AIのパワーを活用する

コアピラー、実用的なアプリケーション、責任のある実装

クナル アナンド
チーフテクノロジーオフィサー

アジェンダ

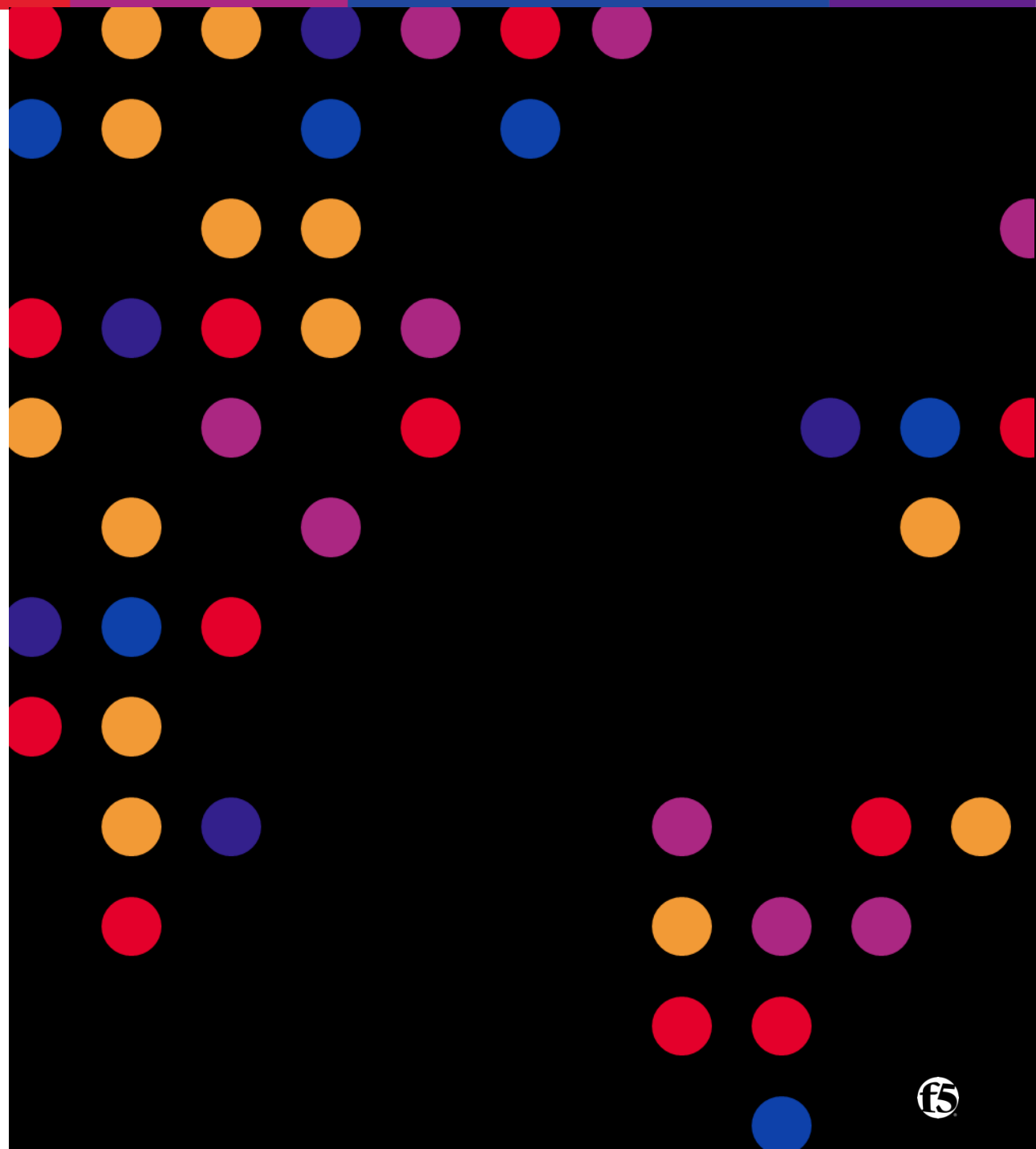
AIの歴史

Transformerとは

モデルの学習と推論

責任ある実装

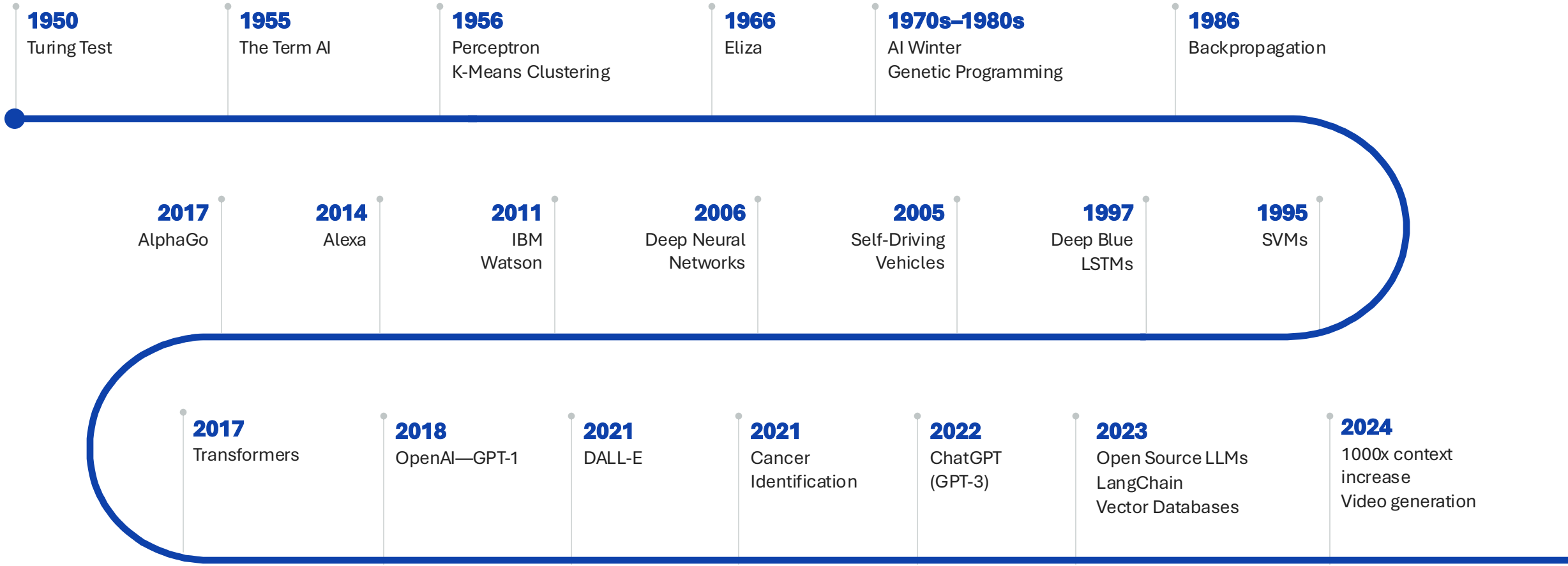
AIの次のステップは？



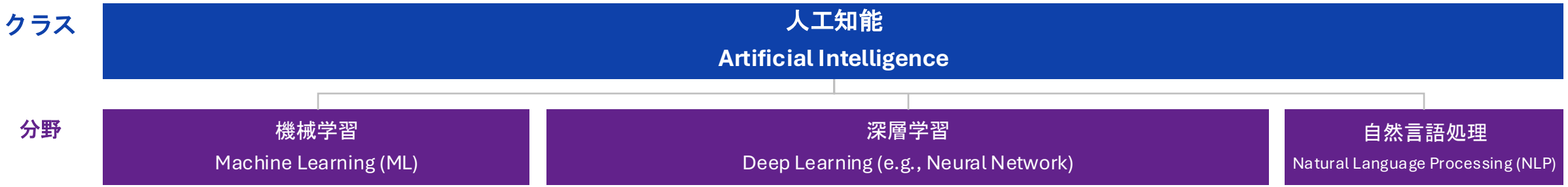
AIの歴史

AIはこの数十年で進化したが、アプリケーションによる活用は限定されたユースケースに留まっている

AIの歴史

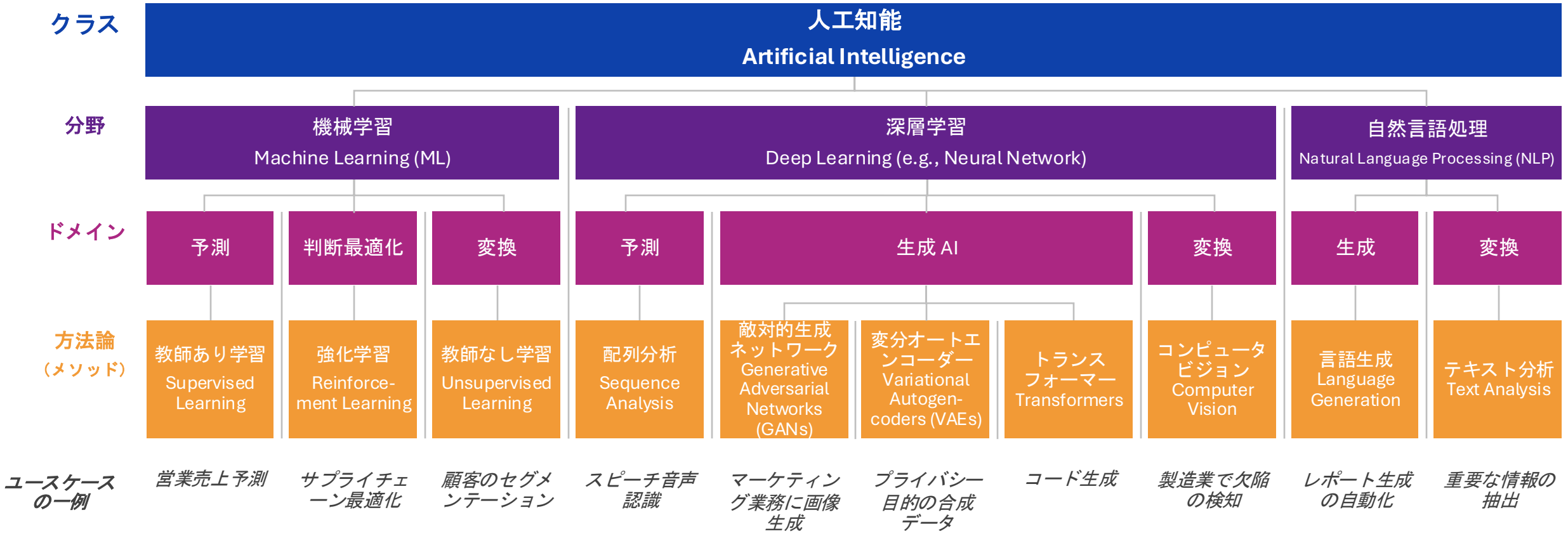


AIには複数の分野が存在する



Source: "Artificial Intelligence: A Modern Approach" by Stuart Russell and Peter Norvig (2021)

AIには複数の分野が存在する



Source: "Artificial Intelligence: A Modern Approach" by Stuart Russell and Peter Norvig (2021)

重要なブレイクスルーの起点となったのは、Googleが**Transformer**に関する先進的な資料を2017年に公開したタイミングである

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 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

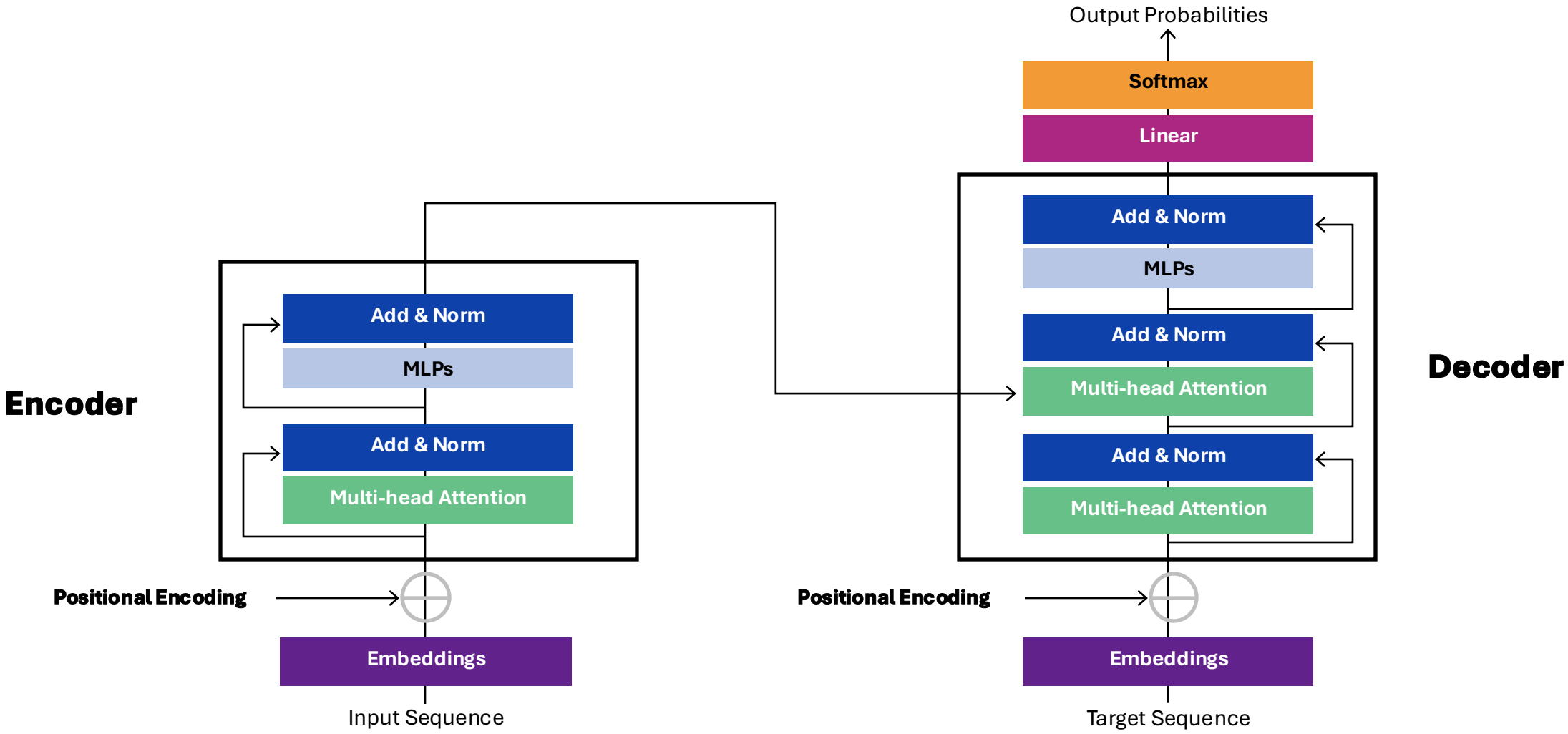
1 Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].

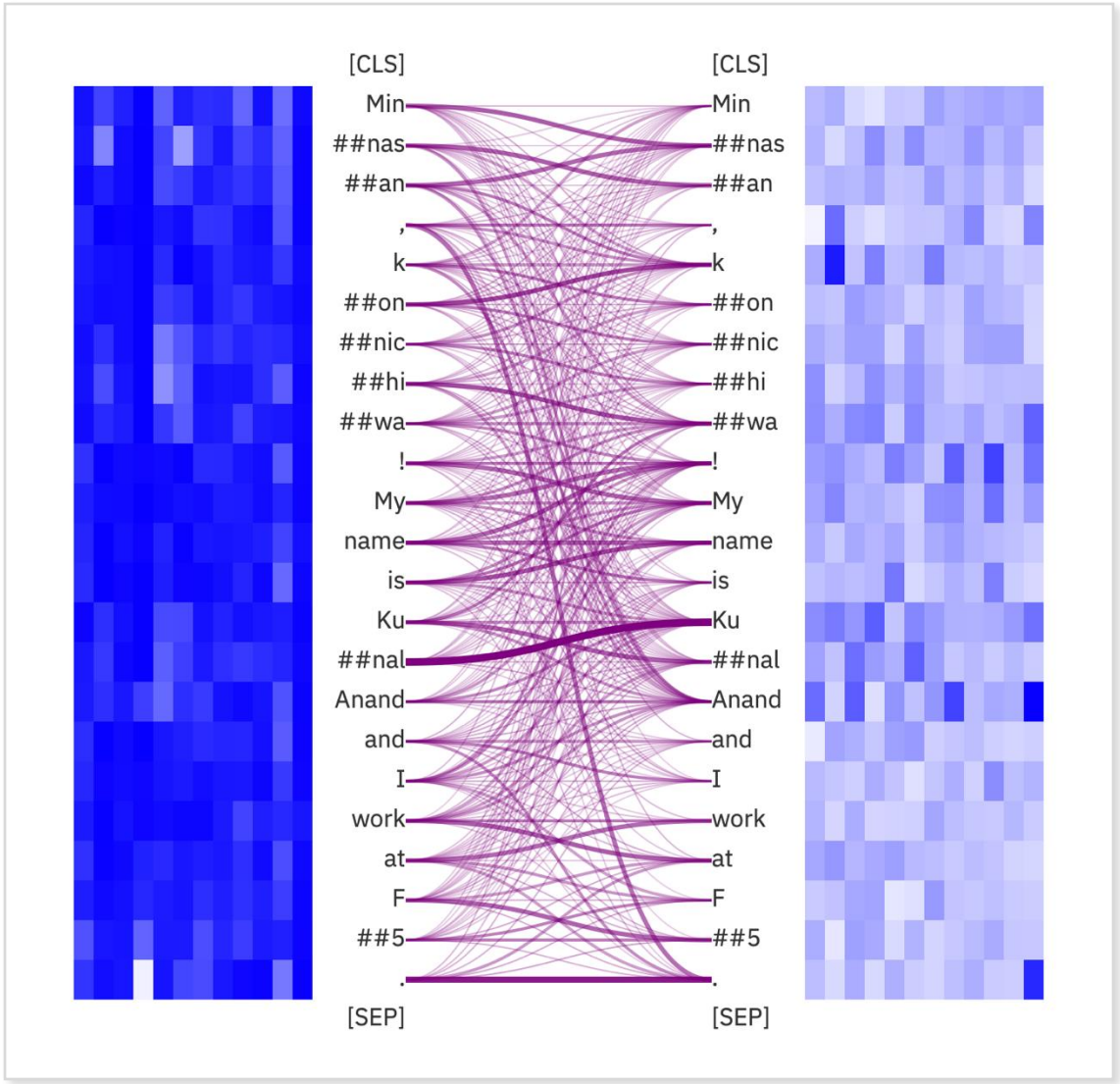
*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every

Transformerとは

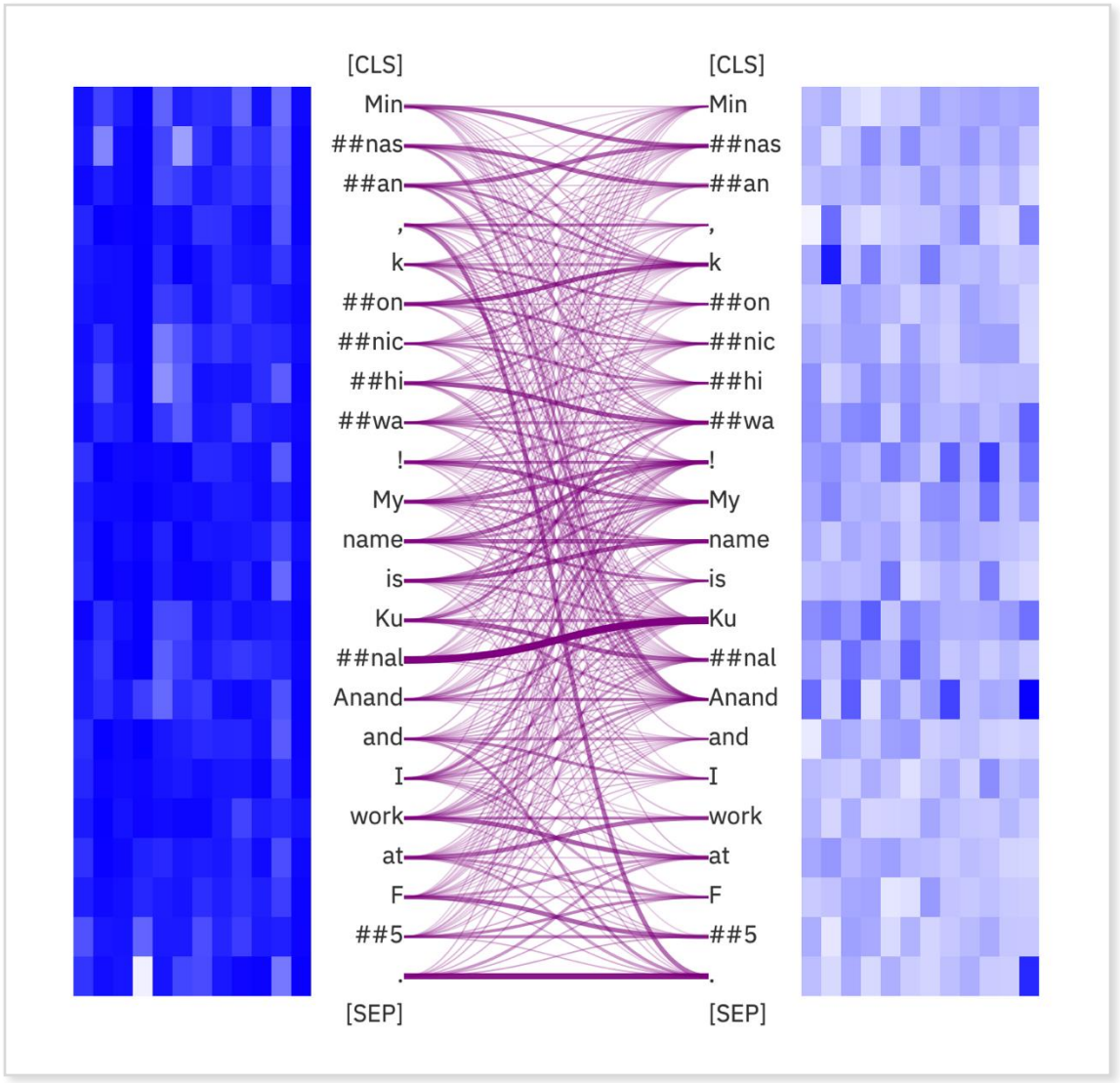
Transformerとは、多層的なAttentionを活用してインプットとアウトプット生成処理を行うニューラルネットワークのアーキテクチャである



“Minnasan, konnichiwa! My name is Kunal Anand and I work at F5.”



Attentionによって
ニューラルネットワークは
「意味のあるインプットデータ」のみに
対応することができる



“Minnasan, konnichiwa! My name is Kunal Anand and I work at F5.

I'm excited to be here today to talk about web application security and how we can protect our digital assets from evolving cyber threats. At F5, we specialize in application delivery and security solutions, and I'm looking forward to sharing some insights from our experience in the field.”

Transformerが機能する流れを喩えるならば

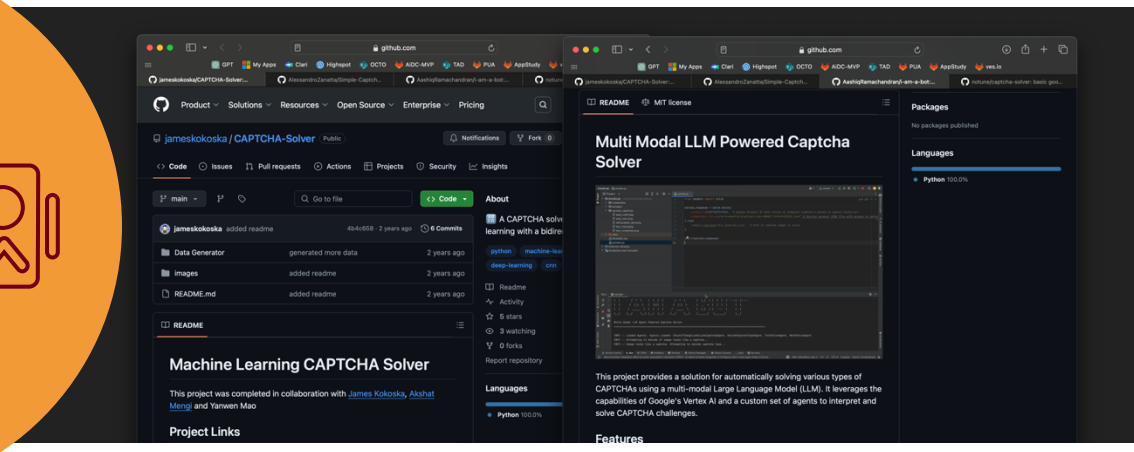
1. ブロックを見て、
   色と形を認識する

2. どのブロック同士が
   より良い組み合わせかを考える

3. ブロックを組み合わせて
   何かを作り出す

テキスト、画像、音声、動画、時系列、遺伝子、化学、地理空間、センサー、ネットワーク、金融、医療、科学機器、製造業、教育分野、環境、振る舞い、、、

Like any other technology, transformers can be used to engineer both good and bad outcomes



バイオテクノロジー企業

創薬やAIを活用したパーソナライズされた治療

数千人の社員を活用して数十もの新薬を市場に展開

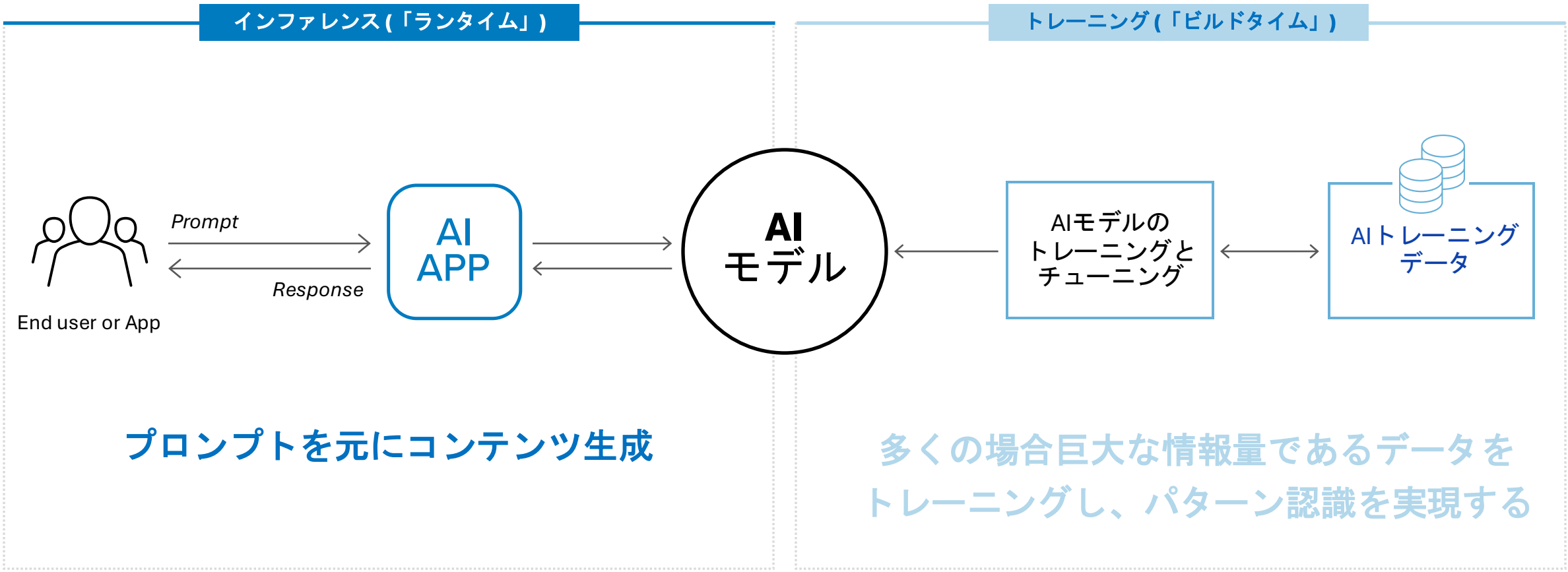
悪意のあるボット

キャプチャ認証をAIを使って突破

一般的なオープンソースのAIを使って
キャプチャ認証を数秒で突破

モデルのトレーニング（学習）とインファレンス（推論）

AIの重要な2つの手順:トレーニング（学習）とインファレンス（推論）



トレーニングにおいて特筆すべき側面:非常に大量なデータ量とマルチモーダルなコンテンツ

1

基礎的な生成AIデータは
データ量が大量である[1]

	パラメータの数
GPT-4o	1.8 兆
Gemini	1.6 兆
Claude	1750億

2

生成AIトレーニングデータは
マルチモーダルとなる

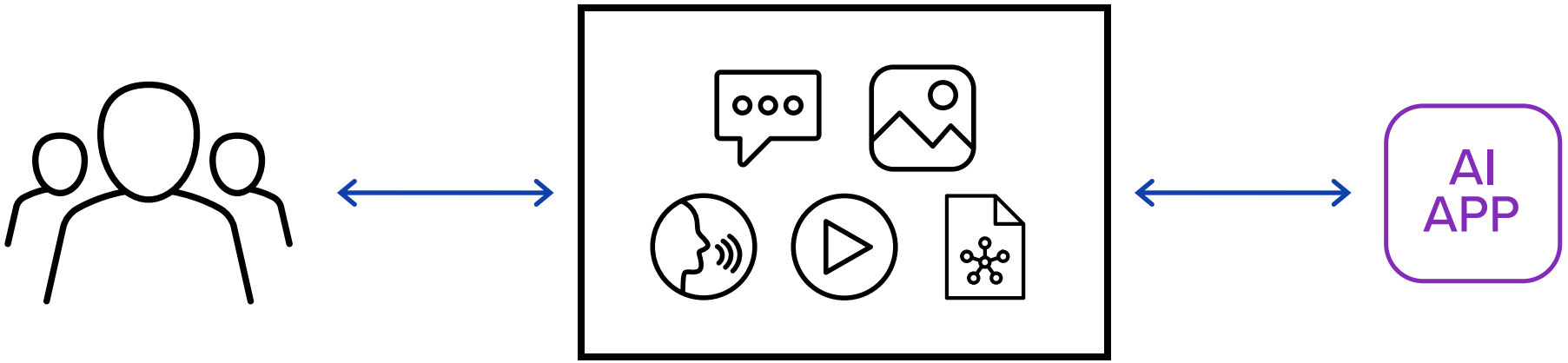
- テキスト
- 画像
- 音声
- 動画
- 構造化データ

Notes: 1 Chat GPT 4 was trained on the “whole internet”, Chat GPT 5 may need to expand to synthetic data
Source: deepchecks.com LLM model comparison

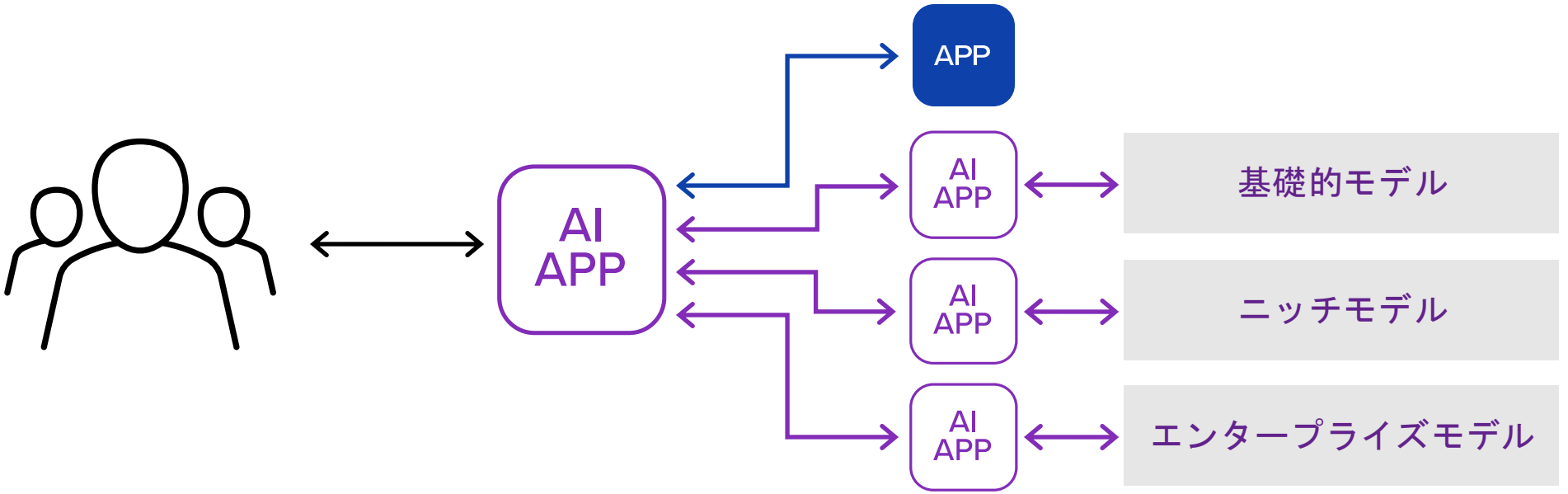
学習における

トレーニング → チューニング → 再トレーニング

AIアプリケーションの体験はマルチモーダルになる：映像、テキスト、画像、音声、ビデオ

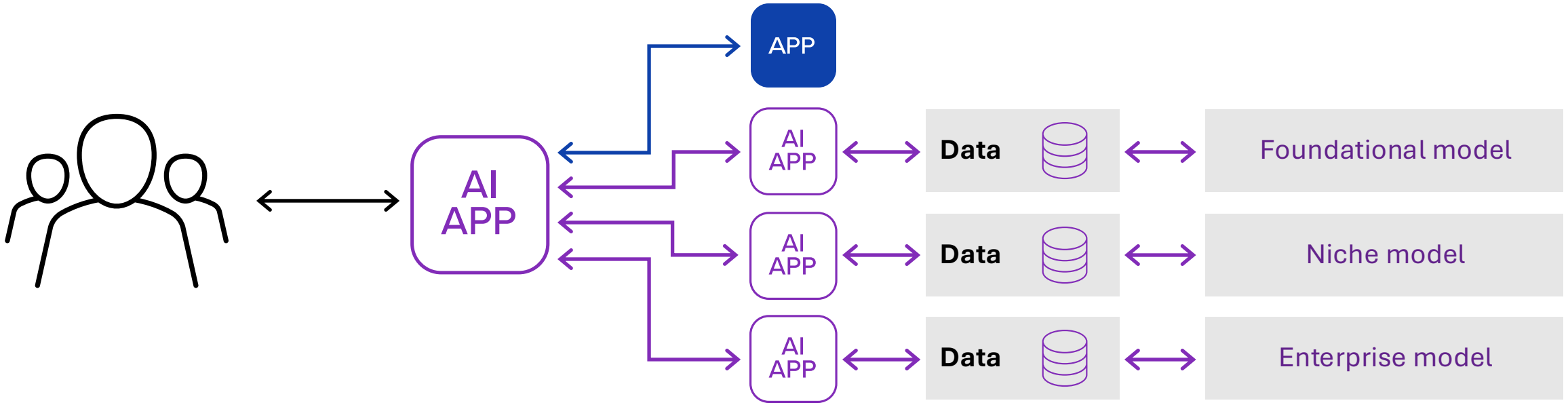


AIアプリケーションは複数のAIモデルを含む、多様なコンポーネントにより成り立つ



AIはハルシネーションや捏造するケースもある

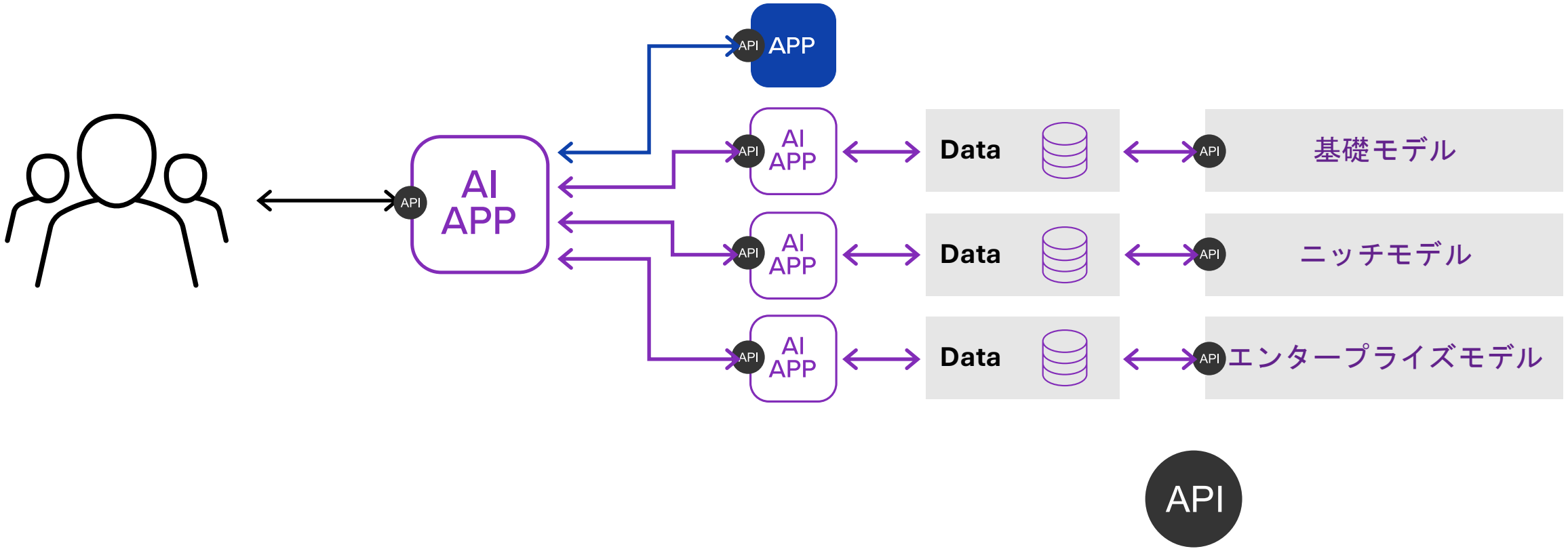
企業は検索拡張生成（Retrieval Augmented Generation=RAG）を用いて自社独自のデータを活用してAIモデルの出力改善を模索している

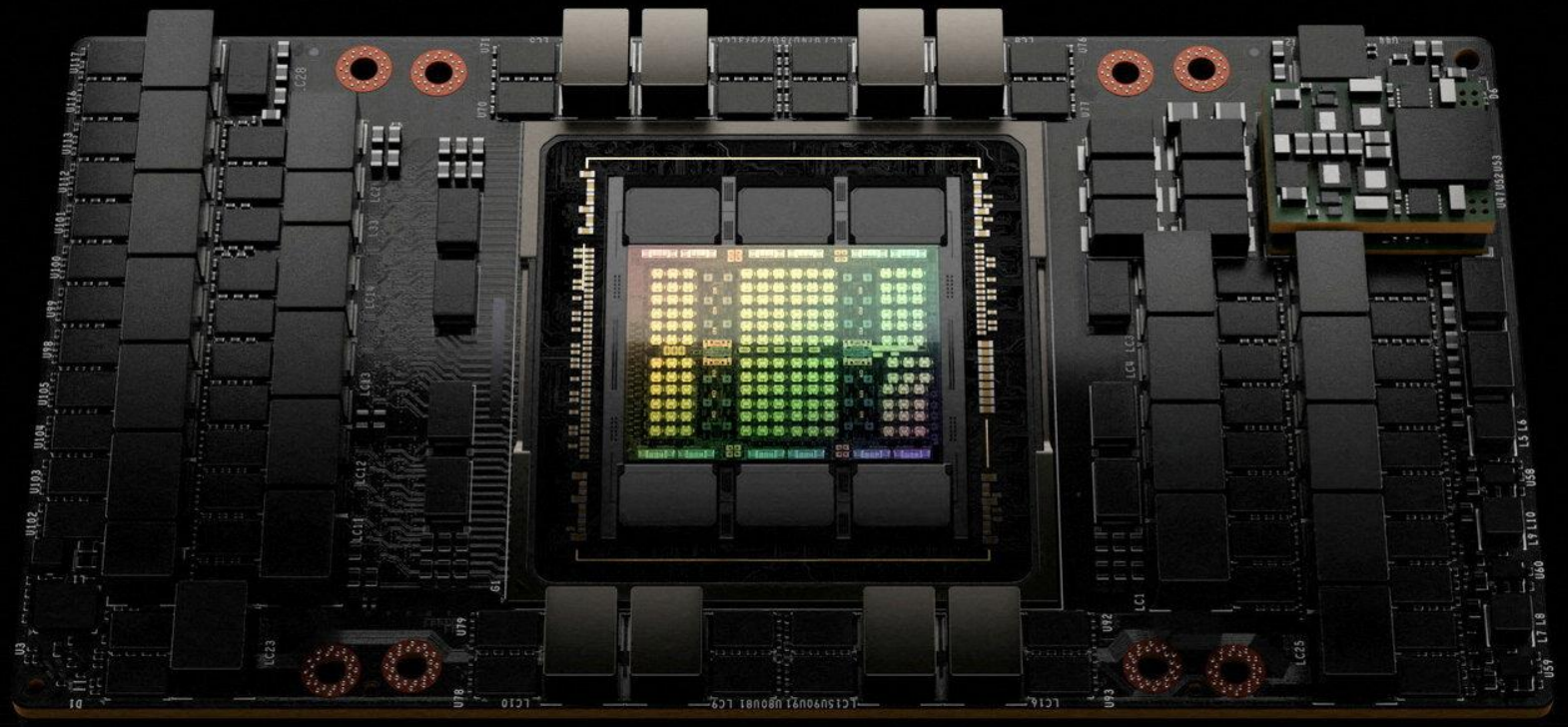


検索拡張生成（RAG）のようなパターンは
検索されたデータを
コンテキストウィンドウに
活用して出力を改善する

一般的なAIモデルはコンテクストとして数百のトークンを保持することができる

AIアプリケーションは従来以上にAPIに依存することになる





Source: NVIDIA H100 Tensor Core GPU, NVIDIA CORPORATION

GPUは並列化を前提に設計されており、Transformerのアーキテクチャに最適である

今時点のモデルを賄うために GPUへの投資は巨額になっている：

168GB

for Llama 70B

972GB

for Llama 405B

Calculating GPU memory for serving LLMs

November 16, 2023 by Sam Stoelinga

How many GPUs do I need to be able to serve Llama 70B? In order to answer that, you need to know how much GPU memory will be required by the Large Language Model.

The formula is simple:

$$M = \frac{(P * 4B)}{(32/Q)} * 1.2$$

Symbol	Description
M	GPU memory expressed in Gigabyte
P	The amount of parameters in the model. E.g. a 7B model has 7 billion parameters.
4B	4 bytes, expressing the bytes used for each parameter
32	There are 32 bits in 4 bytes
Q	The amount of bits that should be used for loading the model. E.g. 16 bits, 8 bits or 4 bits.
1.2	Represents a 20% overhead of loading additional things in GPU memory.

Now let's try out some examples.

GPU memory required for serving Llama 70B

Let's try it out for Llama 70B that we will load in 16 bit. The model has 70 billion parameters.

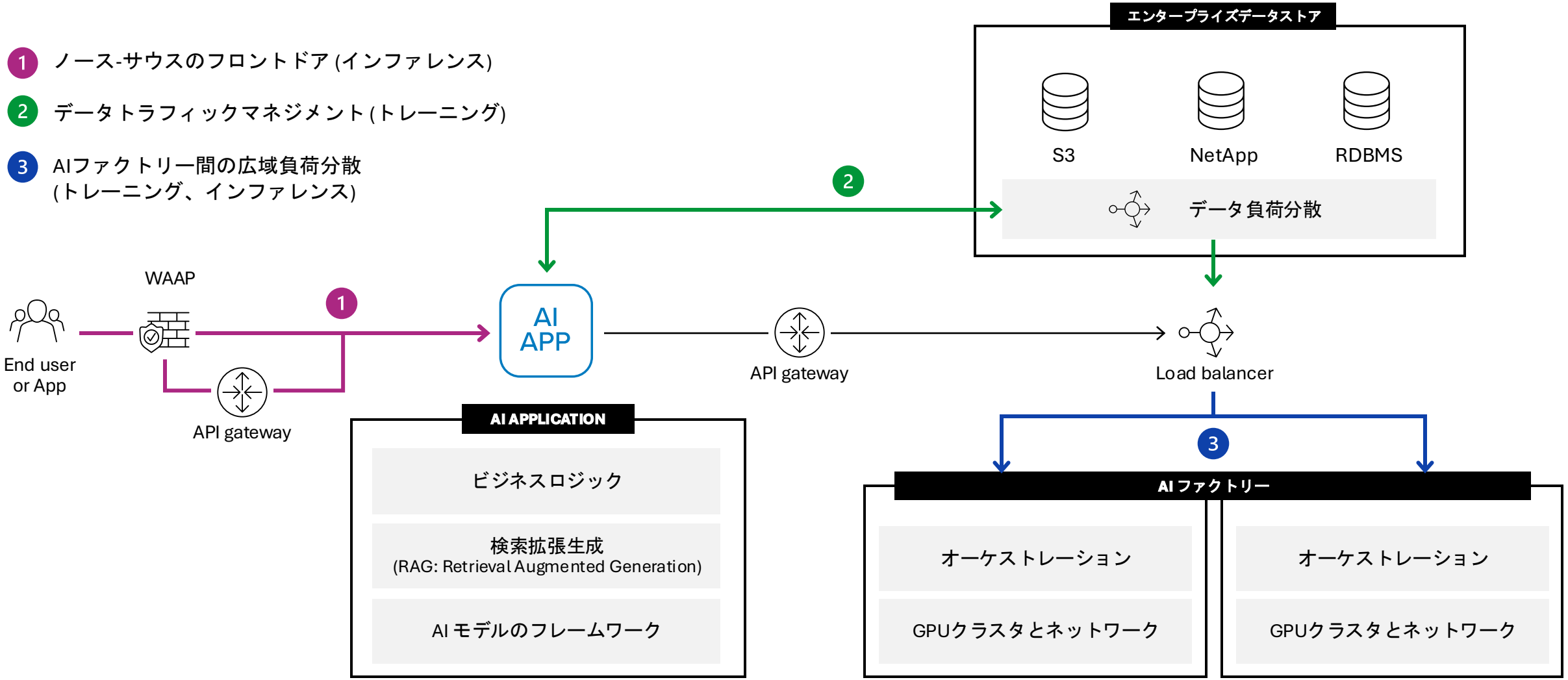
$$\frac{70 * 4\text{bytes}}{32/16} * 1.2 = 168\text{GB}$$

That's quite a lot of memory. A single A100 80GB wouldn't be enough, although 2x A100 80GB should be enough to serve the Llama 2 70B model in 16 bit mode.

AIサーバは最終的には
AIファクトリーとなる

AIファクトリーと関連する情報フローのアーキテクチャ概要

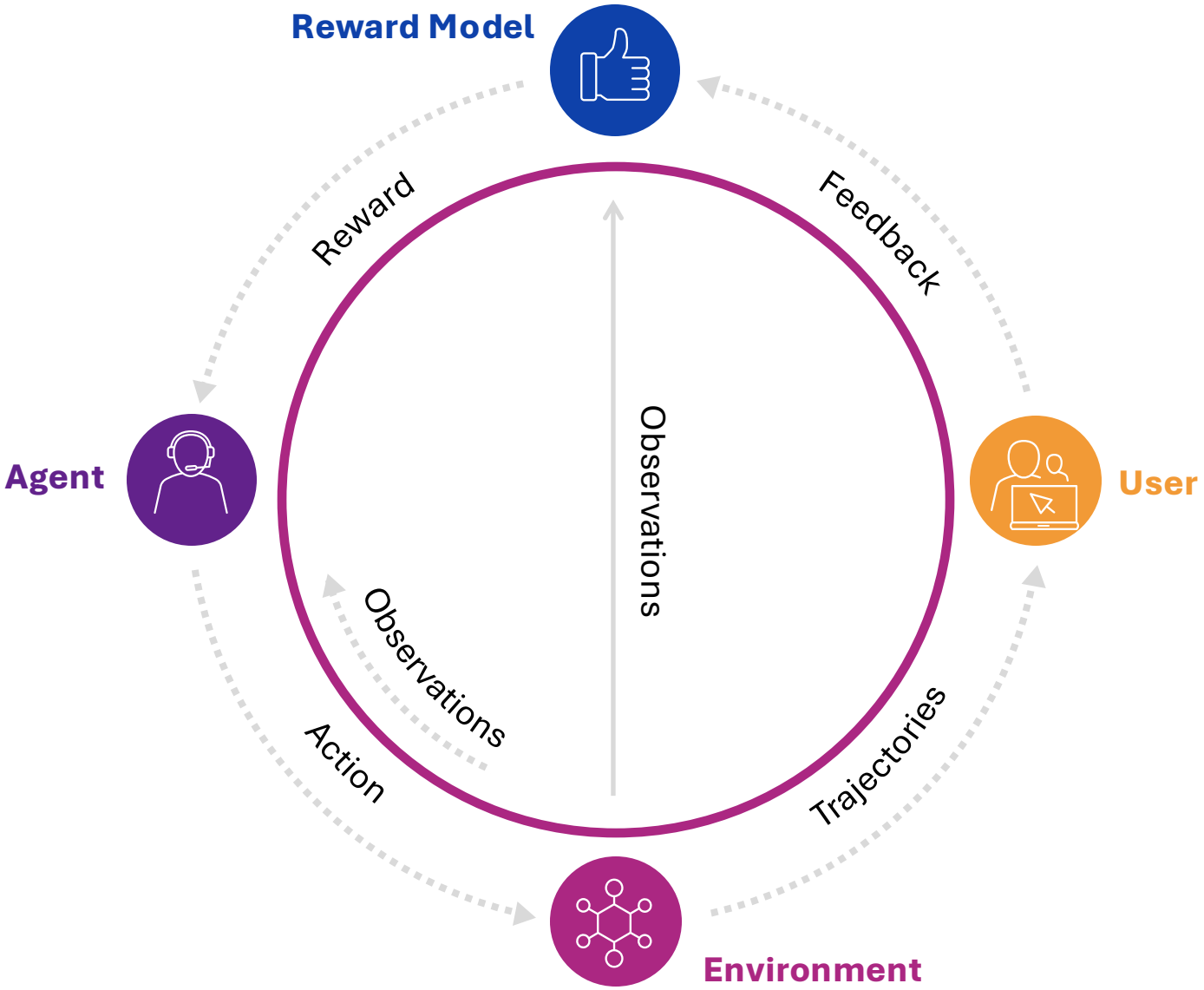
1. ノース-サウスのフロントドア (インファレンス)
2. データトラフィックマネジメント (トレーニング)
3. AIファクトリー間の広域負荷分散 (トレーニング、インファレンス)



責任ある実装

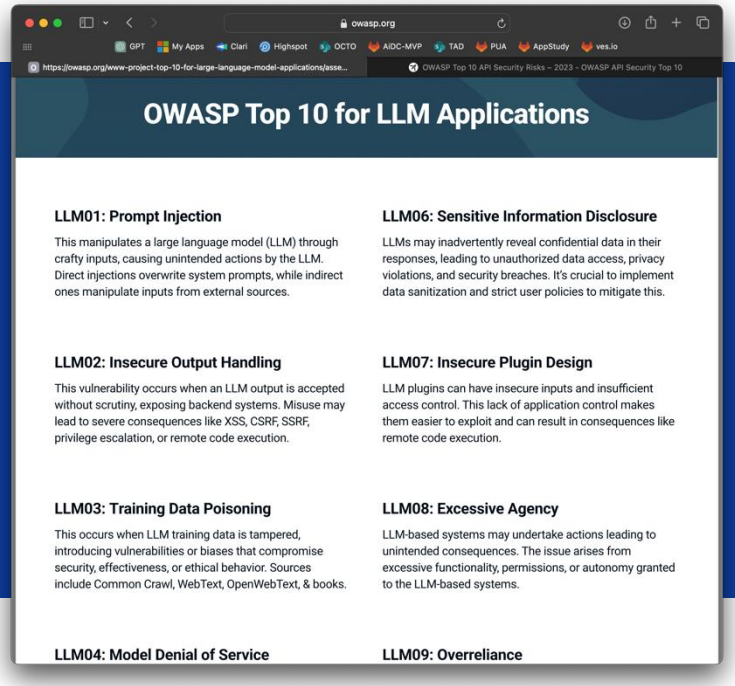
ユーザ企業はAI CoE
（全社的なAIビジョンの実現に特化した組織単位）
に投資注力し、
セキュリティ・安全性・品質改善を進めている

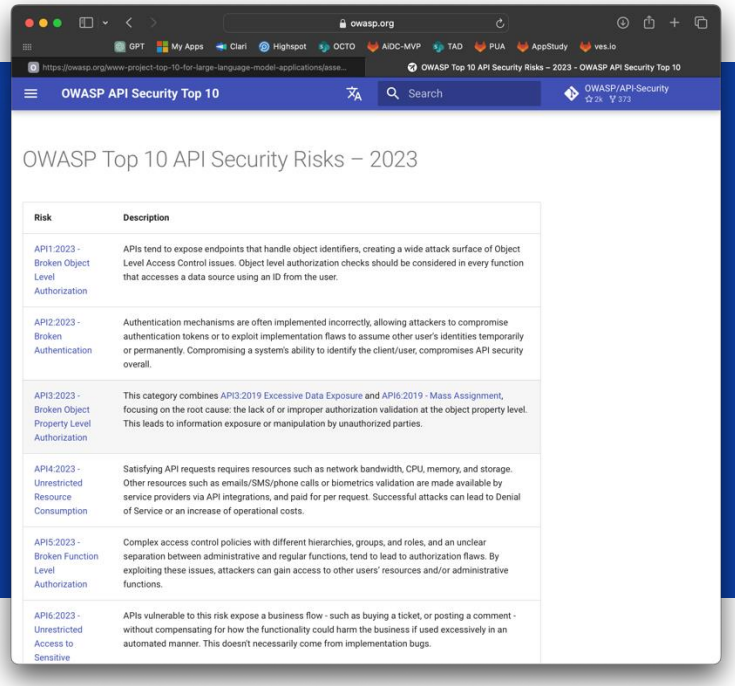
企業はエージェントアーキテクチャにおいて重要となる、人間にとっての価値中心のモデルへと調整を進めている



Source: Image from Google DeepMind

セキュリティはモデルとAPIの保護と合わせて考慮される必要がある

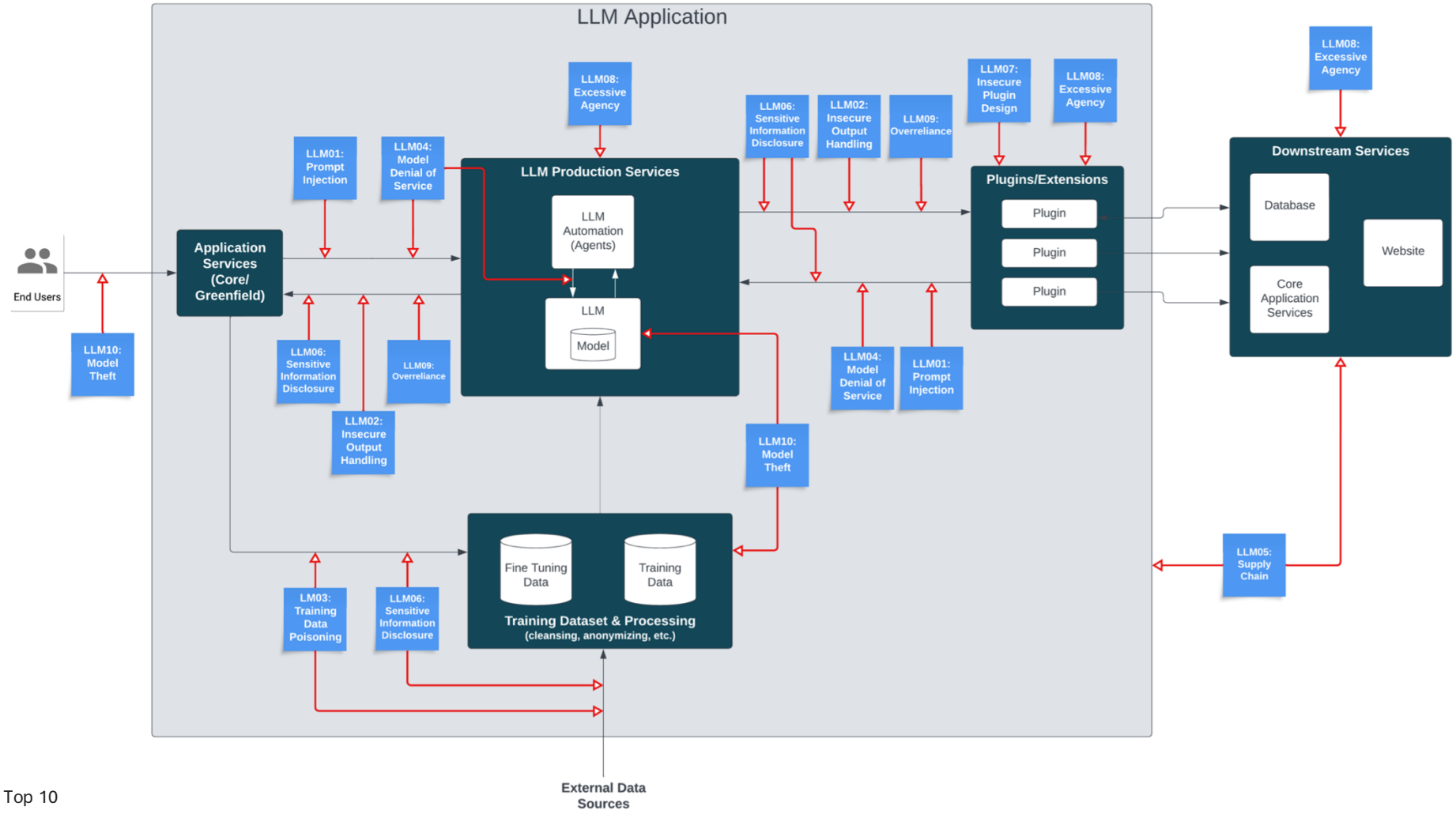




OWASP LLM Top 10
モデル（データ）の保護

OWASP API Top 10
API（アクセスとデータ）の保護

大規模言語モデルアプリケーションにおけるOWASP LLM Top 10 呼び出し概略図



Source: Image from OWASP LLM Top 10

AIの次のステップは？

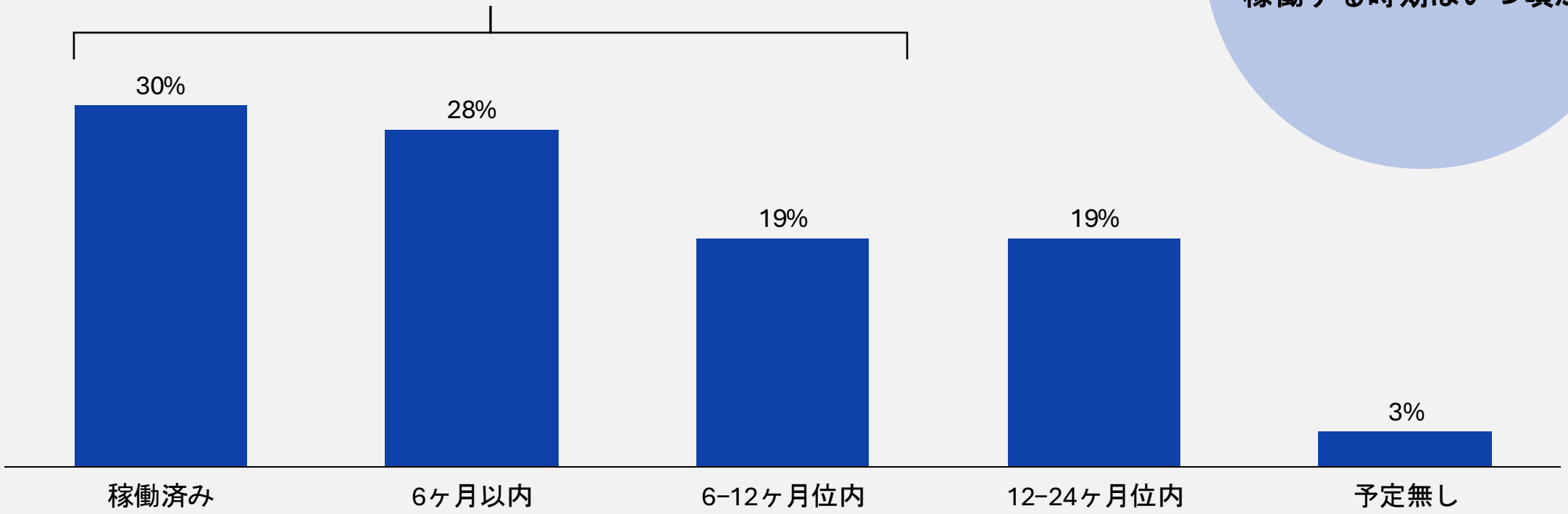
予測

2027年までにモダンなアプリケーションの50%以上はAIの技術を活用するようになる

AIは理論上のものではない。

生成AIは本番環境で稼働している。

AI・大規模言語モデル関連プロジェクトが完了し、稼働する時期はいつ頃か？

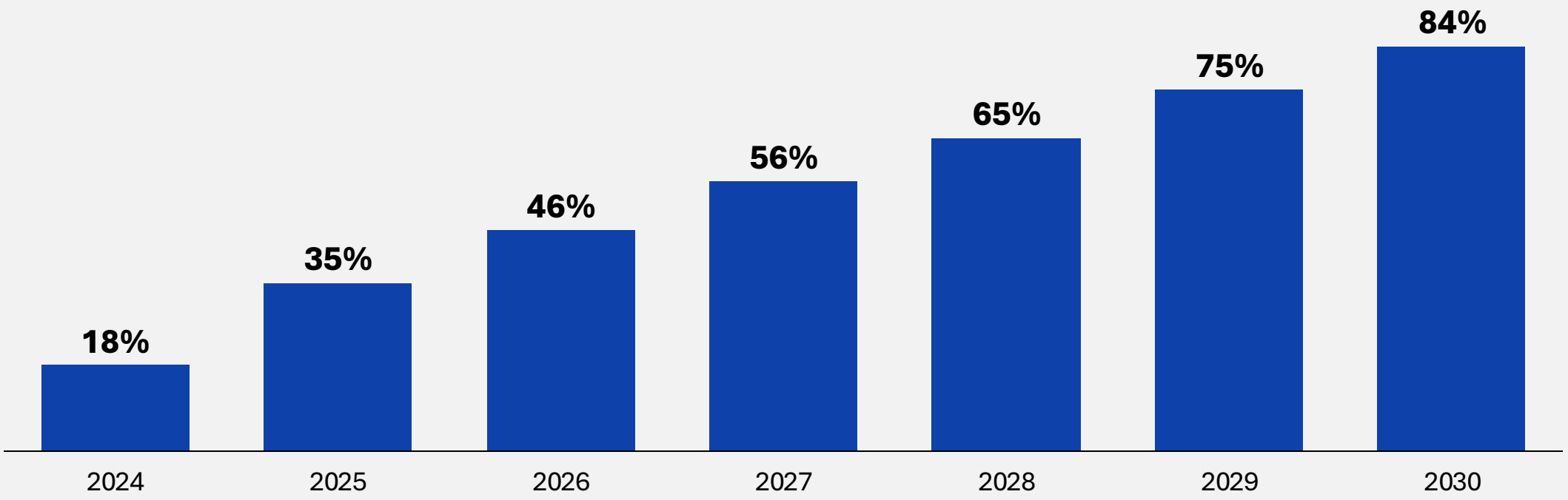


Source: RBC CIO Survey 2024

社外の予測データと組み合わせた弊社社内の試算では
ほぼ全てのアプリケーションがAIアプリかAIのインターフェースを実装する

年ごとのAIアプリの数

全てのアプリケーションの中で占める割合



Source: IDC, Gartner, F5 Corporate Strategy judged

予測

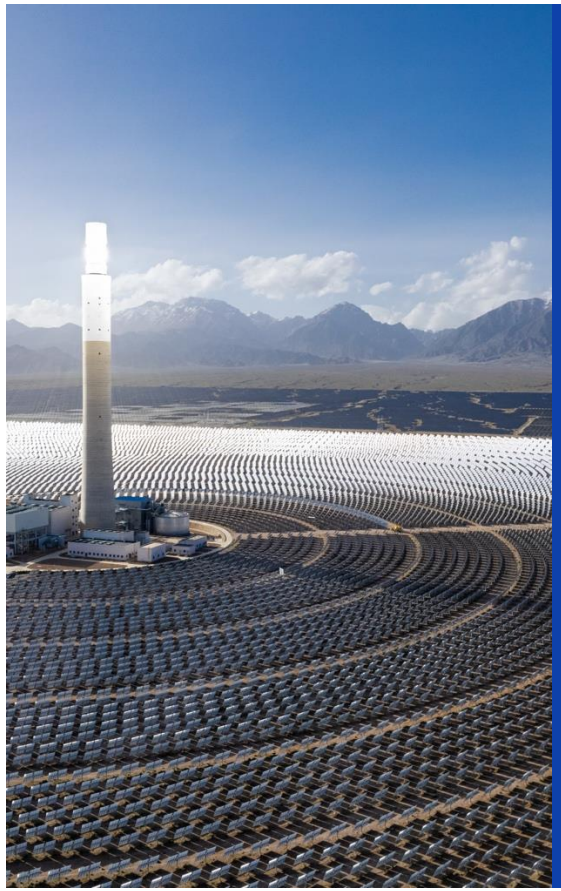
「インテリジェントな」未来を実現するために
3倍以上のエネルギーが必要となる

インテリジェントなアプリケーションに支えられた未来を実現するためには全世界のエネルギー供給が現在比3倍必要

現時点:

7テラワットのエネルギー

ほとんどのエネルギーが石炭、自然ガスその他の水力などの代替エネルギーで賄われている



FUTURE:

23テラワットのエネルギー

ソフトウェアとハードウェアの効率化とともにソーラーや原子力技術などの次世代エネルギーの活用が必須

現実

AIの安全性とセキュリティは最重要である

この先数年のAIは
エネルギー、システム、セキュリティの
スケーリングが重要となる

F5の取り組み

F5はポートフォリオ全体で、生成AI ソリューションを導入

顧客体験をさらにリッチに、
セキュリティ体制を強化し
重要なインサイトを提供する
ために、生成AIソリューション
の追加を進めています。

BIG-IPのためのiRuleコード生成

AIを搭載したWAF

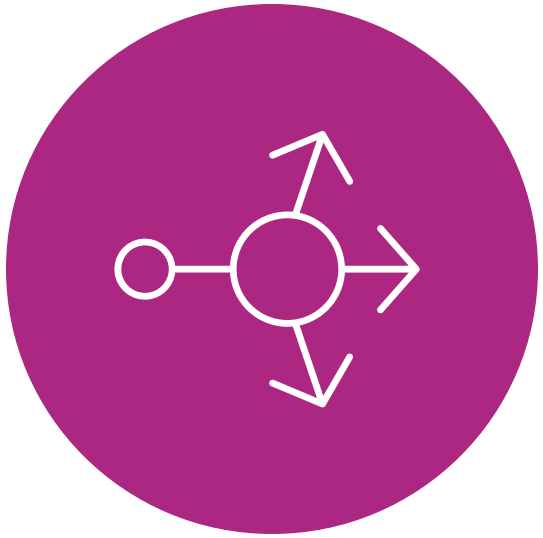
NGINXとDistributed Cloud Services向けのAIアシスタント

NGINXの予測運用

F5 は、あらゆる領域でAI アプリケーションの提供と保護を可能に

AIデータの取り込み

AIモデルのトレーニングおよび
RAGのために、大量かつ多様な
データを効率的に収集・準備



AIファクトリーの
ロードバランシング

高度なトラフィック管理で
AIファクトリーのパフォーマンスと
規模を最適化

安全性の高いAI推論

AIアプリの展開と実行時に
機密データとモデルを保護


App
World
2024