



Enterprise Technical Workshop
次世代 ITインフラエンジニア育成ワークショップ

次世代インフラエンジニアのためのAI基礎

Network × AI × Programmability

Toshiya Kawata (CCIE #51381, AI Evangelist)
Solutions Engineer
Japan Enterprise Systems Engineering 2
2024/12/11



Agenda

- 1 自己紹介・略歴
- 2 AIの歴史と現在地
- 3 LLMの長所と短所
- 4 LLMに翼を授けるRAGとは
- 5 単語や文章のベクトル化とは
- 6 次世代のAIOpsを実現しよう！
- 7 AI × Programmability のデモ

自己紹介・略歴

■ 川田 俊也（かわた としや）

- 専門分野: 知能情報系 (ロボット工学, 計算幾何学)
- 大学講師 (CCNA, C言語, Java, 組み込み)
- 2011年度 Cisco入社
 - ✓8年4ヶ月: Advanced Service/Professional Service
 - ✓2年5ヶ月: Customer Success
 - ✓2年9ヶ月: Solutions Engineer

■ Networking 領域における知識や経験

- Router, Switch, WLAN, DC, Security全般
- 保有資格
 - ✓CCNA/CCAI (2009)
 - ✓CCIE Routing & Switching (2015)
 - ✓Devnet Professional (2020)



■ Programming 領域における知識や経験

- Assembler, C/C++, C#, Java, Python, etc...
 - ✓ザラ場 simulator (Java, 2005-2009)
 - ✓Multicast test tool (Java, 2013)
 - ✓Endpoint diff tool for ISE (Python, 2019)

■ AI 領域における知識や経験

- 行動規範型AI・モデル規範型AI
- 強化学習, パターン分類, 統計, 高度な数学
- 保有資格
 - ✓Certify AI検定
 - ✓JDLA ディープラーニング G検定



AIはすべてを変える

\$15.7T

2030年までに世界経済に寄与する
潜在力

\$300B

2026年までのAIへの世界支出

75%

2026年までに大企業の70%がAIを
活用したプロセスを導入



ヘルスケアと
ライフサイエンス
診断
創薬
個別化医療



金融サービス
不正検知
リスク評価
トレーディング



リテール
パーソナライゼーション
在庫最適化
仮想エージェント



製造業
予知保全
品質管理
需要予測



農業
収穫量の最適化
自動灌漑
害虫の予測と予防



交通機関
ルート最適化
自動運転車
予知保全



エネルギー
配電最適化
障害予測
需要予測



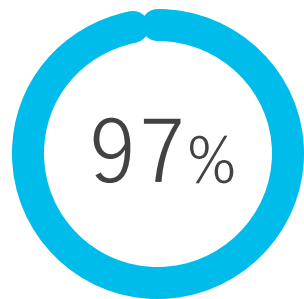
公的機関
スマートシティ
セキュリティ
サービスの改善

Sources: PWC, IDC



© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

AI成熟度調査から見えてくる現状



グローバル企業の97%がAIを活用したテクノロジーの導入を急ぐと回答している一方、AIの導入と活用に向けた準備が十分に整っている企業はわずか14%にとどまっている*1。

41.6%

のグローバル企業が今後18ヶ月*2のインフラストラクチャへの投資で生成AIが最優先課題であると回答。

14%

一方で現状AIを活用する準備が十分に整っている企業は全体の13%にとどまっている。*1

36.2%

のグローバル企業が生成AI活用のためには、IT従事者の雇用/IT技術者の技術力習得のための育成が最優先課題であると回答。*1

*1: [Cisco AI Readiness Index, November 2024](#)

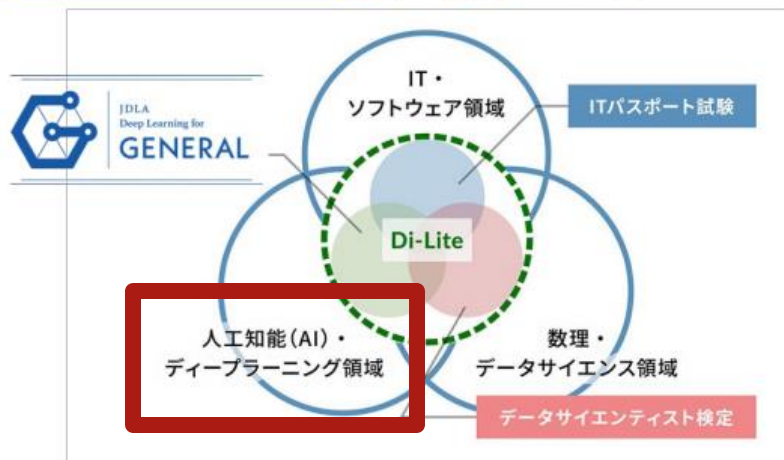
*2: IDC: Future of Enterprise Resiliency & Spending Survey Wave 7, Generative AI Infrastructure Spending Sentiments, November 2024

デジタルリテラシー
とは？



デジタルリテラシーを、全てのビジネスパーソンに

デジタルリテラシー協議会

すべてのビジネスパーソンが学ぶべき
デジタルリテラシー Di Lite を定義

デジタルリテラシー・スキルフレームワーク





・ “DX推進パスポート1・2・3”

「ITパスポート試験」に合格した上で、「DS検定リテラシーレベル」と「G検定」の取得を推奨。

ビジネスパーソン “全員”



ITを活用するすべての社会人・これから社会人となる学生が備えておくべき、ITに関する基礎的な知識が証明できる国家試験です。
ITの知識だけでなく、経営全般の知識、プロジェクトマネジメントの知識など、ITを活用するために必要な知識を習得できます。

※相当するITSSレベル定義 レベル1
情報技術に携わる者に最低限必要な基礎知識を有します。情報技術に携わる者に最低限必要な基礎知識を有します。

DX推進を目指す“全員”

DX推進を行う職場において、上位者の指導の下に、要求された作業を担当する人。
DXを推進するプロフェッショナル人材となるために必要な基本的知識・技能を有する。

JDLA
Deep Learning for
GENERAL

AI・ディープラーニング領域について体系的に学び、適切な活用方針を決定して、事業活用する能力や知識を有しているかを証明する資格試験です。技術的手法から、法律、倫理、事例まで、活用のための知識を幅広く学ぶことができます。

データサイエンティスト検定
リテラシーレベル★

データサイエンティストに必要なデータサイエンス力・データエンジニアリング力・ビジネス力についてそれぞれ見習いレベルの実務能力や知識、また、数理・データサイエンス・AI教育のリテラシーレベルの実力を有していることを証明する資格試験です。

※相当するITSSレベル定義 レベル2
上位者の指導の下に、要求された作業を担当します。プロフェッショナルとなるために必要な基本的知識・技能を有する。

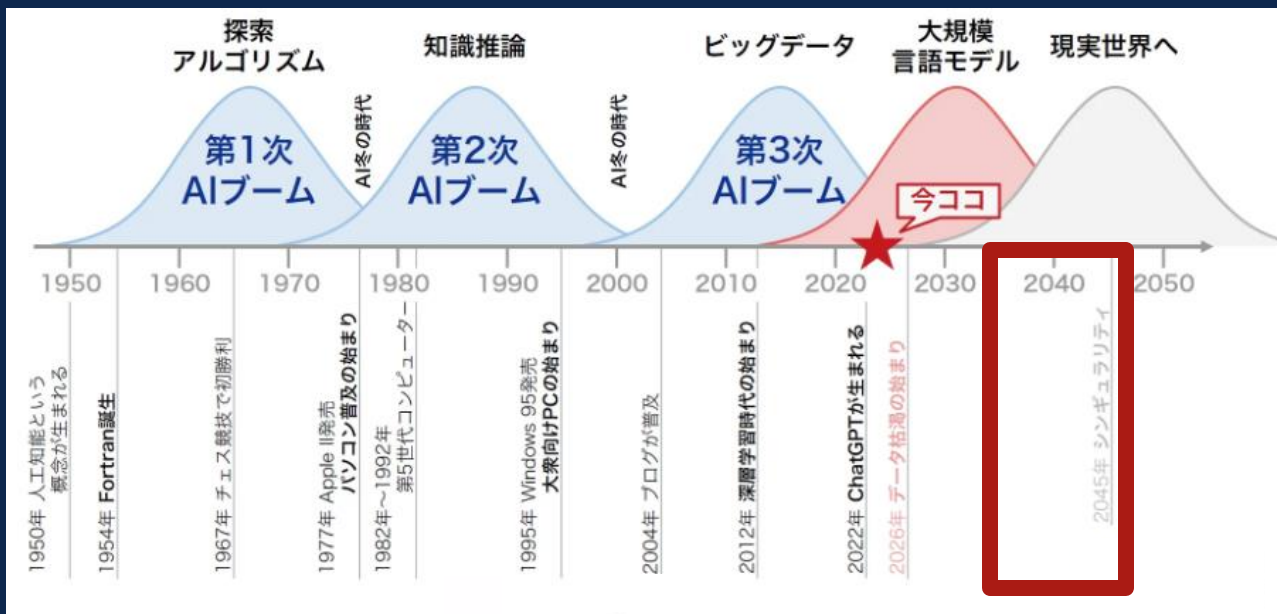
AIさん

手となり足となり

翼になってください!

AIの概要と歴史

- AI: Artificial Intelligence(日本語では「人工知能」)
- 概念は1950年頃から存在しており、人工知能(AI)という用語が造られたのは1956年
- データ量増大、アルゴリズム高度化、コンピューティング性能の発展により、近年AIは急速に発展



AIの概要と歴史

ブーム	時期	主軸のテーマ	概要	主な衰退要因
第1次	1950年代後半 ～ 1960年代	探索と推論	データ構造の工夫によって、オセロやチェスなどのテーブルゲームをプレイできるAIや、簡単な受け答えができる対話型AIの元祖「ELIZA」の登場	- コンピュータの性能が低く、複雑な計算をリアルタイムで行うことが困難 - 複雑な知識をシステム内で表現する方法が未確立 - 探索で解ける問題が世の中に意外と少ない - トイ・プロブレムは解けても現実の複雑な問題は解けない
第2次	1980年代	知識表現	ビジネス（株価予測）や、医療（病理診断）などのエキスパートシステム（ルールベースの知識エンジン）の登場	- 世の中の知識は無限、知識を蓄積・管理すること自体が大変 - エキスパートシステムのルールの追加・修正が非常に困難 - 当時のハードウェアでは、トレーニングに必要な計算能力が不足 - 知識同士の矛盾の可能性
第3次	2010年～	機械学習 特徴表現学習	ビッグデータを用いて知識を定義する要素（特徴量）を自ら習得するディープラーニングが登場し、AlphaGo（2016年）が囲碁のプロ棋士に勝利	- 前処理や多様な教師データの整備が非常に困難 - ディープラーニングモデルの解釈が難しくブラックボックス - 技術の限界や実用化の困難さに対する認識不足
第4次？	現在	超大規模DL 生成AI	2017年に発表のTransformerに端を発し、GPUを利用した並列計算能力の向上もあり歴史的なブレイクスルー	- プライバシーの侵害、バイアスの問題等の倫理的問題？ 「超知性」の誕生（シンギュラリティ）への不安？

Transformerの登場

- ✓ RNNやCNN無しの自己注意機構のみ
- ✓ 比較的少ないコストで並列演算学習が可能
- ✓ 他のモデルに匹敵する性能を発揮

Attention Is All You Need			
Abhimanyu Vaswani* Google Brain avaswani@google.com	Noam Shazeer* Google Brain noam@google.com	Niki Parmar* Google Research nikip@google.com	Jakob Uszkoreit* Google Research uszkoreit@google.com
Llion Jones* Google Research llion@google.com	Aidan N. Gomez*† University of Toronto aidan@cs.toronto.edu	Lukasz Kaiser* Google Brain lukaszkaiser@google.com	
Illa 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 [15, 21, 13].

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Abhimanyu, with Illia, designed and implemented the first Transformer models and has been critically 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 detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and wrote the paper. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualization. Lukasz and Aidan spent countless long days designing various parts of and implementing later versions, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

1st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

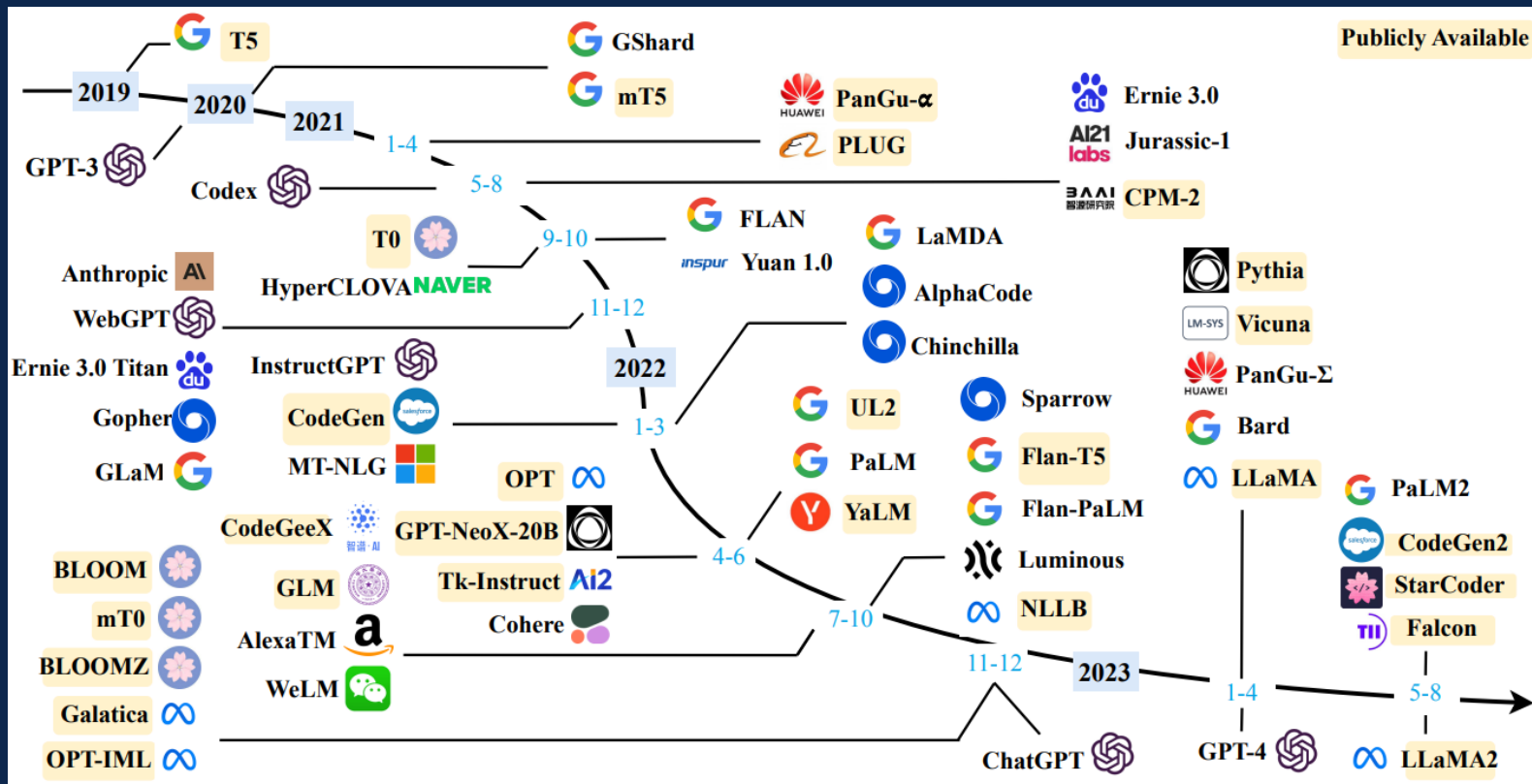
Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	41.29	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$3.3 \cdot 10^{18}$	
Transformer (big)	28.4	41.0	$2.3 \cdot 10^{19}$	

極めて広範囲の基盤技術として活躍



Large Language Models (10B以上) の変遷



AIの進歩と未来

	PRE - 2020	2020	2022	2023	2025?	2030?
TEXT	Spam detection Translation Basic Q&A	Basic copy writing First drafts	Longer form Second drafts	Vertical fine tuning gets good (scientific papers, etc)	Final drafts better than the human average	Final drafts better than professional writers
CODE	1-line auto-complete	Multi-line generation	Longer form Better accuracy	More languages More verticals	Text to product (draft)	Text to product (final), better than full-time developers
IMAGES			Art Logos Photography	Mock-ups (product design, architecture, etc.)	Final drafts (product design, architecture, etc.)	Final drafts better than professional artists, designers, photographers
VIDEO / 3D / GAMING			First attempts at 3D/video models	Basic / first draft videos and 3D files	Second drafts	AI Roblox Video games and movies are personalized dreams
Large model availability:					● Almost there	● Ready for prime time

AI (Large Language Models) の現在

GLUE SuperGLUE

Paper </> Code Tasks Leaderboard FAQ Diagnostics Submit

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI
1	Microsoft Alexander v-team	Turing ULR v6		91.3	83.3	97.5	94.2/92.3	93.5/93.1	76.4/90.9	92.5	92.1	96.7
2	JDEExplore d-team	Vega v1		91.3	83.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7
3	Microsoft Alexander v-team	Turing NLR v5		91.2	82.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9
4	DIRL Team	DeBERTa + CLEVER		91.1	84.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7
5	ERNIE Team - Baidu	ERNIE		91.1	85.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3
6	AliceMind & DIRL	StructBERT + CLEVER		91.0	85.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4
7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	81.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2
8	HFL iFLYTEK	MacALBERT + DKM		90.7	84.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8
9	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	83.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5
10	T5 Team - Google	T5		90.3	81.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9
23	GLUE Human Baselines	GLUE Human Baselines		87.1	86.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2

だが、しかし . . .

Large Language Models の短所

1. ハルシネーションの懸念がある

- 結局人間がダブルチェック
- 質問内容に関連する学習データが乏しい

2. 学習データが公開情報のみ

- 非公開情報の知識は無い
- 学習には大量のデータセットが必要

3. カットオフ日が存在する

- カットオフ日以降の最新知識は無い
- 新しい知識を必要としたければ新バージョンのリリースを待つしかない

モデル	ナレッジカットオフ日
GPT-3.5 (ChatGPT)	2021年9月
GPT-4	2023年4月
GPT-4 Turbo	2023年12月
GPT-4o	2023年10月
Claude 3	2023年8月
Claude 3.5 (Sonnet)	2024年4月
Llama 3.1	2023年12月
Llama 3.2	2023年12月

Large Language Models の特徴的な挙動



知らないことは
答えられない

承知之助って知ってる？

「承知之助（しょうちのすけ）」は、江戸時代のしゃれ言葉の一種で、「承知している」「了解した」といった意味を、お洒落に言い換えた表現です。「承知」と「之助」を組み合わせ、軽い調子で「わかりました」「お任せください」といった意味を伝える言い回しです。



プロンプトに追加知識を挿入すれば
答えられる

承知之助は、シスコSEの川田が飼っているウサギです。
彼は1歳のホーランドロップです。

承知之助は、シスコのSE、川田が飼っている1歳のホーランドロップのウサギです。すごく可愛い名前ですね。川田さんがこの名前をつけた理由は、江戸時代のユーモラスな言い回しである「承知之助」に由来するのかもしれませんが。

Large Language Models の特徴的な挙動

承知之助って知ってる？



「承知之助（しょうちのすけ）」は、江戸時代のしゃれ言葉の一種で、「承知している」「了解した」といった意味を人名風に言い換えた表現です。「承知」と「之助」を組み合わせ、軽い調子で「わかりました」「お任せください」といった意味を伝える言い回しです。

ココを外部文書から探して
自動で挿し込みたい

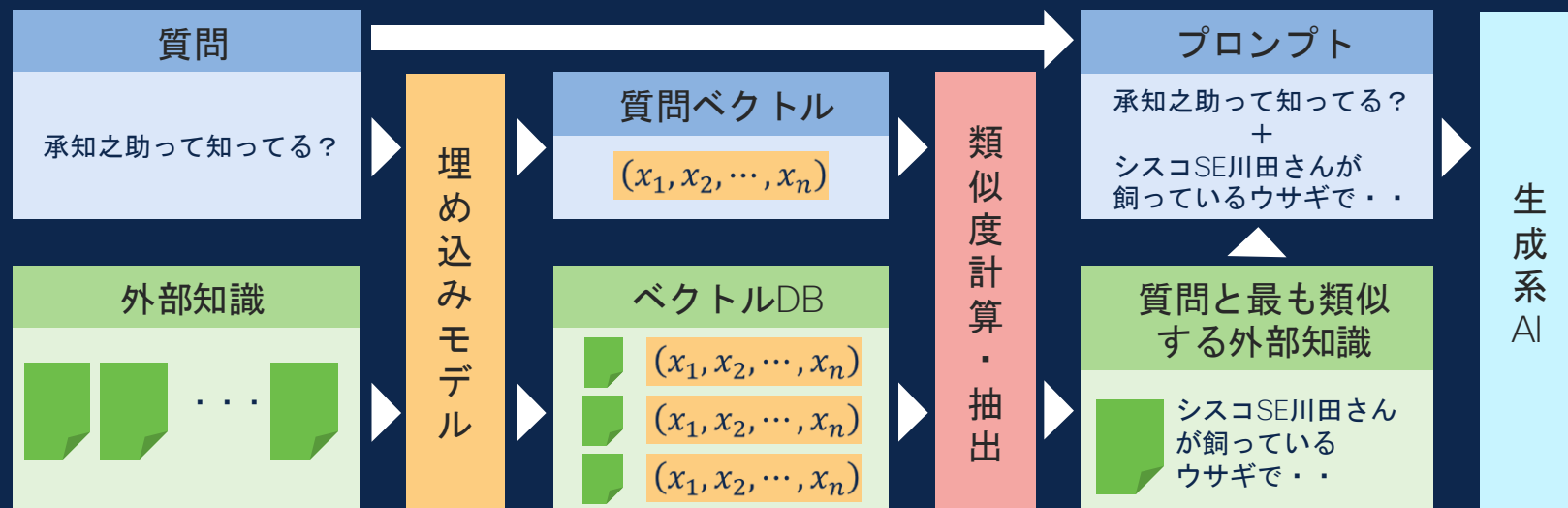
承知之助は、シスコSEの川田が飼っているウサギです。
彼は1歳のホーランドロップです。
承知之助って知ってる？



承知之助は、シスコのSE、川田さんが飼っている1歳のホーランドロップのウサギです。すごく可愛い名前ですね。川田さんがこの名前をつけた理由には、江戸時代のユーモラスな言い回しである「承知之助」に由来するのかもしれませんがね。

RAG: Retrieval-Augmented Generation

外部資源から知識を抽出し、LLMの知識を補完した上で生成する技術

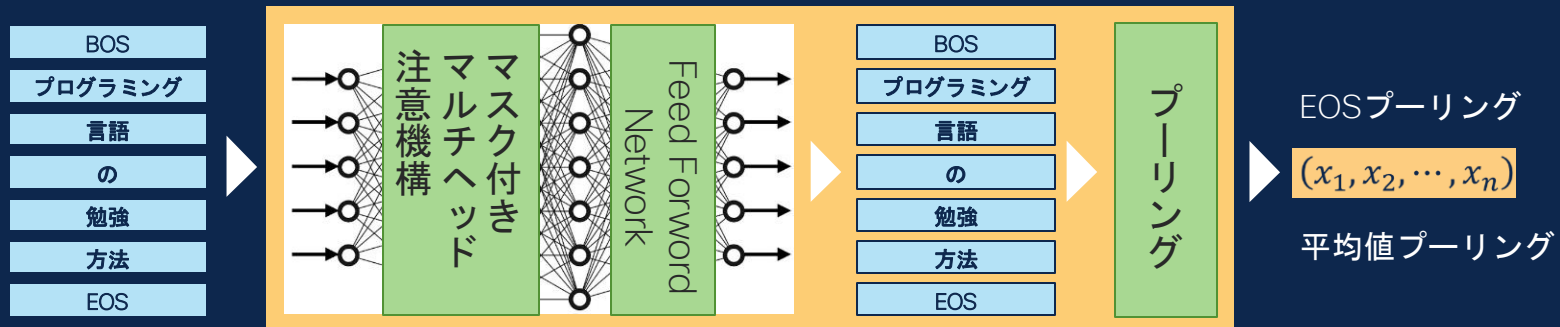
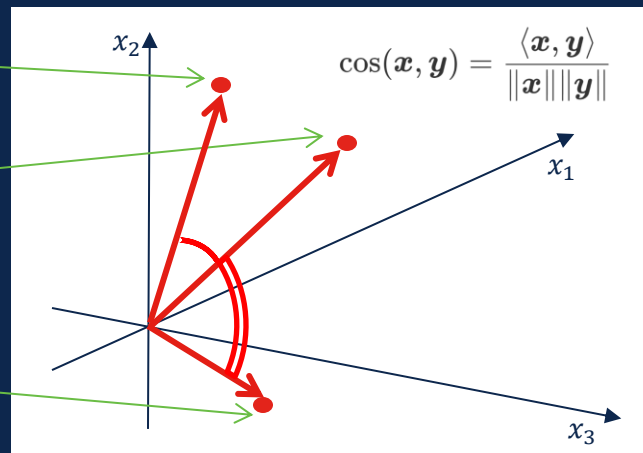


質問と類似する外部知識を探し出すために
テキスト埋め込みベクトルを用いて類似度計算を行う

テキストの埋め込み

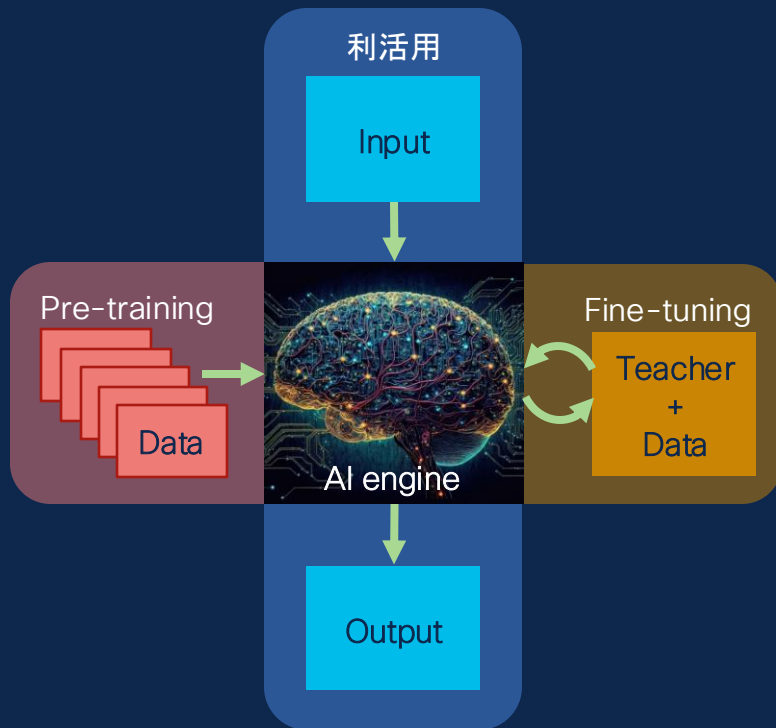
- ・プログラミング言語の勉強方法
- ・効率的なプログラミングの習得法
- ・プログラミングの言語理論

埋め込みモデル

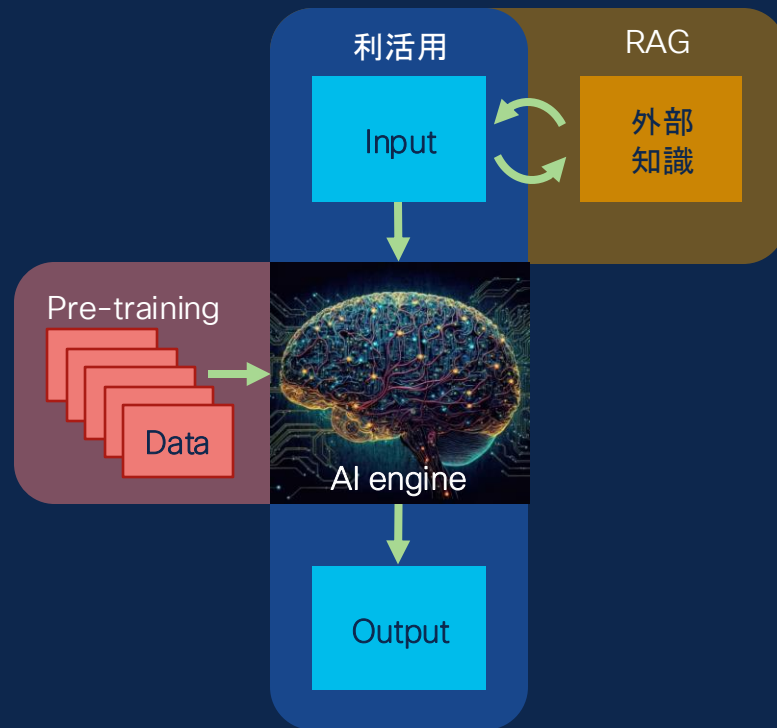


Fine-tuning と RAG の違い

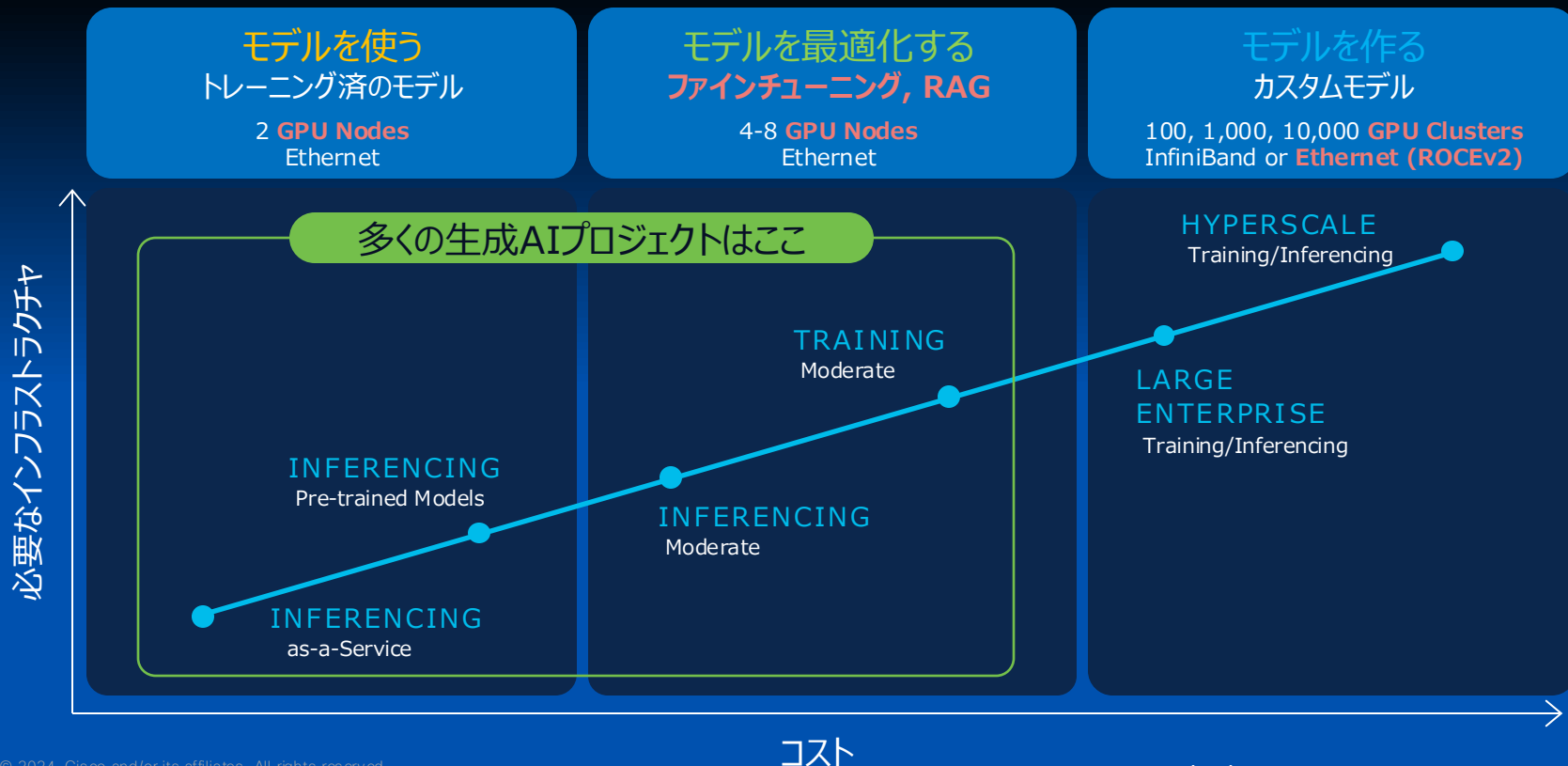
Fine-tuningの概念



RAGの概念



生成AIの全体像



GPU のみの超大規模 AI インフラストラクチャ

H100 GPU X 8 をマウント可能



NVIDIA DGX H100



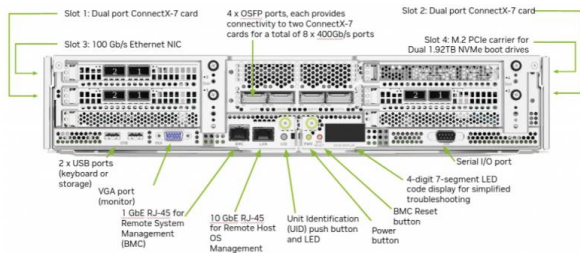
Supermicro H100 サーバー



PowerEdge XE9680



Apollo 6500 G10 Plus

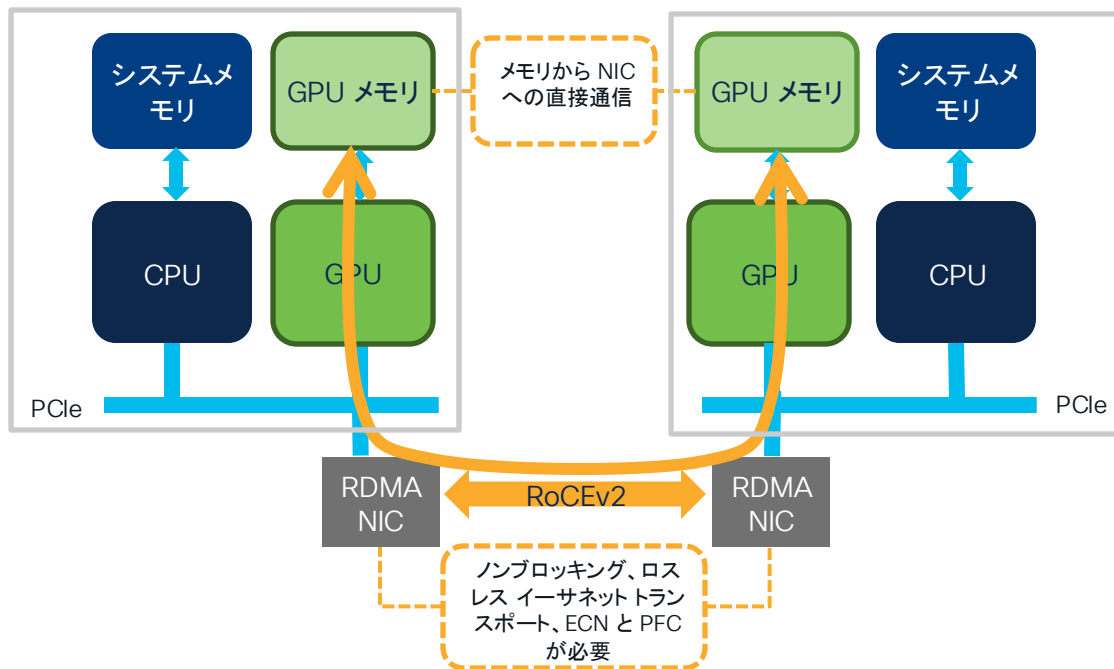


GPU Cluster: 800Gbps 4 X OSPF ポート (IB または ETH)
Storage: 200Gbps 4 X QSPF ポート (IB または ETH)
Option: 100 Gbps NIC

NVIDIA は、マルチ GPU、マルチノード コンピューティング用の N VLINK、NVSWITCH など、さまざまなネットワーク アーキテクチャを提供しています。(32 DGX H100 ノード最大 256 GPU のクラスタリングで最大 57.8 TB/秒のデータをコンピューティング)

GPU クラスタ通信: RDMA

- RDMA により、AI/ML ノードは RAM 内のバイトに直接アクセスすることで、ネットワークを介してデータを交換できます。
- CPU とカーネルをバイパスできるため、遅延が非常に少ない
- RDMA データは InfiniBand ファブリックを介してネイティブに交換される
- その後、RoCE (RDMA over Converged Ethernet) プロトコルにより、イーサネットファブリックを介した交換が可能になりました。



AI にとってネットワークが重要な理由

転送されるデータが
モデルサイズの約6倍に
比例して増加

「モデルサイズが 1 MB 増加すると、
ステップごとに送信されるデータが
5.7 MB 増加」

パケット損失により
トレーニングが8倍遅くなる
可能性がある

「0.01% の損失で、時間はすでに 2
倍になります。2% の損失は、0% の
損失の場合の 8 倍のトレーニング時
間につながる」

Cisco Nexusシリーズ - AI プラットフォーム

Industry leading
25.6Tb switch on a chip



ハイパフォーマンス・低遅延

- 400G、ネイティブ200G対応、800Gを見据えたポートフォリオ
- インテリジェントバッファリング

ロスレスファブリックの大規模展開

- RoCEv2、PFC/ECN対応
- NDFCによる大規模環境でのファブリック自動展開

輻輳状態の可視化

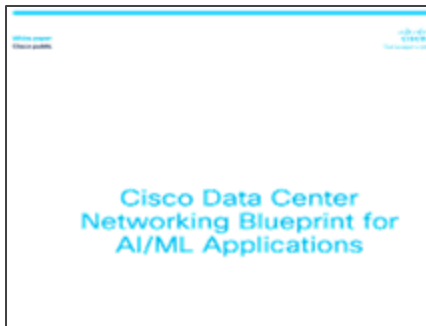
- ハードウェアに実装したテレメトリ機能による可視化
- プラットフォームの輻輳状態の把握と改善



Cisco Data Center Networking Blueprint for AI/ML



ホワイトペーパー
AI/ML向けネットワークの設計
留意点



検証済みデザイン
テスト済みコンフィグ



設定の自動化
可視化



サステナビリティ

AI/MLワークロード要件を満たすソリューション | 検証済み構成 | 自動化、可視化、サステナビリティ



© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

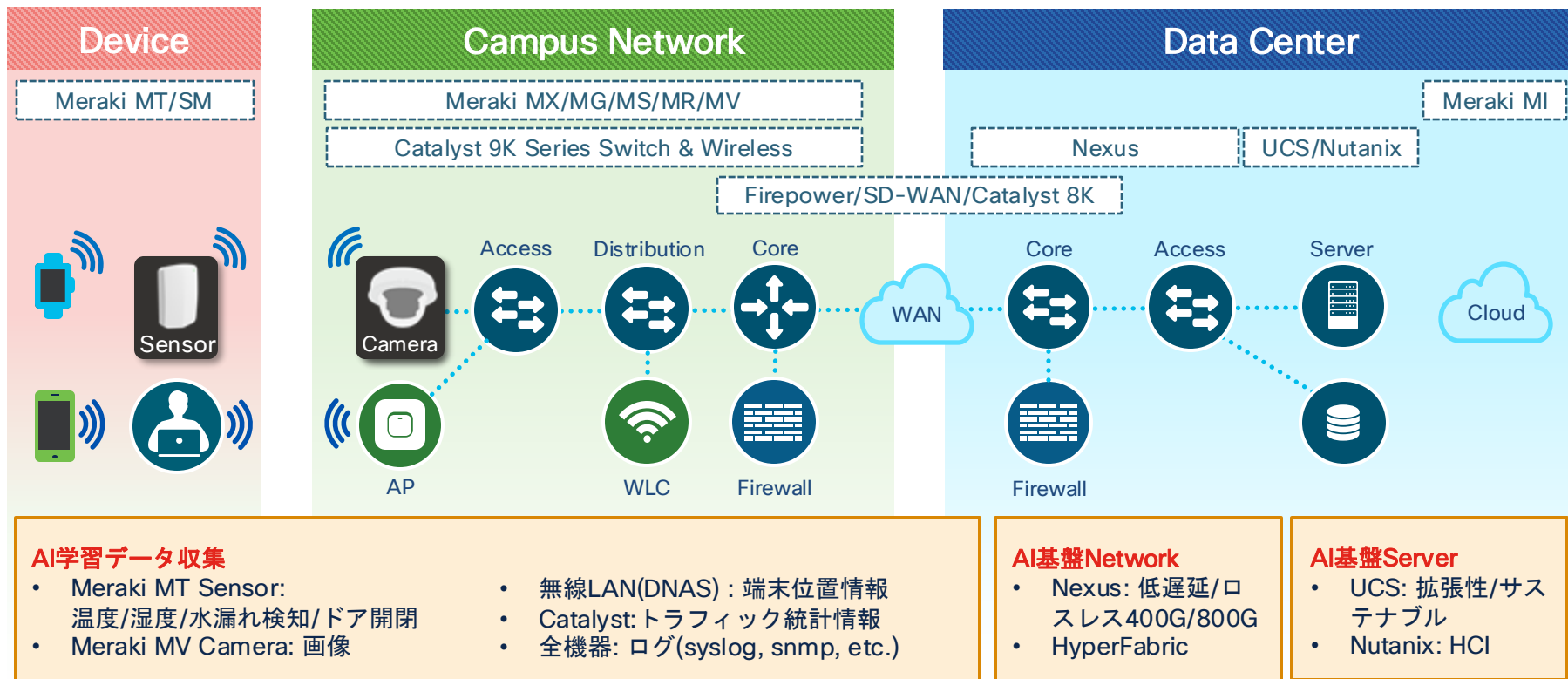
Whitepaper: <https://www.cisco.com/c/en/us/td/docs/dcn/whitepapers/cisco-data-center-networking-blueprint-for-ai-ml-applications.pdf>

Validated Design: <https://www.cisco.com/c/en/us/td/docs/dcn/whitepapers/cvt-for-data-center-networking-blueprint-for-ai.pdf>

Press Release: <https://newsroom.cisco.com/c/en/newsroom/en/aly/2023/m06/cisco-showcases-vision-to-simplify-networking-and-securely-connect-the-world.html>

Cisco Public

AI活用にご利用可能な製品例



プログラミング、、
何かから始めよう、、

プログラミング言語の進化と種類

■ 1950年代以前

- 19世紀にイギリスの数学者チャールズ・バベジが考案される機械式の解析機関を考案
- 1942年にPlankalkül (計画計算書) が整い始める

■ 1960年代前後

- FORTRAN、ALGOL、COBOL

■ 1970年代前後

- Pascal、B言語、C言語
- BASIC (1978年シャープのBASIC-8000)

■ 1980年代

- C++、Objective-C、Pascal

■ 1990年代

- Python、PHP、Ruby、Perl

■ 2000年代

- C#、Scala、D言語、Clojure



される機械式の解析機関を考案
が整い始める

通「FM-8」に標準搭載)

習うより慣れろ

とりあえず Hello World ?

安心してください
まもなく2025年ですよ!

AIの進歩と未来

	PRE - 2020	2020	2022	2023	2025?	2030?
TEXT	Spam detection Translation Basic Q&A	Basic copy writing First drafts	Longer form Second drafts	Vertical fine tuning gets good (scientific papers, etc)	Final drafts better than the human average	Final drafts better than professional writers
CODE	1-line auto-complete	Multi-line generation	Longer form Better accuracy	More languages More verticals	Text to product (draft)	Text to product (final), better than full-time developers
IMAGES			Art Logos Photography	Mock-ups (product design, architecture, etc.)	Final drafts (product design, architecture, etc.)	Final drafts better than professional artists, designers, photographers
VIDEO / 3D / GAMING			First attempts at 3D/video models	Basic / first draft videos and 3D files	Second drafts	AI Roblox Video games and movies are personalized dreams
Large model availability:					● Almost there	● Ready for prime time

言語は何が良い？

Pythonを使うべき理由①

■ Pythonは人気で学習しやすい

- 言語が簡単で書きやすく読みやすい
- ライブラリが豊富※、汎用性が高い
- 年収が高い、グローバルな言語、学習情報量が豊富

※ Requests: Python の HTTP ライブラリ

※ pyATS:

- ❖ シスコが開発しているネットワーク自動化ツール
(もともとはシスコ社内のテスト自動化フレームワーク)
- ❖ Pythonライブラリとして一般公開されており無料で使用可能
- ❖ テストの実行・結果レポートが自動化できる



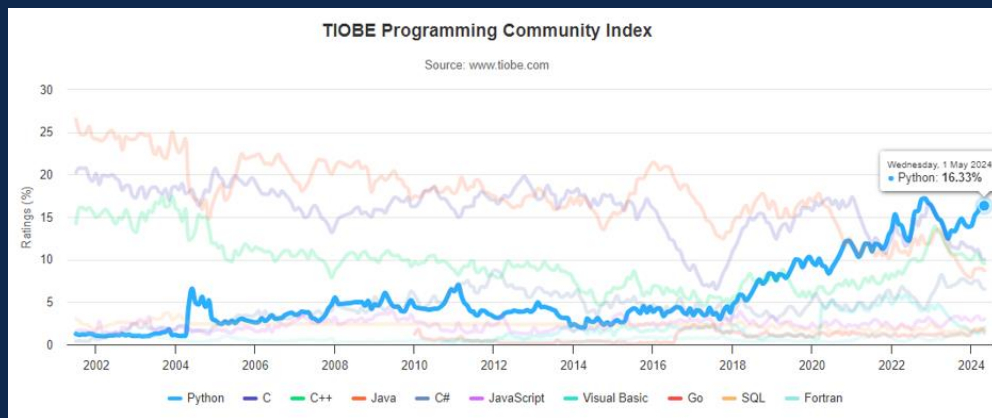
Pythonを使うべき理由②

■ 2024年現在のTIOBEインデックスで第1位

➤ プログラミング言語の数 約250種中1位

➤ <https://www.tiobe.com/tiobe-index>

- ❖ 独立したWikipediaページが存在するか ≒ 十分な知名度があるか
- ❖ 一般的なコンピュータで処理可能かどうか
- ❖ Google検索で5,000以上の参照先があるか



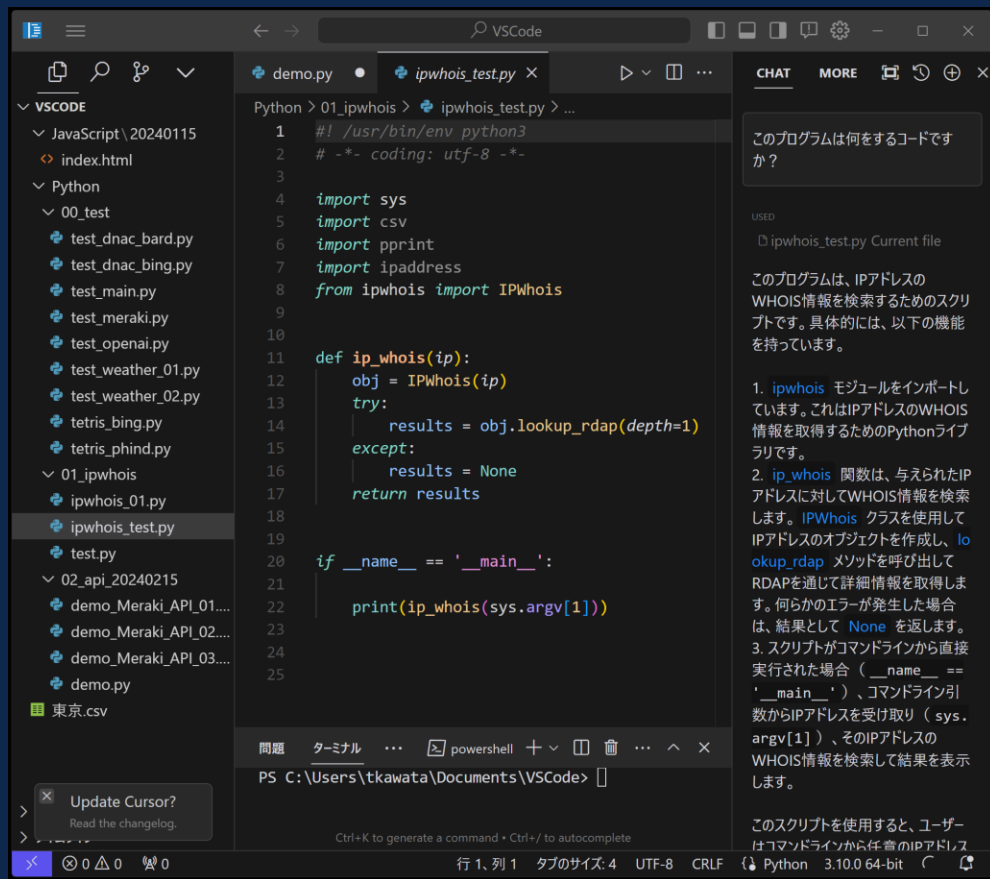
May 2024	May 2023	Change	Programming Language	Ratings	Change
1	1		 Python	16.33%	+2.88%
2	2		 C	9.98%	-3.37%
3	4	▲	 C++	9.53%	-2.43%
4	3	▼	 Java	8.69%	-3.53%
5	5		 C#	6.49%	-0.94%
6	7	▲	 JavaScript	3.01%	+0.57%
7	6	▼	 Visual Basic	2.01%	-1.83%
8	12	▲▲	 Go	1.60%	+0.61%
9	9		 SQL	1.44%	-0.03%
10	19	▲▲	 Fortran	1.24%	+0.46%

AIを活用すれば
誰でも本格的なコードが
すぐに書けます!



The AI-first Code Editor: Cursor

- プログラミング関連の業務全般を効率化できるエディタ
 - 「VSCode (Visual Studio Code)」を基に、新たな機能や改善を加えて開発
 - 「ChatGPT」を搭載し、プログラミング関連の業務を自然言語で効率化できる
 - Open AIのAPI連携も可能
 - コードの自動生成だけでなく、自動デバッグやチャットへの質問など、豊富な機能を搭載
 - AIに参考文献の提供が可能
 - 対応するプログラミング言語は拡張機能次第



LLM+RAGで次世代のAIOpsを実現しよう！

外部知識

- ・ DC基本設計書
- ・ 運用マニュアル
- ・ インシデント管理表
- ・ 問題管理表

:

Primary DC のコアスイッチ1号機でパケットドロップが発生しています。考えられる原因と対処方法を教えて。

コアスイッチはCat9500スイッチ

正常系ではすべてのトラフィックは1号機を通過

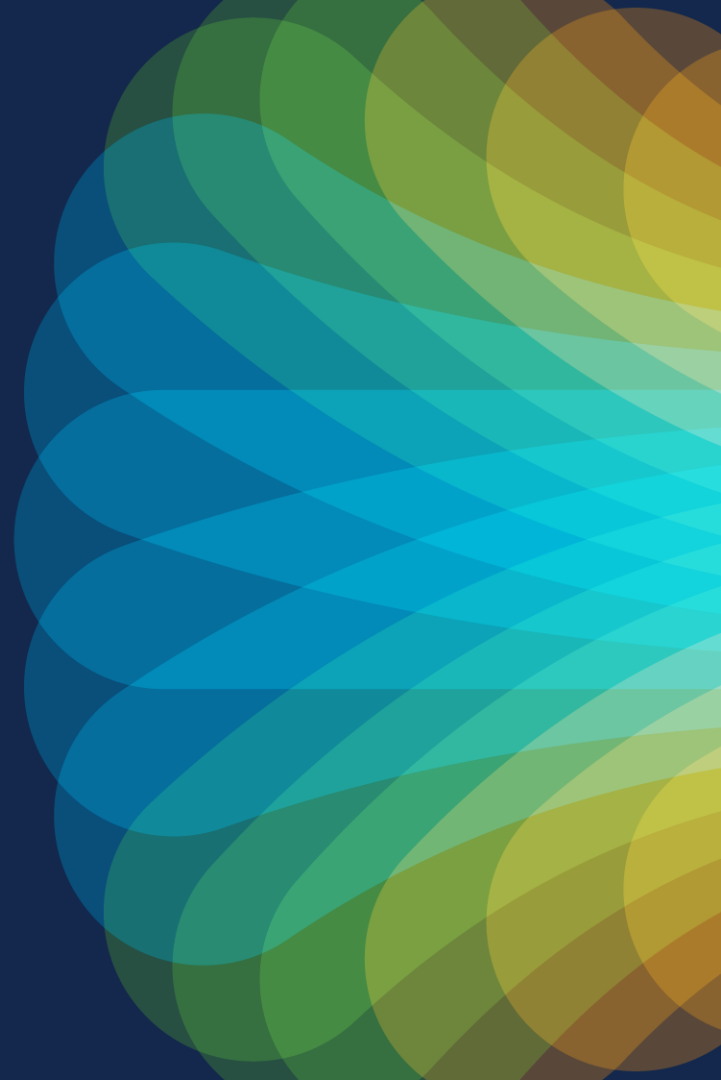
2023年10月にマイクロバーストによってTCPパケットのドロップが発生し、再送による通信遅延が発生



2023年10月に Primary DC のコアスイッチ1号機でパケットドロップが発生していました。その原因はマイクロバーストで、トラフィックを2号機と負荷分散することによって改善しました。また、一般的にマイクロバーストが原因でパケットドロップが発生している場合は、ポートバッファを増強することによって改善することも期待できます。

Cisco Catalyst 9500 スイッチでポートバッファを増やすコマンドは次の通りです。
`#qos queue-softmax-multiplier 1200`

とにかくデモ!



Agenda

- 1 自己紹介・略歴
- 2 AIの歴史と現在地
- 3 LLMの長所と短所
- 4 LLMに翼を授けるRAGとは
- 5 単語や文章のベクトル化とは
- 6 次世代のAIOpsを実現しよう！
- 7 AI × Programmability のデモ

Key points

- 1 AI とか Programming とか 数学とか
先端技術を楽しんで学んで伝える Cisco SE
- 2 約70年に渡って発展と冬の時代を繰り返し、
Transformerに端を発してブレークスルー
- 3 LLMの性能は高いが
学習していない知識は答えられない
- 4 RAGで外部知識を与えて
カスタムLLMの作成が可能
- 5 文章という類似度を比較しづらい対象を
ベクトル化することで簡単に比較
- 6 RAGを活用したカスタムLLMを作って
時間を取られる日常業務を効率化
- 7 Cursor を使って自然言語で
サクサクコーディング&デバッグ

視座高<

視野広<



次世代インフラエンジニアのためのAI基礎

Network × AI × Programmability

Thank you

$$\text{Network Engineer} \stackrel{\text{def}}{=} (\text{Networking} + \text{Programming})^{AI}$$

$$= \sum_{k=0}^{AI} {}_{AI}C_k \text{Networking}^k \text{Programming}^{AI-k}$$