✓ Scout 109B (17B Active), 16 Experts、10M Context Window
- Behemothはトレーニング中（モンスタモデル/蒸留の教師）
- Resoningモデルは別途リリースすると発表

LLaMAがContext Lengthを拡大した理由

- AI Agent for SWEの為
- Codingタスクを全てのモジュールに拡大
- MetaConでMSのCEOと開発のプロセスに関して対談

AIの攻撃手段の例

- Prompt Injection
- Slopsquatting（Typosquattingの応用）
  - ✓ 悪意のある攻撃者がAIの手抜きを悪用する
- Computer Use（AIの画像認識機能を使ってRPA）
  - ✓ We're bringing Project Mariner's computer use capabilities into the Gemini API and Vertex AI. Companies like Automation Anywhere, UiPath, Browserbase, Autotab, The Interaction Company and Cartwheel are exploring its potential, and we're excited to roll it out more broadly for developers to experiment with this summer. By Google I/O 2025

Vibe Codingの広がりと脅威

- Vibe Codingとは、AIの柔軟性を活用して自然言語を使ってプログラミングをする方法
- Vibe Codingには根強い支持者たちがいる
- 一方で、プログラミング言語についてよく知らない人たちがコーディングを行うことをターゲットとした悪意ある人たちがこれにつけいる攻撃を行う

Slopsquatting

- AIは利用者の要求に対して忠実に答えるために様々なイレギュラーなことをすることが知られている。これは**ハルシネーション**の一種として考えられている。
- 例えば「XXのプログラムを作って」という要求に対して、よく使われるオフィシャルなパッケージを利用せずに、AIが「手抜き」パッケージを作成してレポジトリに登録をし利用するケースがある
- 攻撃者はタイプミス（Typosquatting）を推測し、予測可能なパッケージ名称を使いAIの手抜きを悪用する。
- これをSlopsquattingと呼ばれている。

Slopsquattingの事例

具体的な事例として、「huggingface-cli」のケースが挙げられる。この事例では、LLMがこのパッケージ名をハルシネーションによって生成した。

ある研究者がこのハルシネーションされた名前で空のパッケージをアップロードしたところ、3ヶ月で3万回以上ダウンロードされたという 。

この事例は、Slopsquattingの現実的な脅威と、開発者がいかに容易に罠にはまるかを示している。