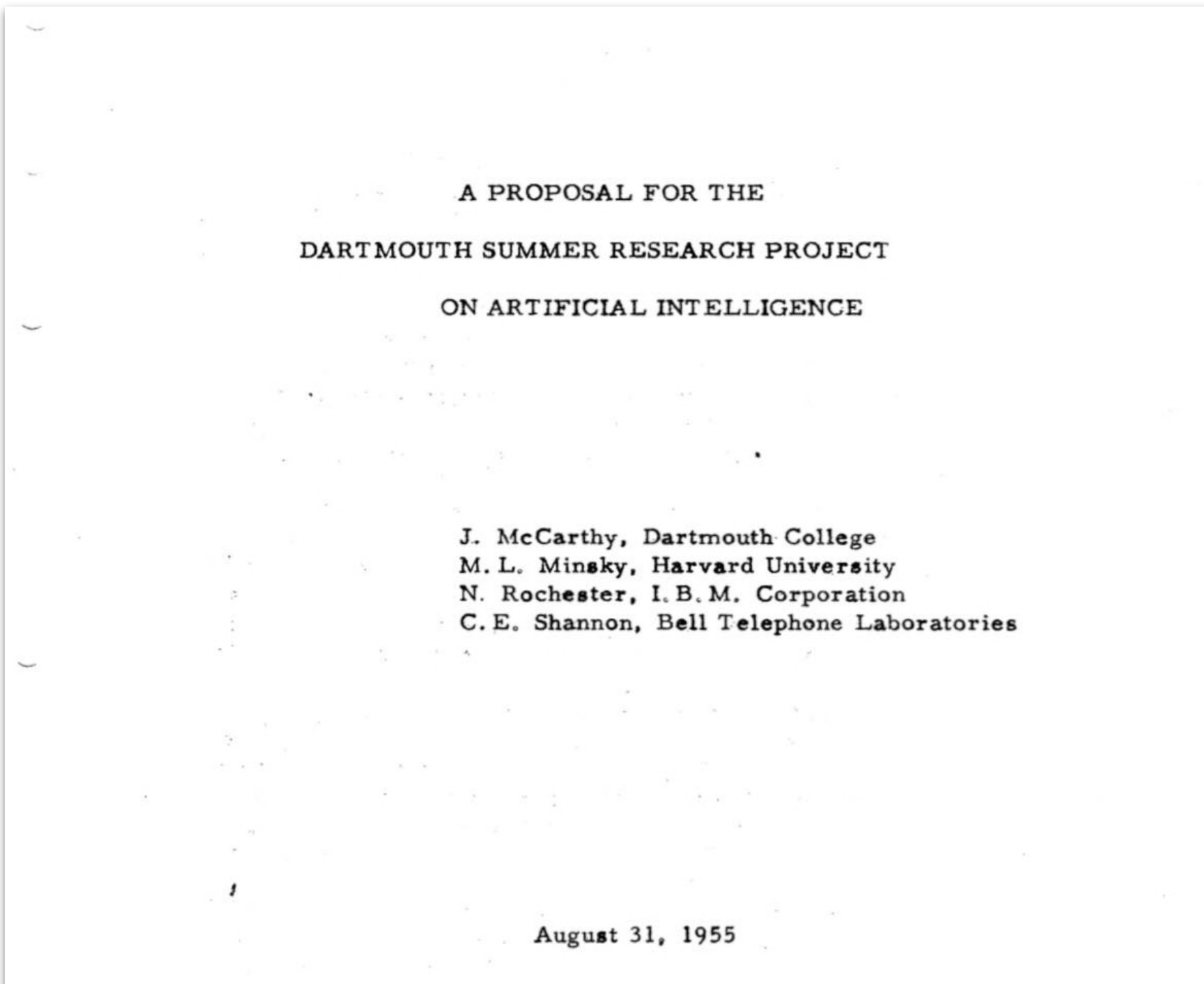
第4世代AIと医学生物学データ

清田　純　医師・博士（医学）・個人情報保護士

理化学研究所
　数理創造研究センター・数理展開部門・医科学深層学習チーム チームディレクター
　生命医科学研究センター・統合ゲノミクス研究チーム チームディレクター
　マルチオミクス基盤部門・プロテオーム解析ユニット ユニットリーダー
　予測医学特別プロジェクト・医療データ深層学習特別チーム チームディレクター
　情報統合本部・基盤研究開発部門・医科学データ共有開発ユニット ユニットリーダー
筑波大学　卓越大学院ヒューマニクス学位プログラム　教授
千葉大学医学部　客員教授
国立遺伝学研究所　特命教授
avatarin株式会社　アドバイザー
日本ディープラーニング協会　有識者会員

AI（人工知能）とは

"the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et al., 1956). Later defined as the "capacity of computers or other machines to exhibit or simulate intelligent behaviour" (OED)



A PROPOSAL FOR THE

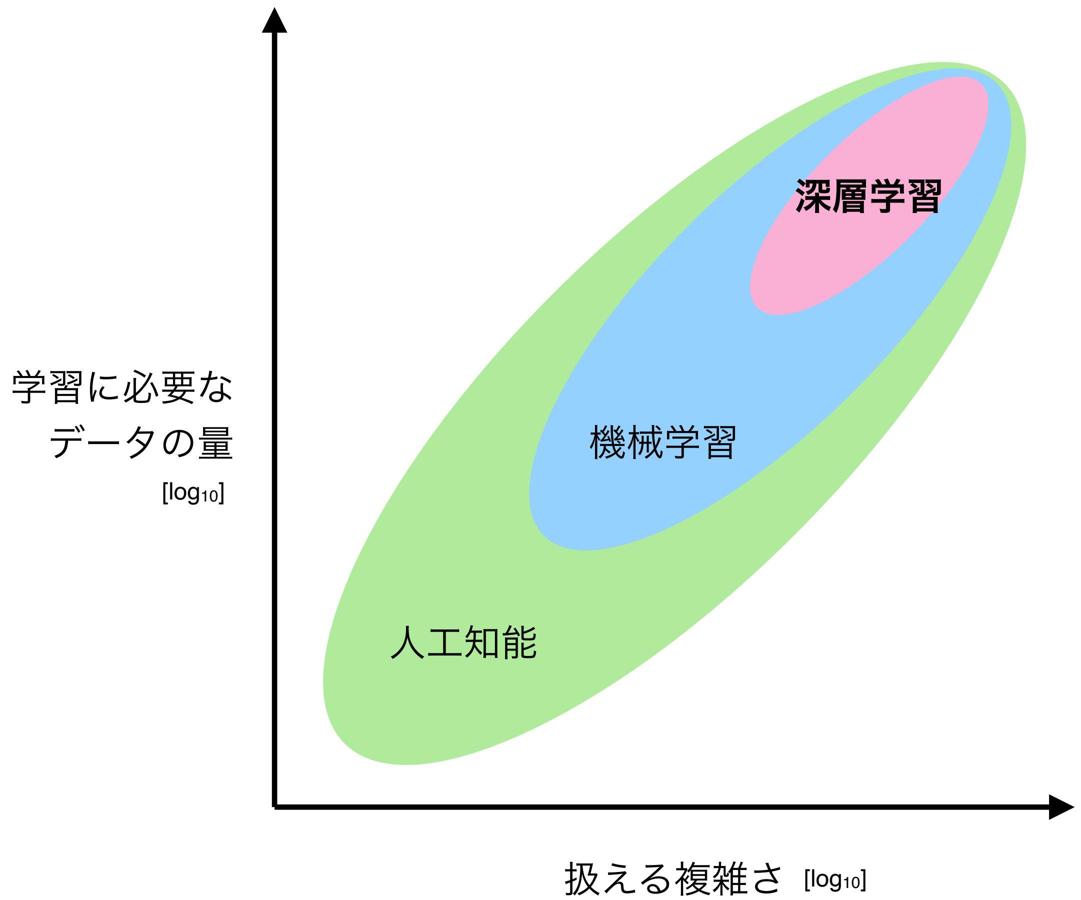
DARTMOUTH SUMMER RESEARCH PROJECT

ON ARTIFICIAL INTELLIGENCE

J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I. B. M. Corporation
C. E. Shannon, Bell Telephone Laboratories

August 31, 1955

3rd Generation AI: Deep Learning



深層学習はこれまでに無い複雑な問題を扱えるが、大量のデータが必要になる。

深層学習のわずか10年での進展

2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023

AlexNet 画像認識

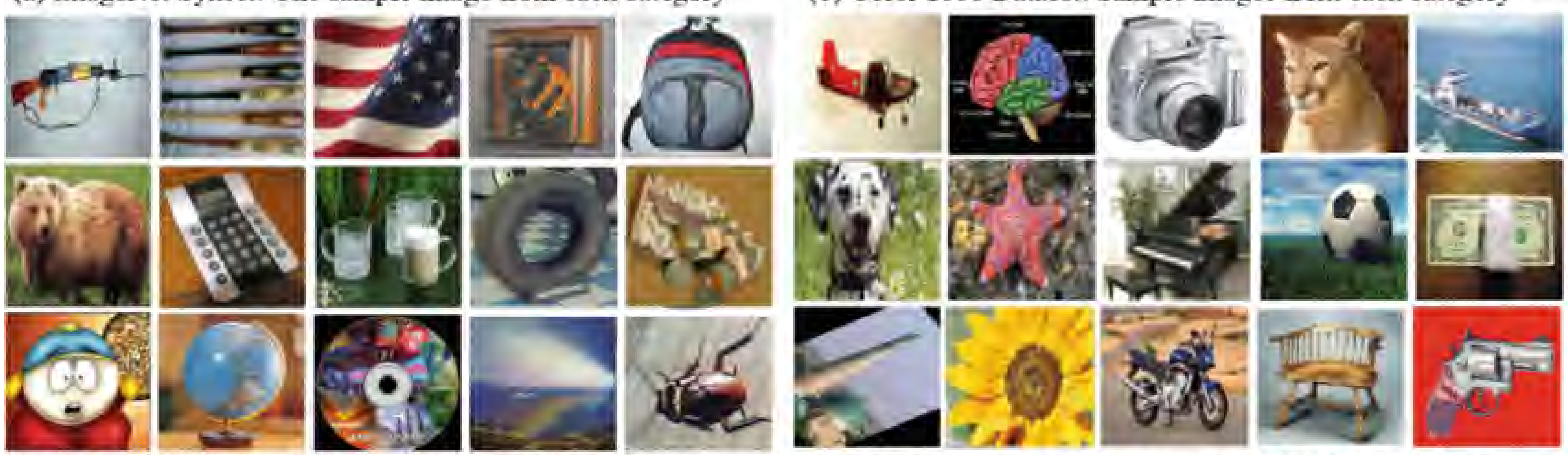
トロント大学

画像認識

ImageNet

14M images tagged by human.

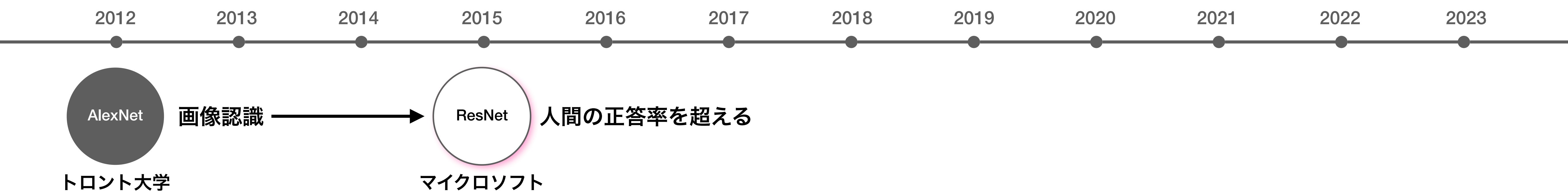
(a) ImageNet Synset: One sample image from each category (b) Corel-1000 Dataset: Sample images from each category



(c) Caltech-256 Dataset: One sample image from each category (d) Caltech-101 Dataset: One sample image from each category

http://image-net.org

深層学習のわずか10年での進展



深層学習による画像診断

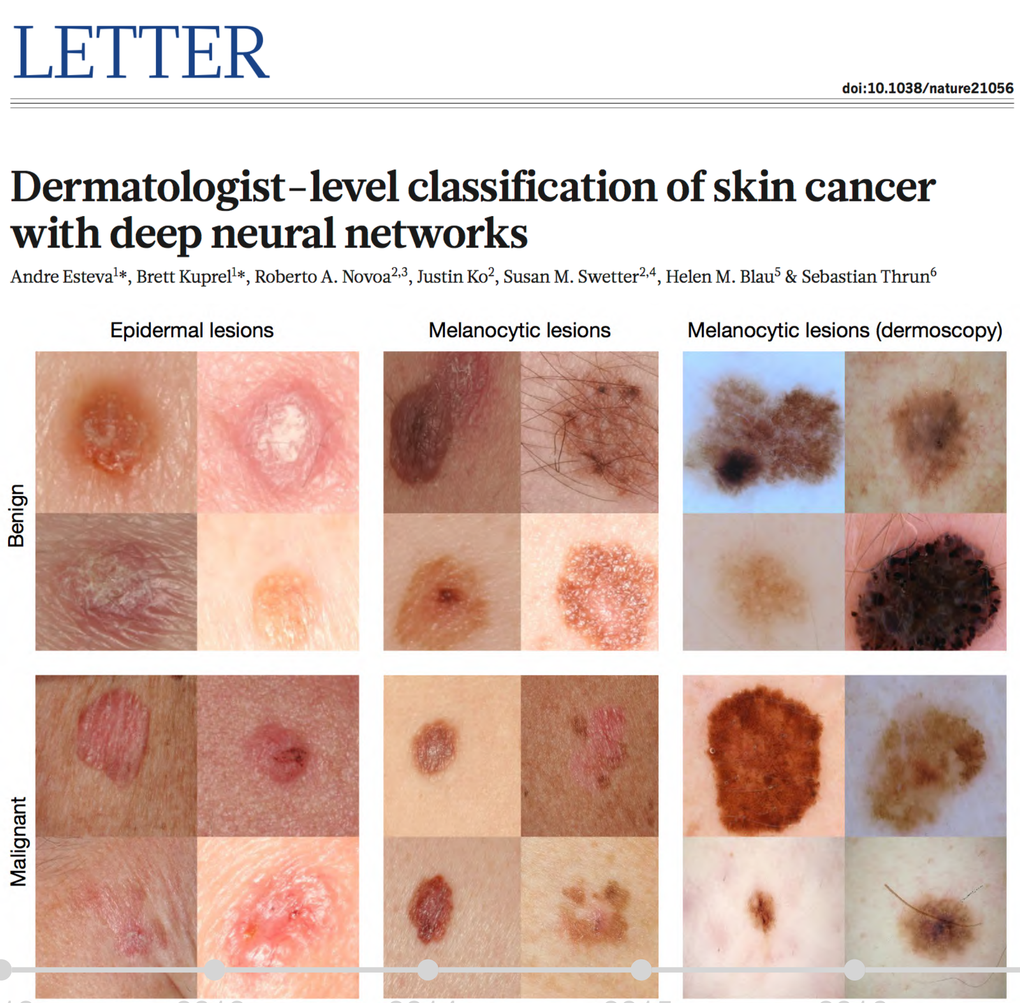
2017

Stanford大学

２０種類の皮膚疾患

約13万枚の画像を学習

専門医と同等の正答率で診断



LETTER

doi:10.1038/nature21056

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

2018

DeepMind / Google

眼底検査の光干渉断層計（OCT）

15000件のデータを学習

約50の疾患を特定可能

DeepMind's AI can detect over 50 eye diseases as accurately as a doctor

The system analyzes 3D scans of the retina and could help speed up diagnoses in hospitals

By James Vincent | Aug 13, 2018, 11:01am EDT

SHARE

2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023

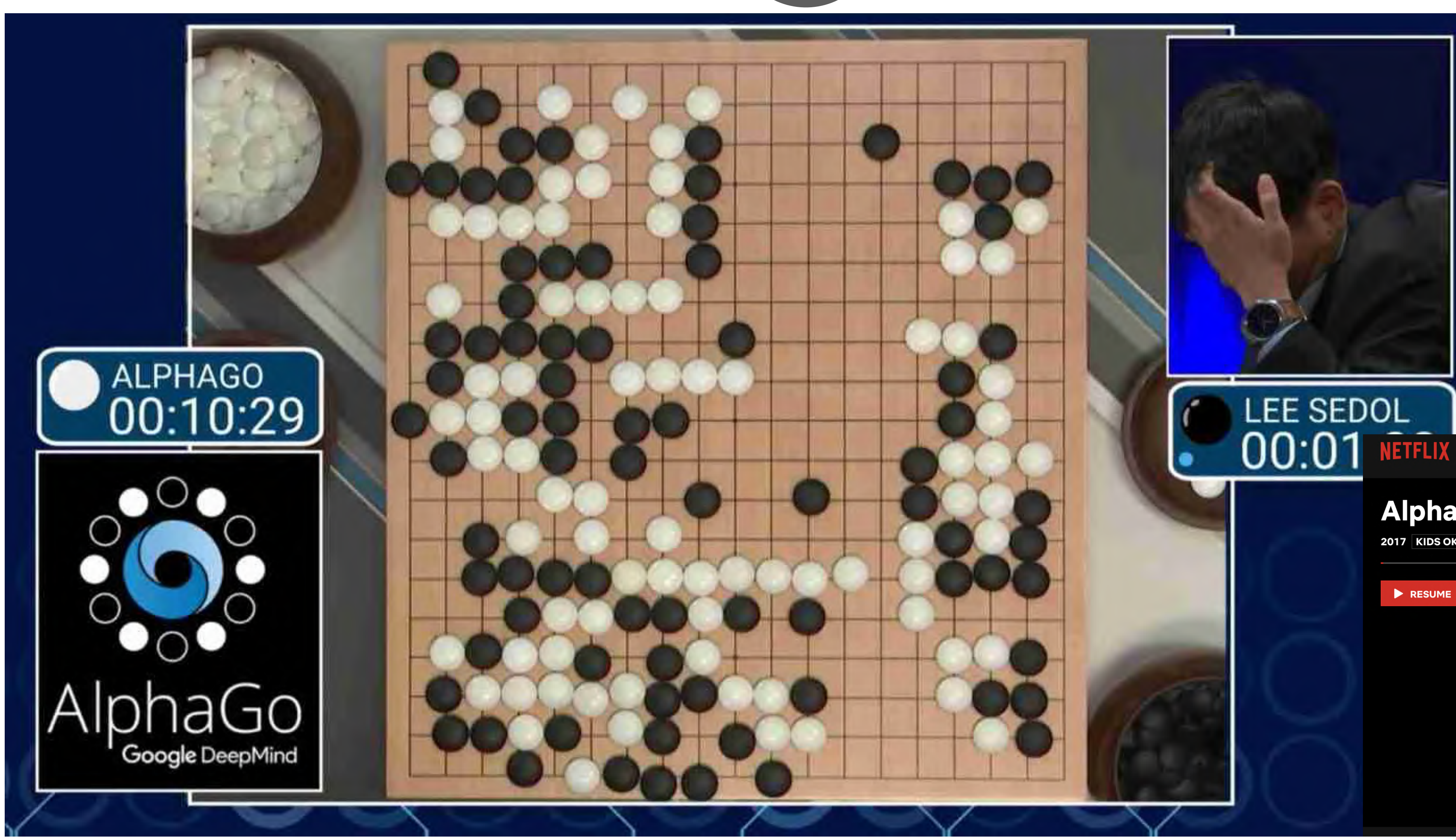
深層学習のわずか10年での進展

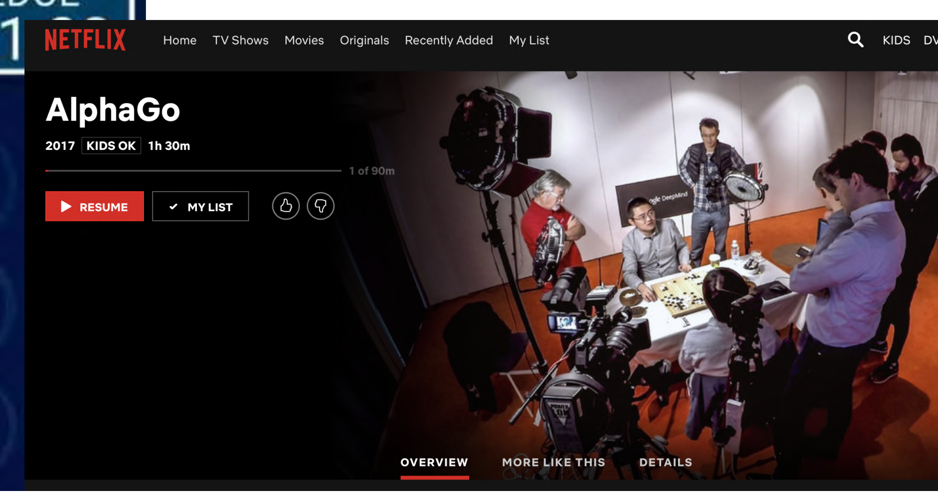
2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023

AlexNet 画像認識 → ResNet 人間の正答率を超える

トロント大学 マイクロソフト

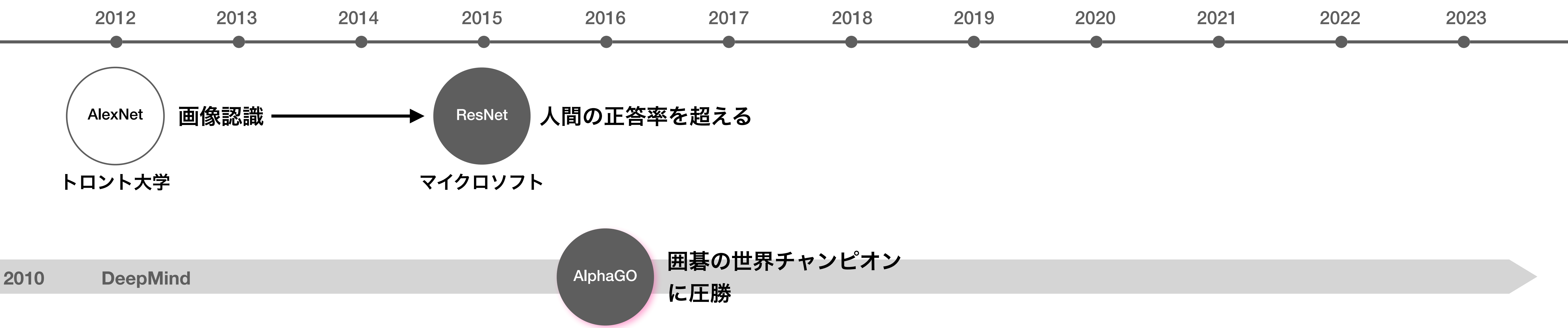
AlphaGO 囲碁の世界チャンピオンに圧勝



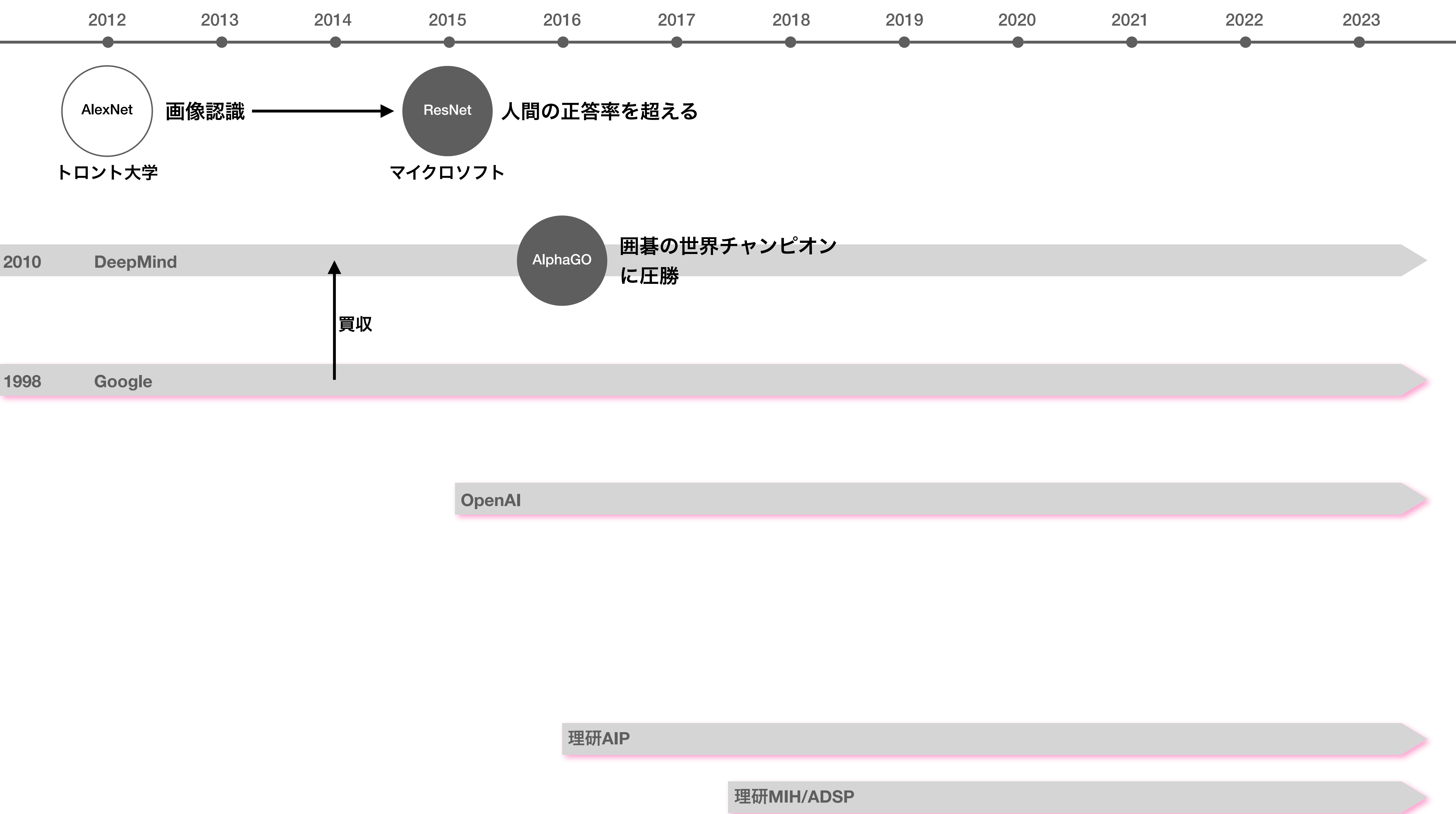


Google DeepMind AlphaGo 2016年3月対李世乭4勝1敗0引き分け

深層学習のわずか10年での進展



深層学習のわずか10年での進展



深層学習のわずか10年での進展

2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023

AlexNet 画像認識 → ResNet 人間の正答率を超え

トロント大学

マイクロソフト

AlphaGO 囲碁の
に圧勝

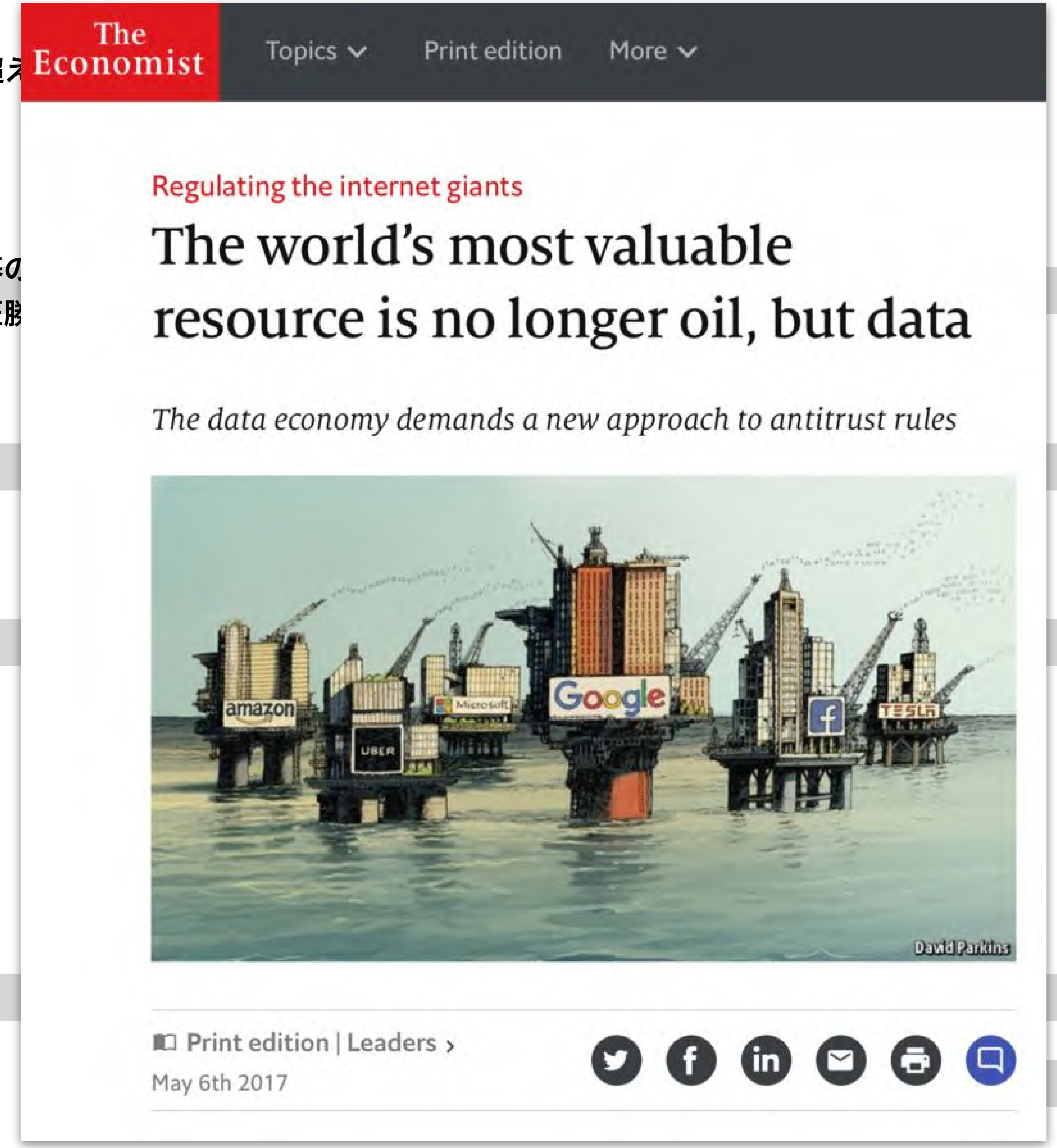
2010 DeepMind

買収

1998 Google

OpenAI

理研AIP


The Economist Topics Print edition More

Regulating the internet giants

The world's most valuable resource is no longer oil, but data

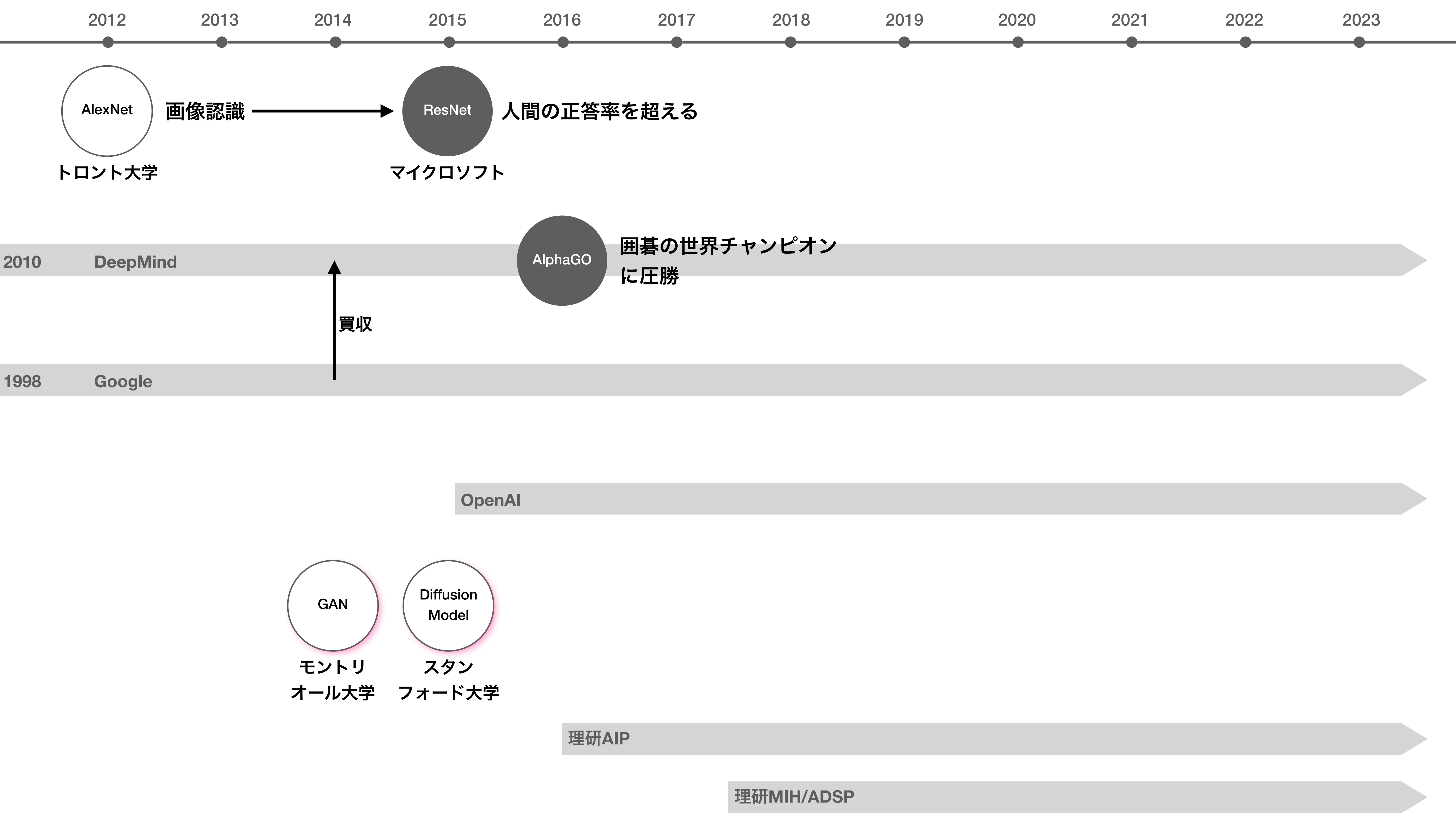
The data economy demands a new approach to antitrust rules

Print edition | Leaders ›

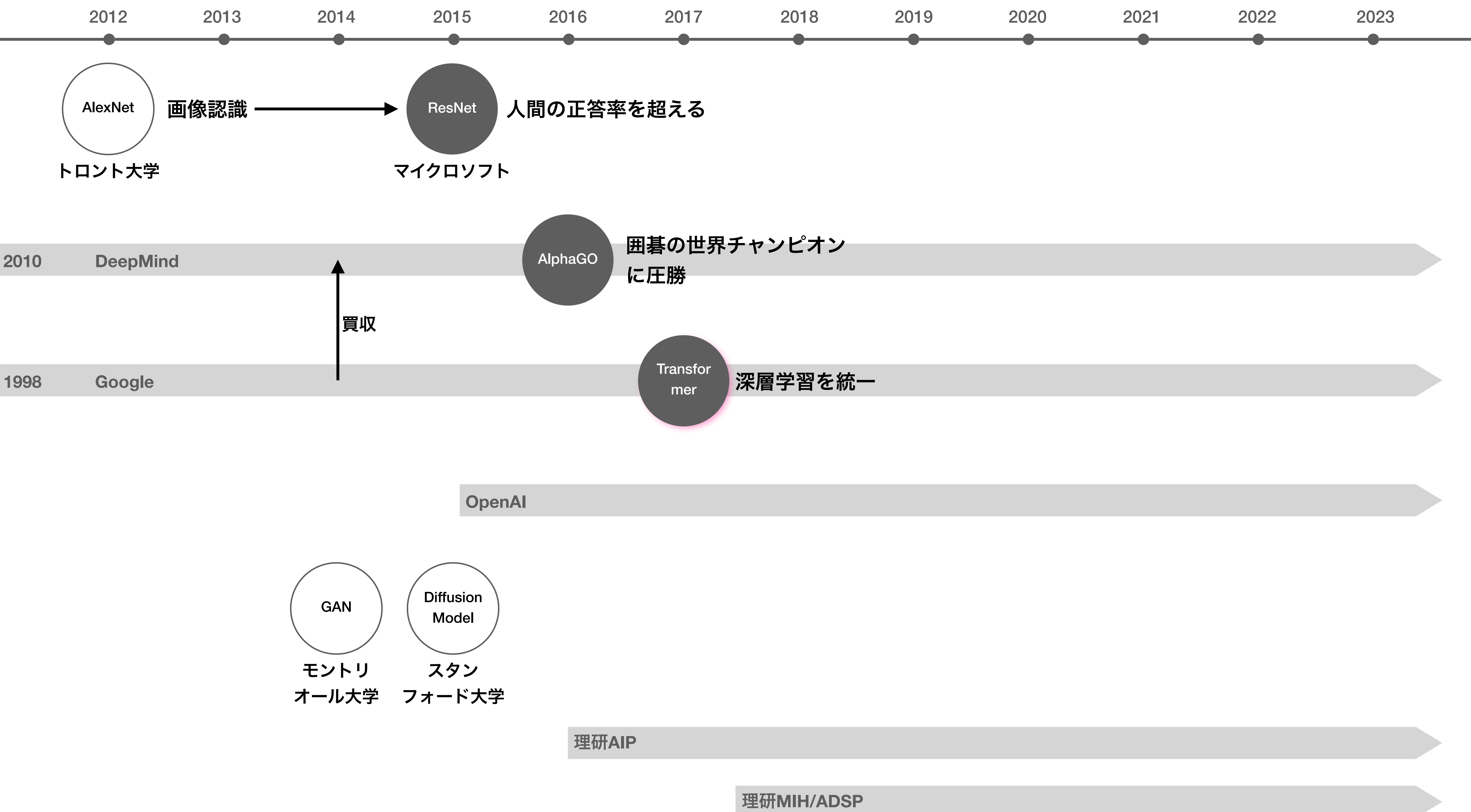
May 6th 2017

The Economist May 6th 2017

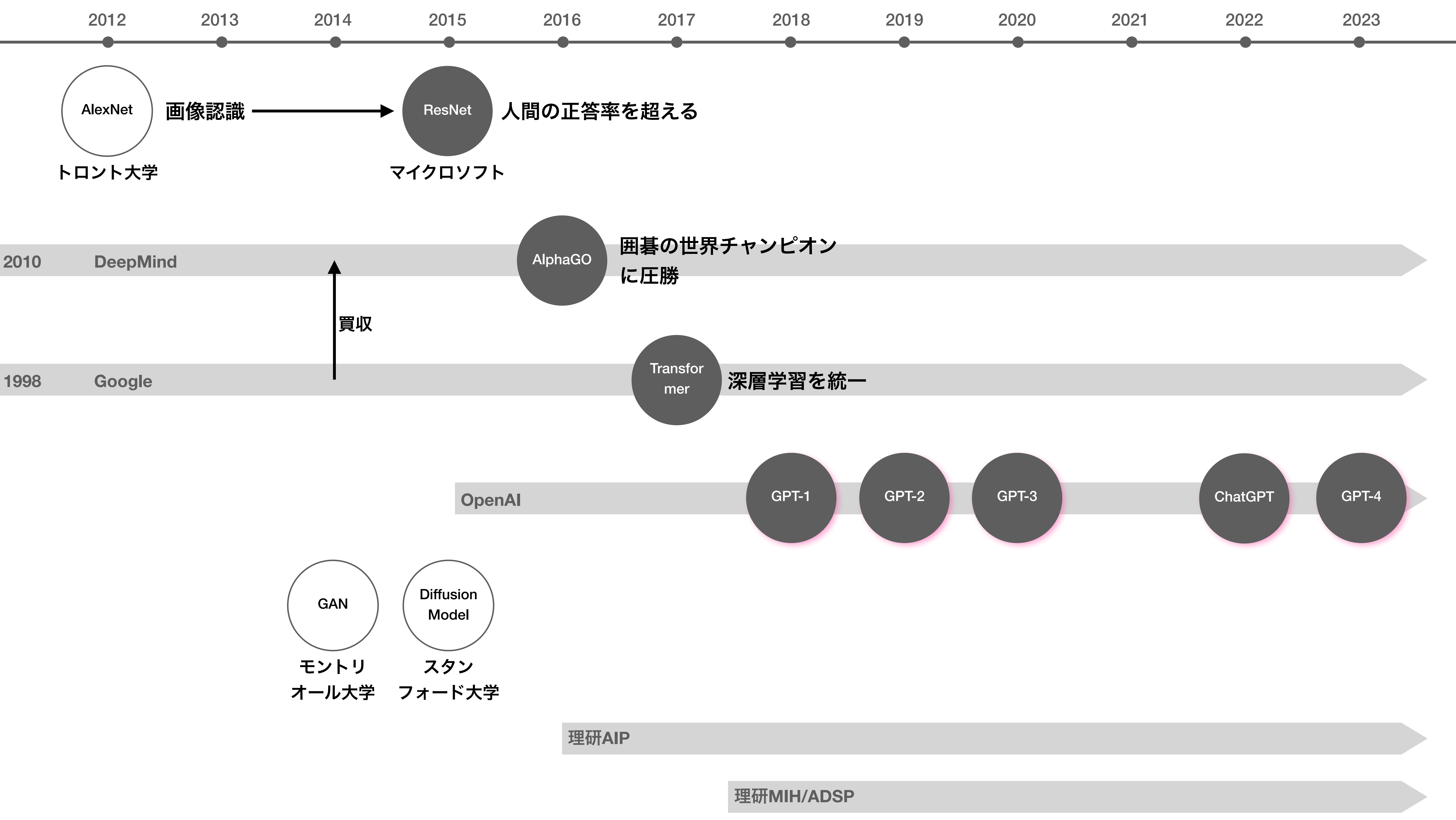
深層学習のわずか10年での進展



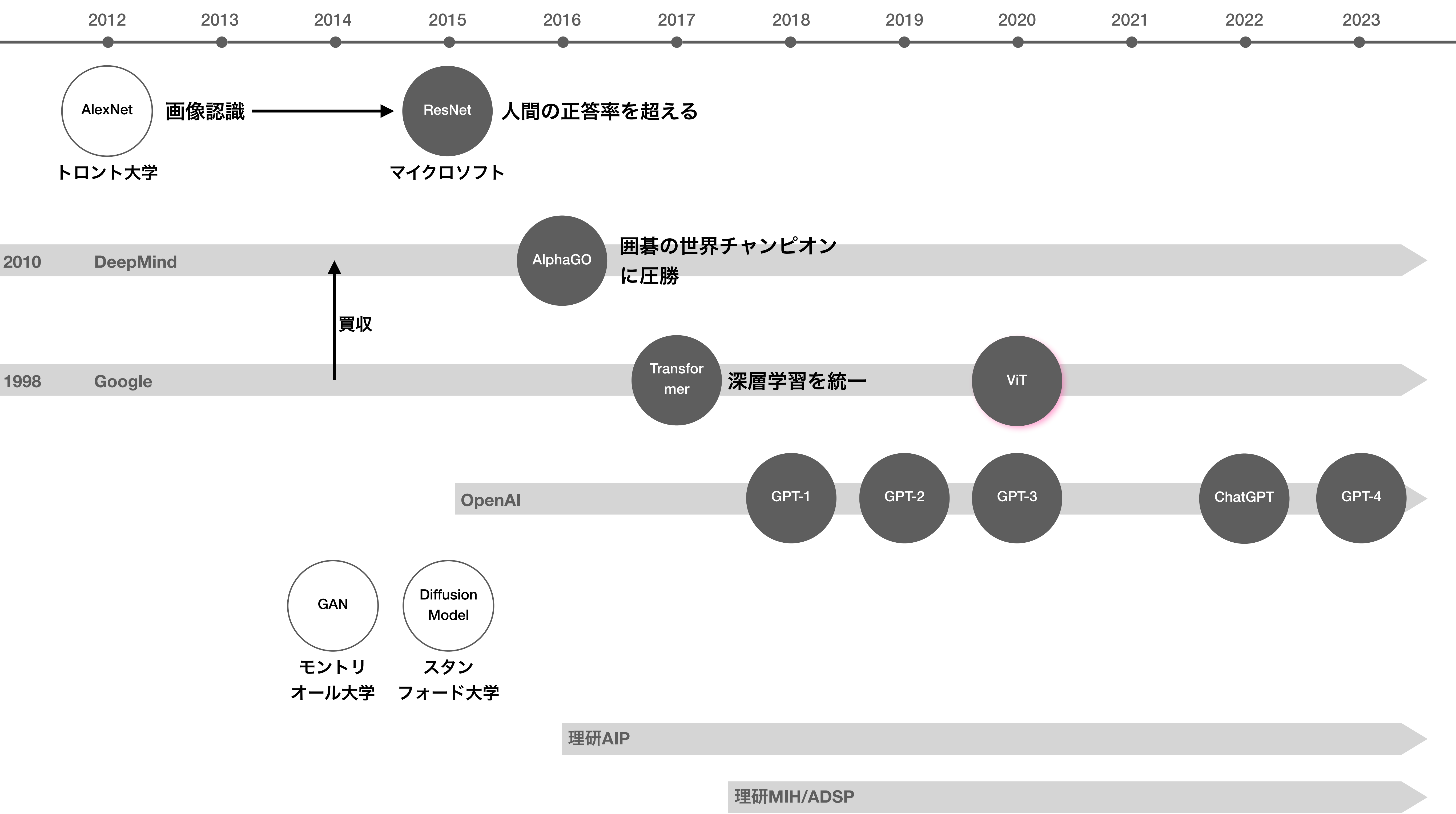
深層学習のわずか10年での進展



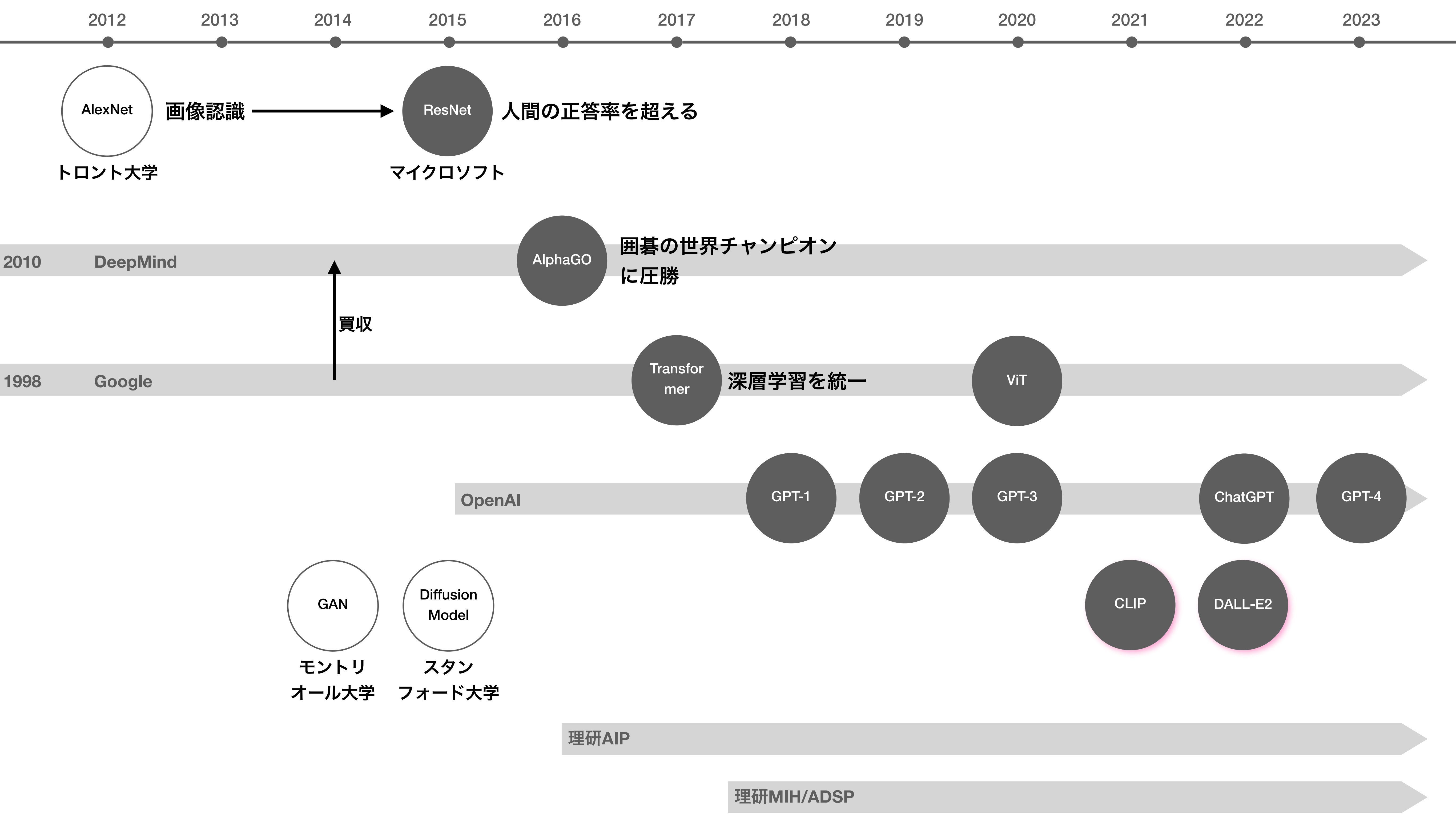
深層学習のわずか10年での進展



深層学習のわずか10年での進展



深層学習のわずか10年での進展



2022年11月にとんでもないことが起こった

JS 人工知能技術のライフサイエンス分野への応用について教えてください

人工知能 (AI) 技術は、ライフサイエンス分野で幅広い応用が可能です。例えば、遺伝子配列解析、薬物開発、疾病の予防や診断、治療の最適化などで使用されます。AIは、大量のデータを処理し、パターンを見つけることができるため、医療イメージ診断や疾患の予測にも使用されます。また、バイオインフォマティクスにおいては、遺伝子やタンパク質の予測、解析にも使用されます。

JS AIをつかった遺伝子配列解析の具体例は？

遺伝子配列解析において人工知能 (AI) 技術は、解析の自動化や精度の向上、新しい発見のために使用されます。

- ゲノムアセンブリ: AIアルゴリズムを使用して、大量の配列リードからゲノム全体を構築することができます。
- 遺伝子の予測: AIを使用して、ゲノム配列から遺伝子を予測することができ、新しい遺伝子を発見することができます。
- 機能の予測: AIを使用して、遺伝子やタンパク質の機能を予測することができます。
- 疾病の予防や診断: AIを使用して、遺伝子配列データから疾病のリスクを予測し、早期発見や予防に役立てることができます。

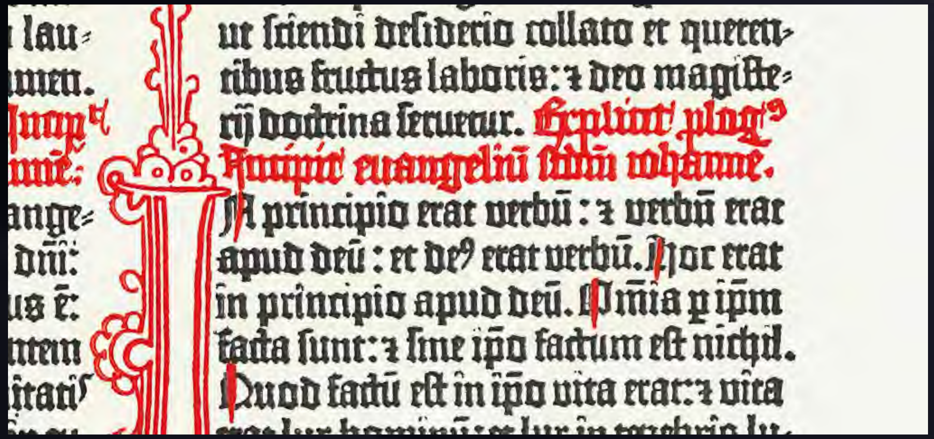
これらは、遺伝子配列解析においてAIが使用される一部の例です。現在も研究が進められており、新しい応用が開発されています。

ChatGPTが明らかにしたこと

- 知能は言語空間に存在した

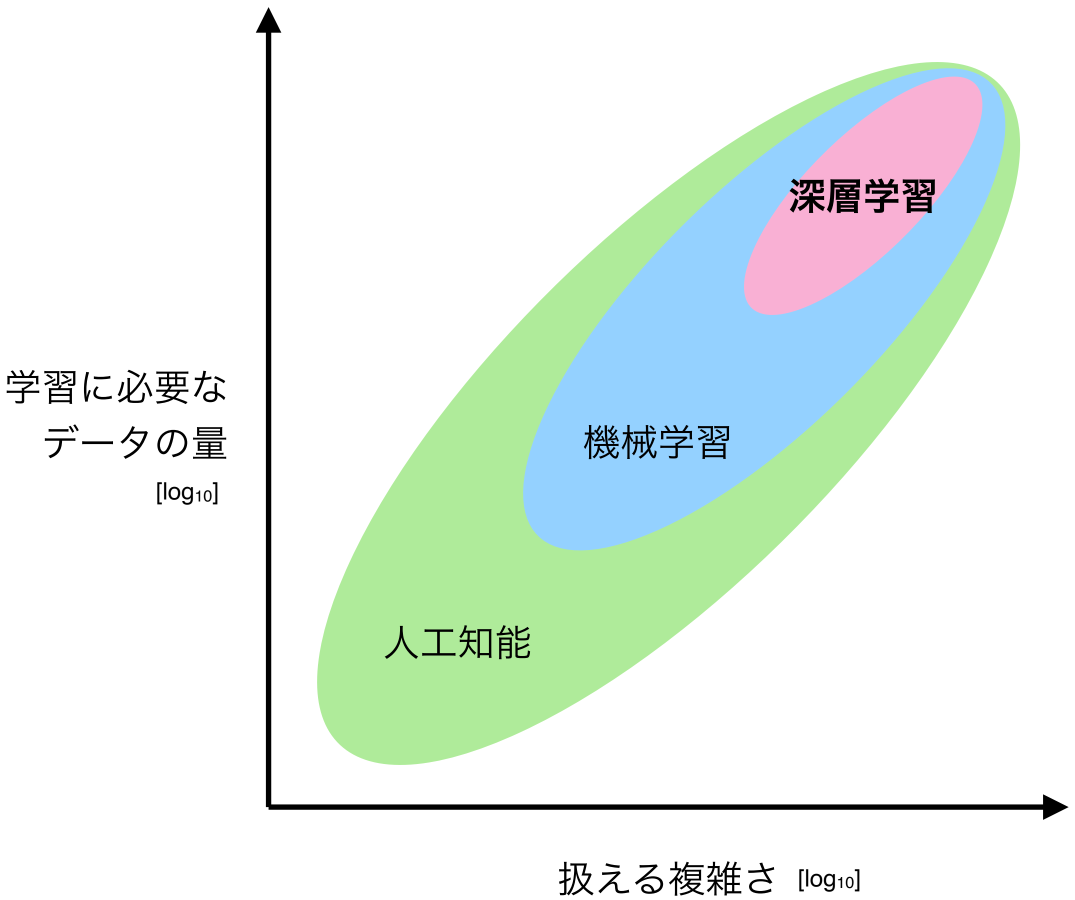
「初めに、ことばがあった。ことばは神とともにあった。ことばは神であった。」

In the beginning the Word already existed. The Word was with God, and the Word was God.


ut ſciendi deſiderio collato et queren-
tibus fructus laboris: ⁊ deo magiſte-
rij doctrina ſeruetur. Explicit plogus
Incipit euangeliũ ſcdm iohãnem.
In principio erat uerbũ: ⁊ uerbũ erat
apud deũ: et de⁹ erat uerbũ. Hoc erat
in principio apud deũ. Omnia p ipm
facta ſunt: ⁊ ſine ipo factum eſt nichil.
Quod factũ eſt in ipo uita erat: ⁊ uita

新約聖書「ヨハネによる福音書」冒頭

3rd Generation AI: Deep Learning



深層学習はこれまでに無い複雑な問題を扱えるが、大量のデータが必要になる。

スケール則はまだまだ有効

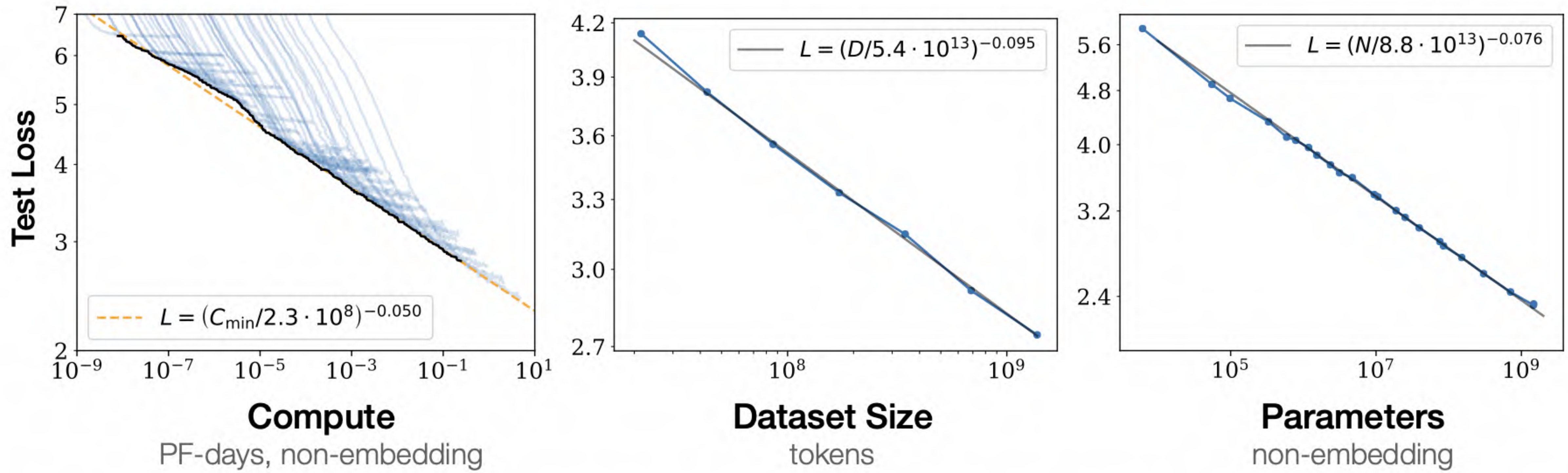
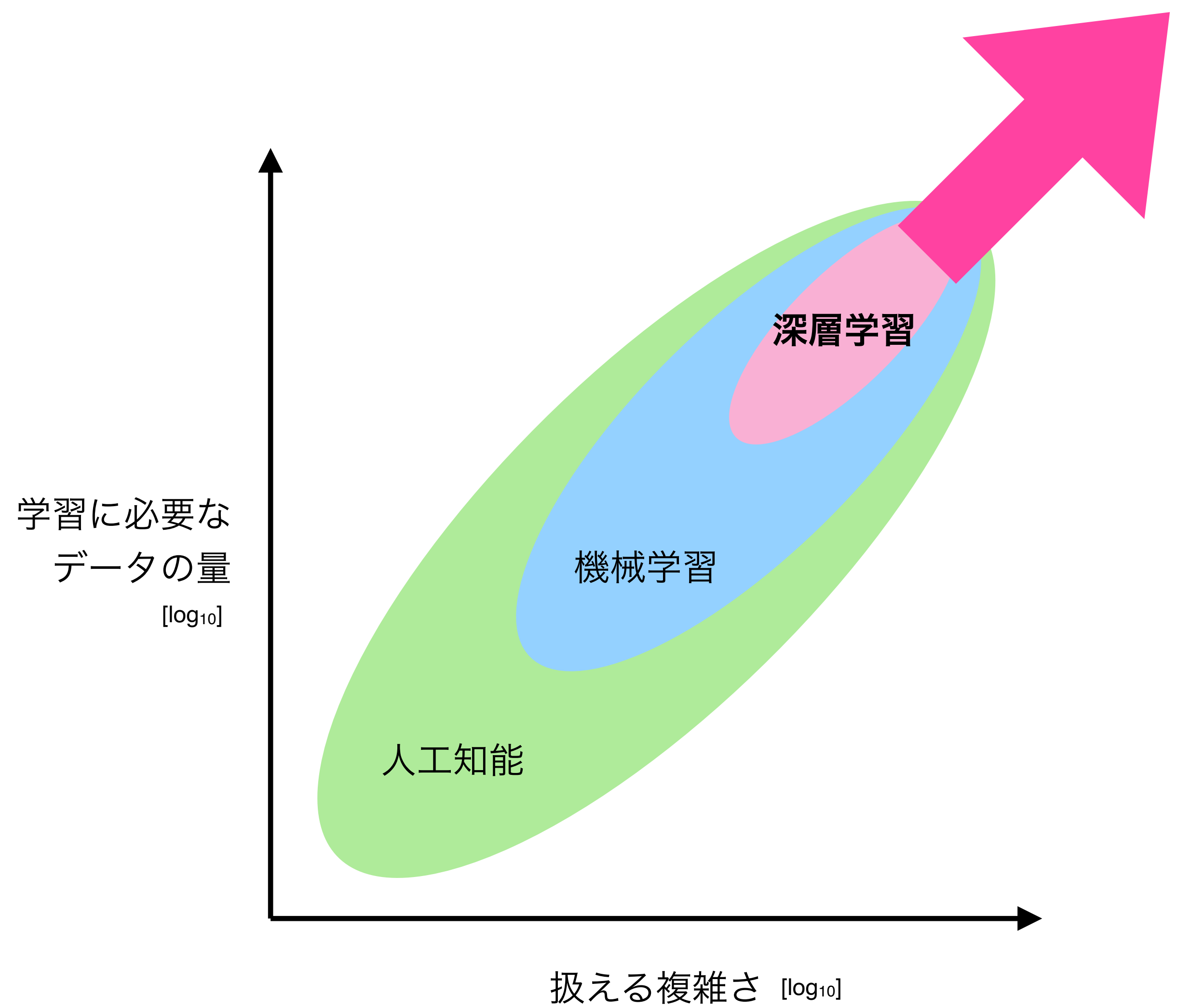


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

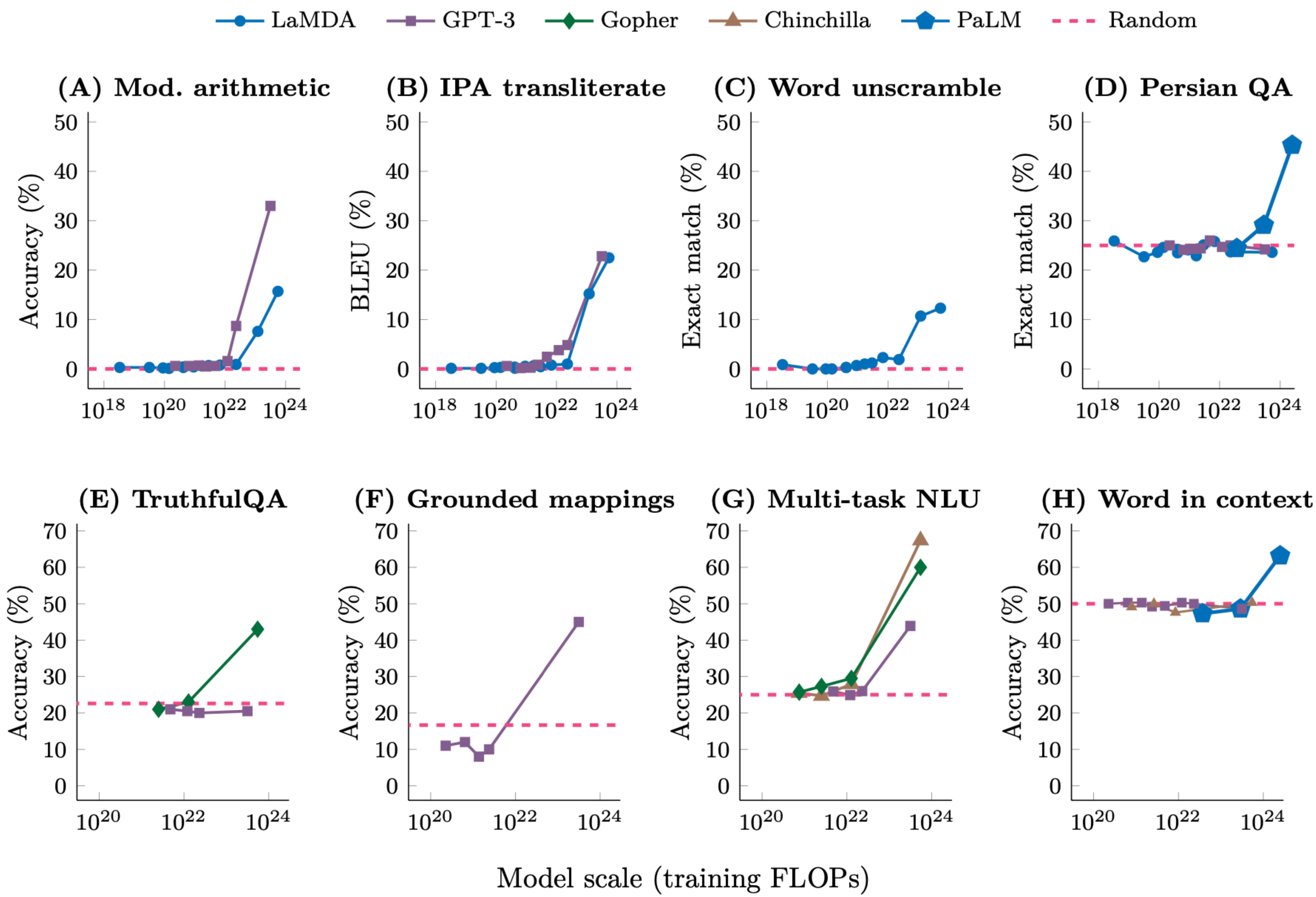
- GPT-3は、インターネットから収集した4兆単語で事前学習。1750億パラメータ。１回の学習に数億〜数十億円

More Data, More Intelligent

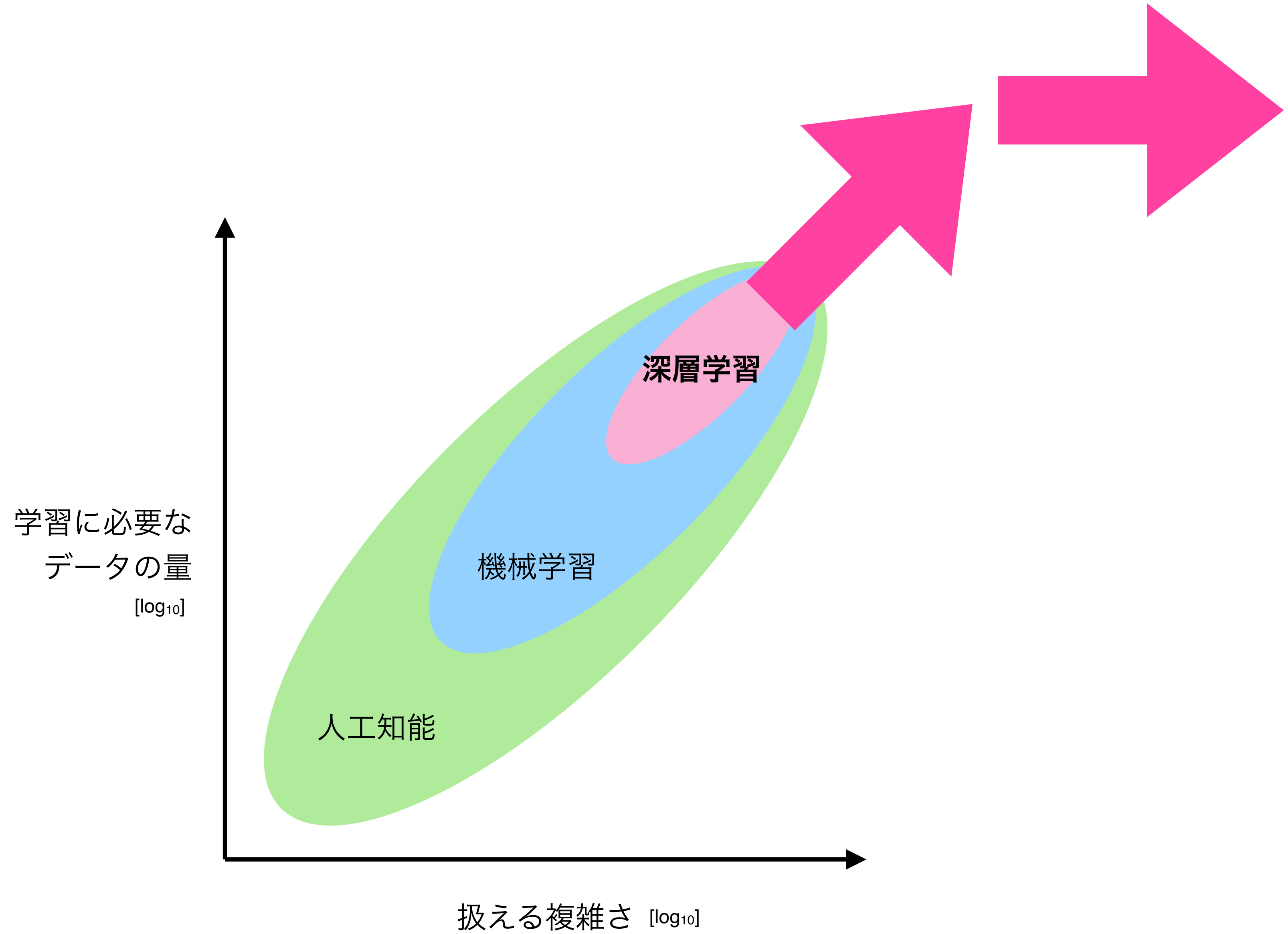


深層学習はこれまでに無い複雑な問題を扱えるが、大量のデータが必要になる。

学習があるスケールを超えるとより高度なことが可能に



Enough Data, More Intelligent



深層学習はこれまでに無い複雑な問題を扱えるが、大量のデータが必要になる。

医学の専門家「とはいえ専門知識は知らないでしょ？」

Article

Large language models encode clinical knowledge

https://doi.org/10.1038/s41586-023-06291-2

Received: 25 January 2023

Accepted: 5 June 2023

Published online: 12 July 2023

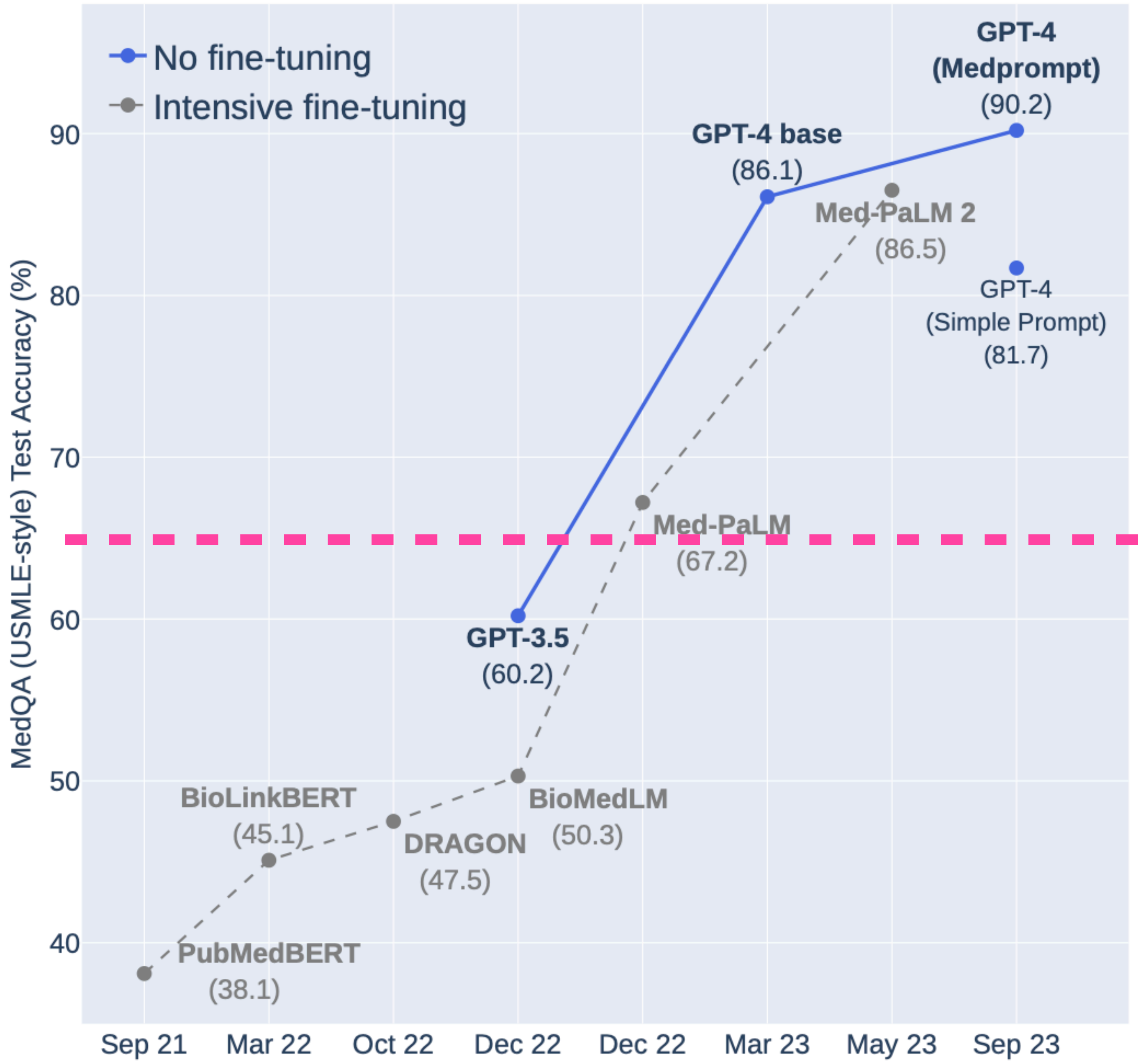
Open access

Check for updates

Karan Singhal[1,4]✉, Shekoofeh Azizi[1,4]✉, Tao Tu[1,4], S. Sara Mahdavi[1], Jason Wei[1], Hyung Won Chung[1], Nathan Scales[1], Ajay Tanwani[1], Heather Cole-Lewis[1], Stephen Pfohl[1], Perry Payne[1], Martin Seneviratne[1], Paul Gamble[1], Chris Kelly[1], Abubakr Babiker[1], Nathanael Schärli[1], Aakanksha Chowdhery[1], Philip Mansfield[1], Dina Demner-Fushman[2], Blaise Agüera y Arcas[1], Dale Webster[1], Greg S. Corrado[1], Yossi Matias[1], Katherine Chou[1], Juraj Gottweis[1], Nenad Tomasev[3], Yun Liu[1], Alvin Rajkomar[1], Joelle Barral[1], Christopher Semturs[1], Alan Karthikesalingam[1,5]✉ & Vivek Natarajan[1,5]✉

Large language models (LLMs) have demonstrated impressive capabilities, but the bar for clinical applications is high. Attempts to assess the clinical knowledge of models typically rely on automated evaluations based on limited benchmarks. Here, to address these limitations, we present MultiMedQA, a benchmark combining six existing medical question answering datasets spanning professional medicine, research and consumer queries and a new dataset of medical questions searched online, HealthSearchQA. We propose a human evaluation framework for model answers along multiple axes including factuality, comprehension, reasoning, possible harm and bias. In addition, we evaluate Pathways Language Model[1] (PaLM, a 540-billion parameter LLM) and its instruction-tuned variant, Flan-PaLM[2] on MultiMedQA. Using a combination of prompting strategies, Flan-PaLM achieves state-of-the-art accuracy on every MultiMedQA multiple-choice dataset (MedQA[3], MedMCQA[4], PubMedQA[5] and Measuring Massive Multitask Language Understanding (MMLU) clinical topics[6]), including 67.6% accuracy on MedQA (US Medical Licensing Exam-style questions), surpassing the prior state of the art by more than 17%. However, human evaluation reveals key gaps. To resolve this, we introduce instruction prompt tuning, a parameter-efficient approach for aligning LLMs to new domains using a few exemplars. The resulting model, Med-PaLM, performs encouragingly, but remains inferior to clinicians. We show that comprehension, knowledge recall and reasoning improve with model scale and instruction prompt tuning, suggesting the potential utility of LLMs in

医学知識でさらに鍛える：USMLEに合格



ところが知らないのは専門家の方だった

Can Generalist Foundation Models Outcompete
Special-Purpose Tuning? Case Study in Medicine

Harsha Nori*‡, Yin Tat Lee*, Sheng Zhang*, Dean Carignan, Richard Edgar,
Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu,
Renqian Luo, Scott Mayer McKinney†, Robert Osazuwa Ness,
Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and
Eric Horvitz‡

Microsoft

November 2023

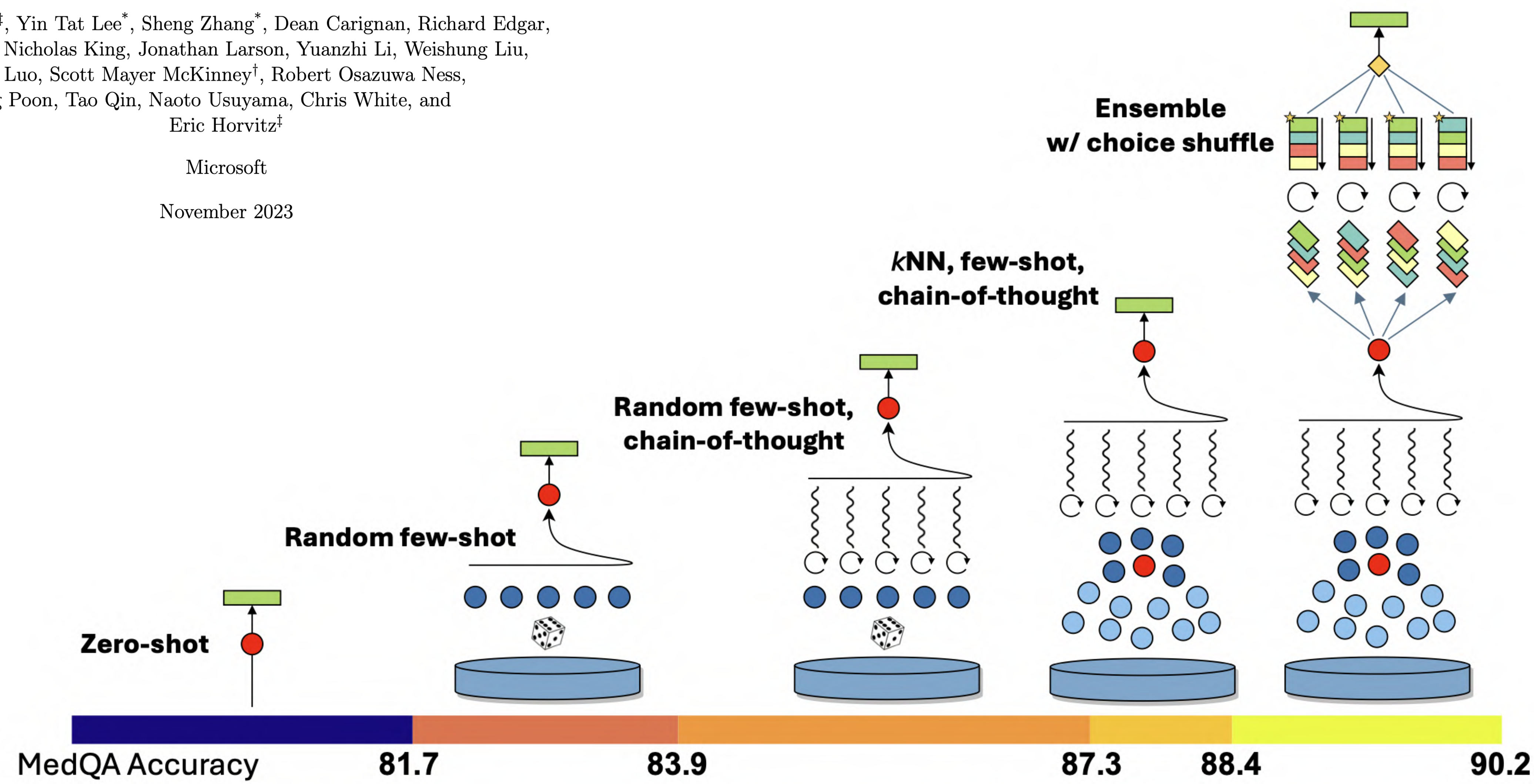


Figure 4: Visual illustration of Medprompt components and additive contributions to performance on the MedQA benchmark. The prompting strategy combines *k*NN-based few-shot example selection, GPT-4–generated chain-of-thought prompting, and answer-choice shuffled ensembling (see details in Section 4). Relative contributions of each component are shown at the bottom (details in Section 5.2).

基盤モデルとは

定義

- Stanford Institute for Human-Centered Artificial Intelligenceが2021年に国際会議を開催し導入した概念。
    - “A large artificial intelligence model trained on a vast quantity of unlabeled data at scale (usually by self-supervised learning) resulting in a model that can be adapted to a wide range of downstream tasks.”
    - 大量のデータを学習することにより、幅広い下流タスクをこなすことができるようになった大型のAIモデル。

なぜ大規模言語モデルは基盤モデルになれたのか

1. 大量のデータが存在した
2. 大量のデータを学習可能な深層学習手法（Transformer）が開発された
3. 「次の単語を予測しろ」というタスクに適した巨大なニューラルネットワークの最終段に、目的のタスクに応じた小さなニューラルネットワークを付加するだけで、様々なタスクが解けるということが（ラッキーにも）発見された（ただしタスク毎に学習し直す必要がある）：広義の基盤モデル
4. タスク（人間がやってほしいこと）も言語で表現して、入力側から加えるだけで、「次の単語を予測しろ」というタスクに適した巨大なニューラルネットワーク１つあれば（学習は1回でOK）、様々なタスクを解けることが発見された：狭義の基盤モデル

基盤モデルとしての大規模言語モデル

- “Let’s think step by step.” と伝えるだけで返答内容が格段と賢くなる。

Large Language Models are Zero-Shot Reasoners

Takeshi Kojima
The University of Tokyo
t.kojima@weblab.t.u-tokyo.ac.jp

Shixiang Shane Gu
Google Research, Brain Team

Machel Reid
Google Research*

Yutaka Matsuo
The University of Tokyo

Yusuke Iwasawa
The University of Tokyo

Abstract

Pretrained large language models (LLMs) are widely used in many sub-fields of natural language processing (NLP) and generally known as excellent *few-shot* learners with task-specific exemplars. Notably, chain of thought (CoT) prompting, a recent technique for eliciting complex multi-step reasoning through step-by-step answer examples, achieved the state-of-the-art performances in arithmetics and symbolic reasoning, difficult *system-2* tasks that do not follow the standard scaling laws for LLMs. While these successes are often attributed to LLMs' ability for few-shot learning, we show that LLMs are decent *zero-shot* reasoners by simply adding “Let’s think step by step” before each answer. Experimental results demonstrate that our Zero-shot-CoT, using the same single prompt template, significantly outperforms zero-shot LLM performances on diverse benchmark reasoning tasks including arithmetics (MultiArith, GSM8K, AQUA-RAT, SVAMP), symbolic reasoning (Last Letter, Coin Flip), and other logical reasoning tasks (Date-

基盤モデルとしての大規模言語モデル

- もっと賢くなる呪文を探す

Google DeepMind

Large Language Models as Optimizers

Chengrun Yang* **Xuezhi Wang** **Yifeng Lu** **Hanxiao Liu**
Quoc V. Le **Denny Zhou** **Xinyun Chen***

{chengrun, xuezhiw, yifenglu}@google.com, 6.hanxiao@gmail.com
{qvl, dennyzhou, xinyunchen}@google.com

Google DeepMind * Equal contribution

Abstract

Optimization is ubiquitous. While derivative-based algorithms have been powerful tools for various problems, the absence of gradient imposes challenges on many real-world applications. In this work, we propose Optimization by PROmpting (OPRO), a simple and effective approach to leverage large language models (LLMs) as optimizers, where the optimization task is described in natural language. In each optimization step, the LLM generates new solutions from the prompt that contains previously generated solutions with their values, then the new solutions are evaluated and added to the prompt for the next optimization step. We first showcase OPRO on linear regression and traveling salesman problems, then move on to prompt optimization where the goal is to find instructions that maximize the task accuracy. With a variety of LLMs, we demonstrate that the best prompts optimized by OPRO outperform human-designed prompts by up to 8% on GSM8K, and by up to 50% on Big-Bench Hard tasks.

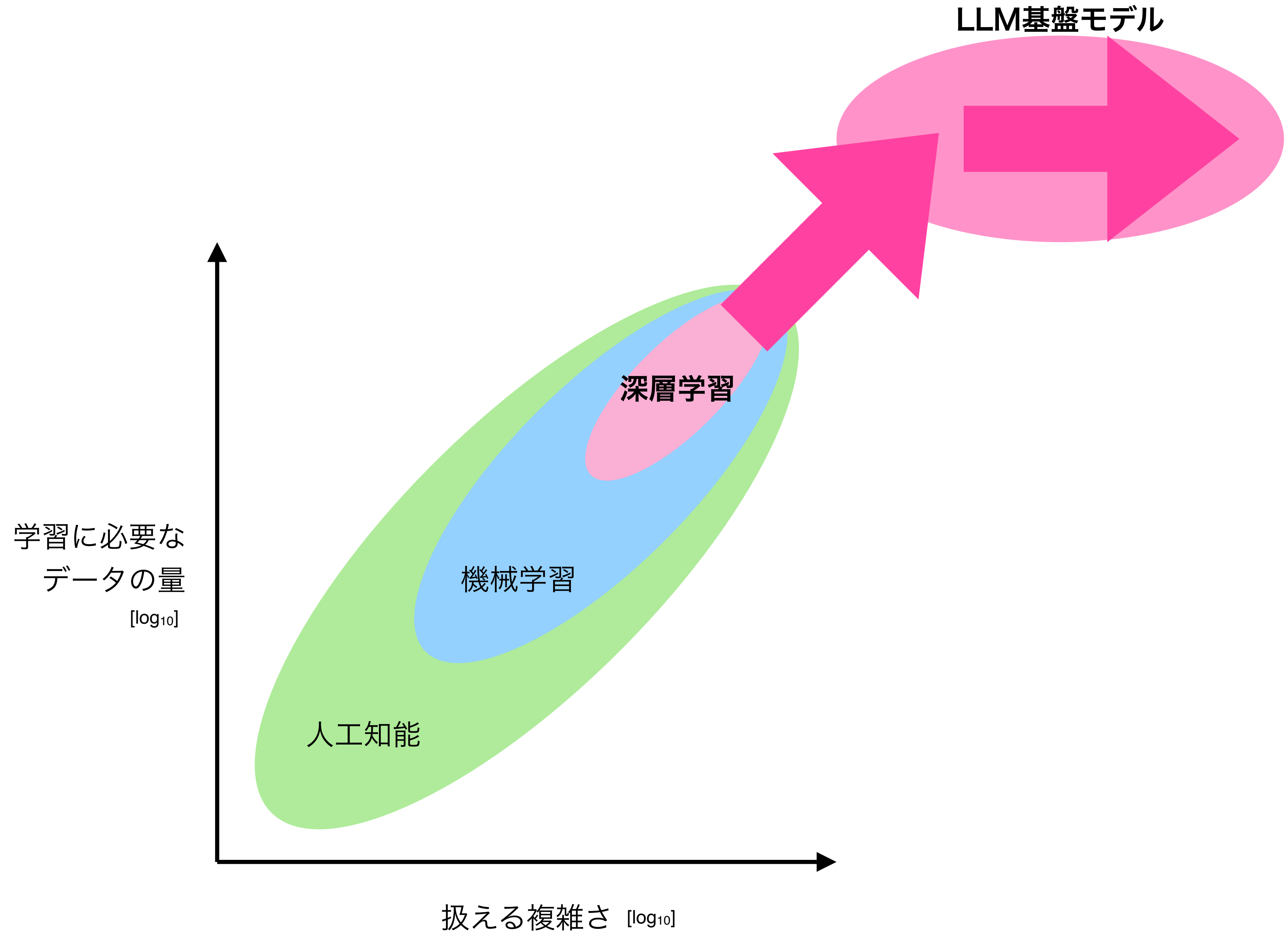
基盤モデルとしての大規模言語モデル

- “Take a deep breath and work on this problem step-by-step.”

Table 4: Test accuracies on GSM8K. We show the instruction with the highest test accuracy for each scorer-optimizer pair.

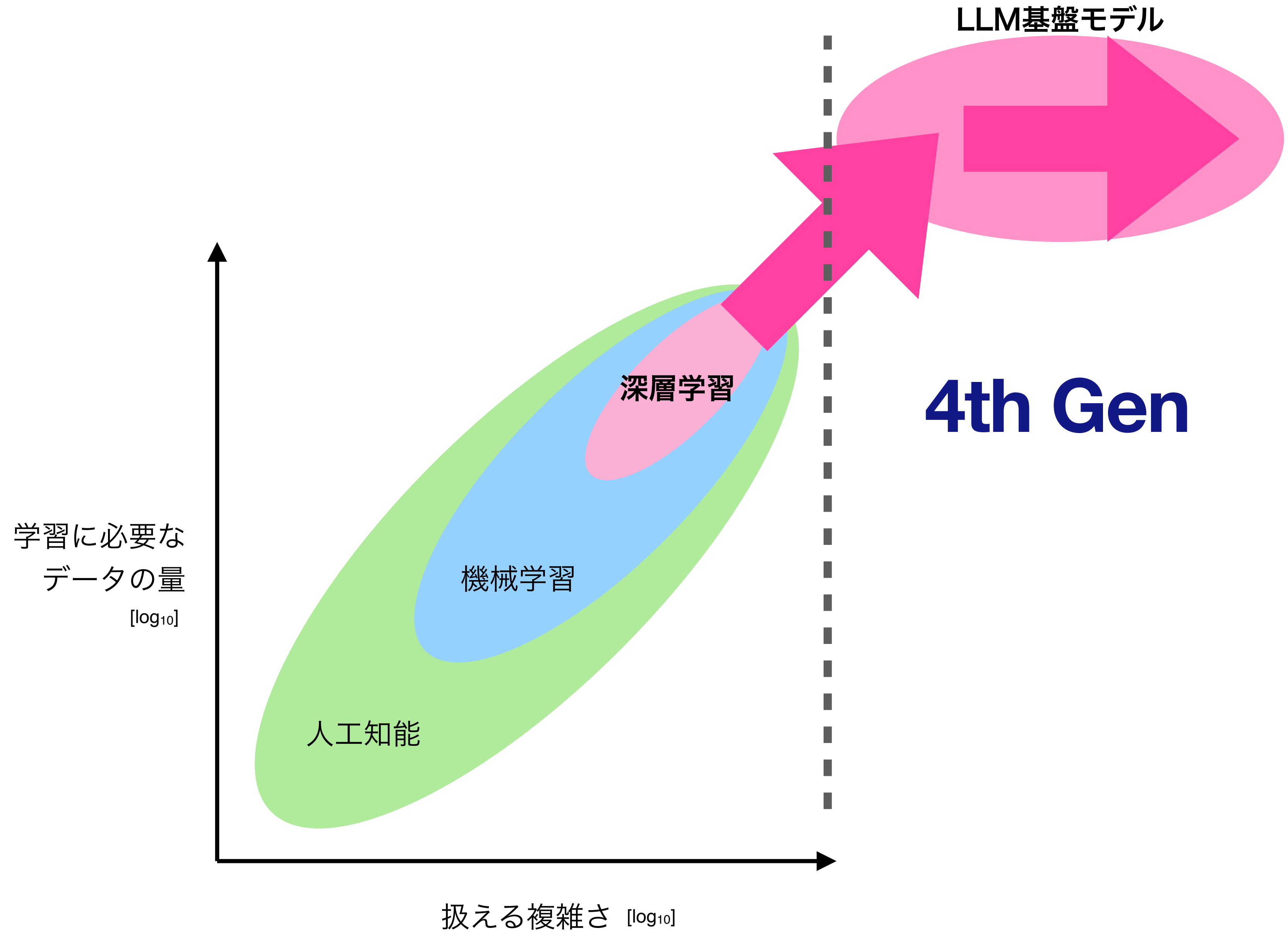
Scorer	Optimizer / Source	Instruction position	Top instruction	Acc
Baselines				
PaLM 2-L	(Kojima et al., 2022)	A_begin	Let's think step by step.	71.8
PaLM 2-L	(Zhou et al., 2022b)	A_begin	Let's work this out in a step by step way to be sure we have the right answer.	58.8
PaLM 2-L		A_begin	Let's solve the problem.	60.8
PaLM 2-L		A_begin	(empty string)	34.0
text-bison	(Kojima et al., 2022)	Q_begin	Let's think step by step.	64.4
text-bison	(Zhou et al., 2022b)	Q_begin	Let's work this out in a step by step way to be sure we have the right answer.	65.6
text-bison		Q_begin	Let's solve the problem.	59.1
text-bison		Q_begin	(empty string)	56.8
Ours				
PaLM 2-L	PaLM 2-L-IT	A_begin	Take a deep breath and work on this problem step-by-step.	**80.2**
PaLM 2-L	PaLM 2-L	A_begin	Break this down.	79.9
PaLM 2-L	gpt-3.5-turbo	A_begin	A little bit of arithmetic and a logical approach will help us quickly arrive at the solution to this problem.	78.5
PaLM 2-L	gpt-4	A_begin	Let's combine our numerical command and clear thinking to quickly and accurately decipher the answer.	74.5
text-bison	PaLM 2-L-IT	Q_begin	Let's work together to solve math word problems! First, we will read and discuss the problem together to make sure we	64.4

LLM is Foundation Model



深層学習はこれまでに無い複雑な問題を扱えるが、大量のデータが必要になる。

LLM is Foundation Model

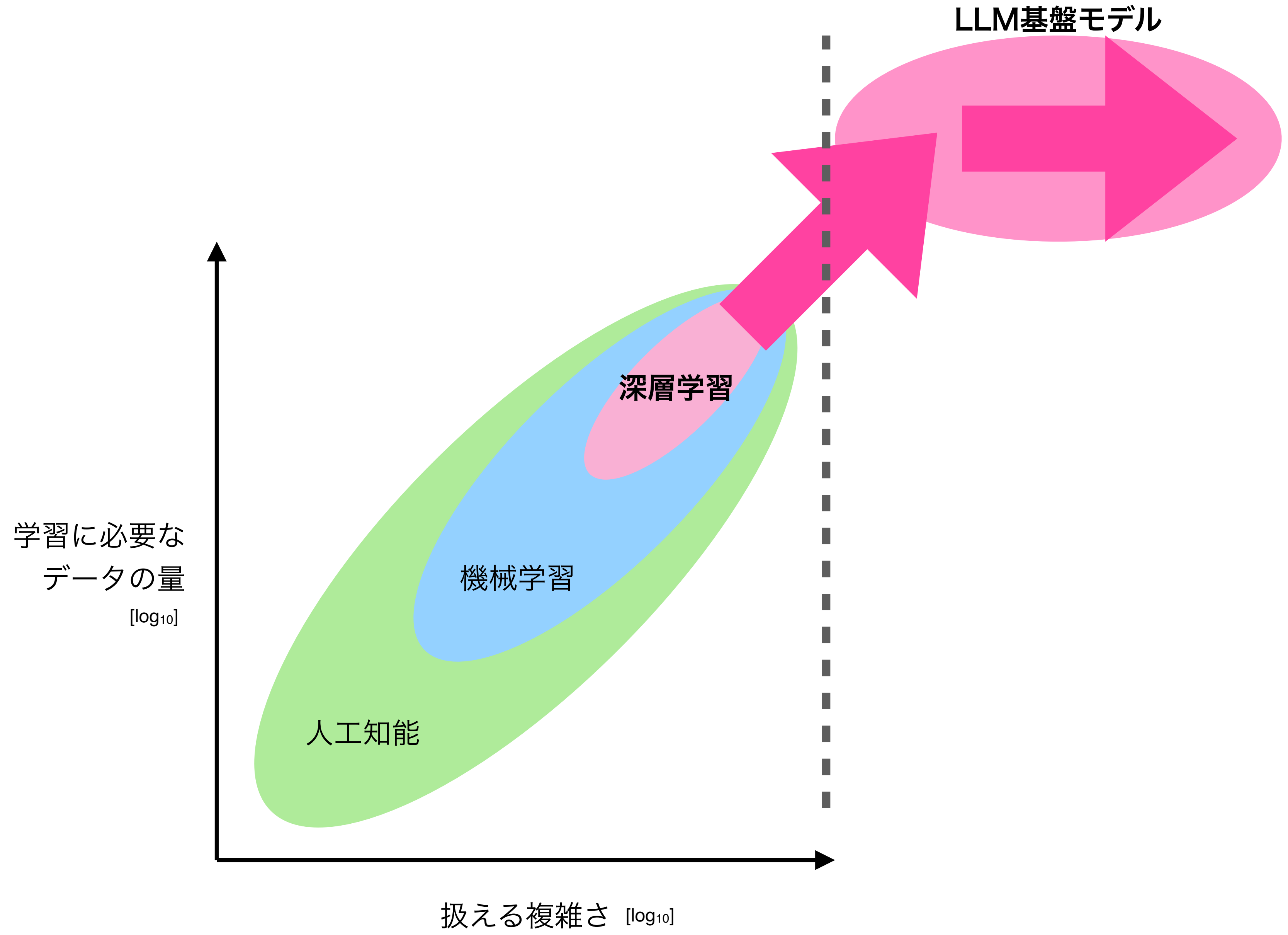


深層学習はこれまでに無い複雑な問題を扱えるが、大量のデータが必要になる。

2017

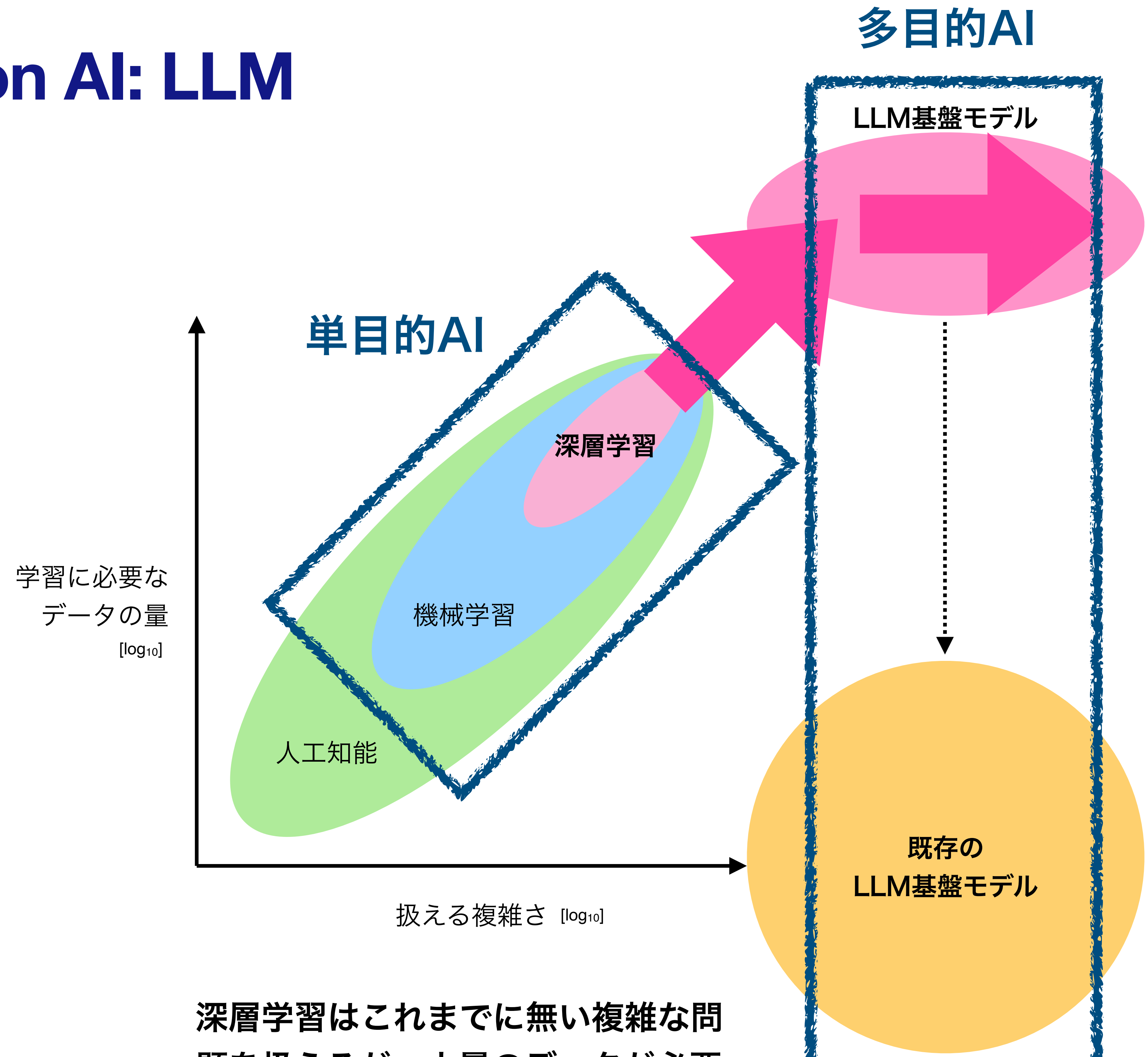
2025

4th Generation AI: LLM



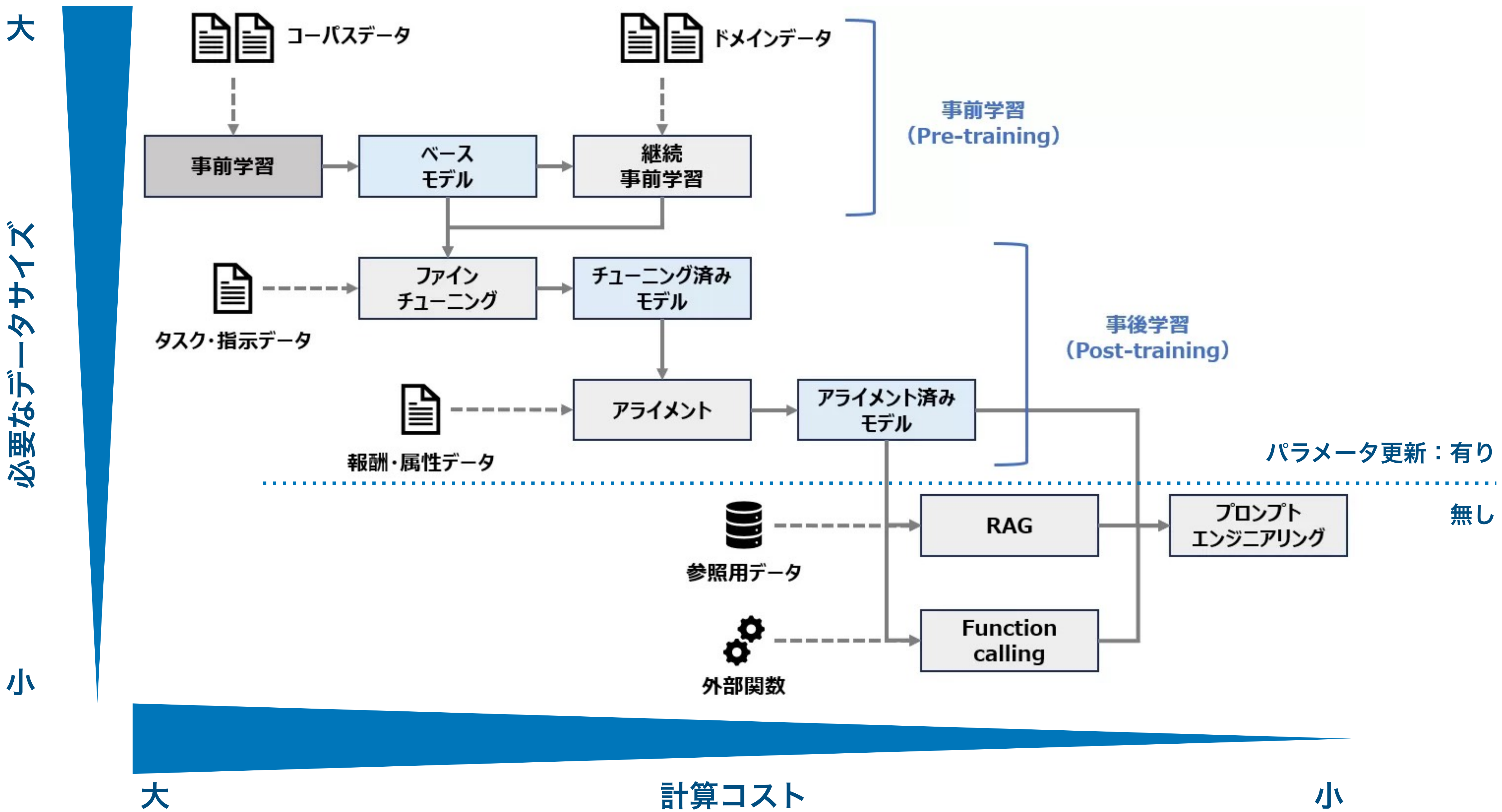
深層学習はこれまでに無い複雑な問題を扱えるが、大量のデータが必要になる。

4th Generation AI: LLM



深層学習はこれまでに無い複雑な問題を扱えるが、大量のデータが必要になる。

LLMとデータ



https://dalab.jp/archives/journal/llm-finetuning-part1/

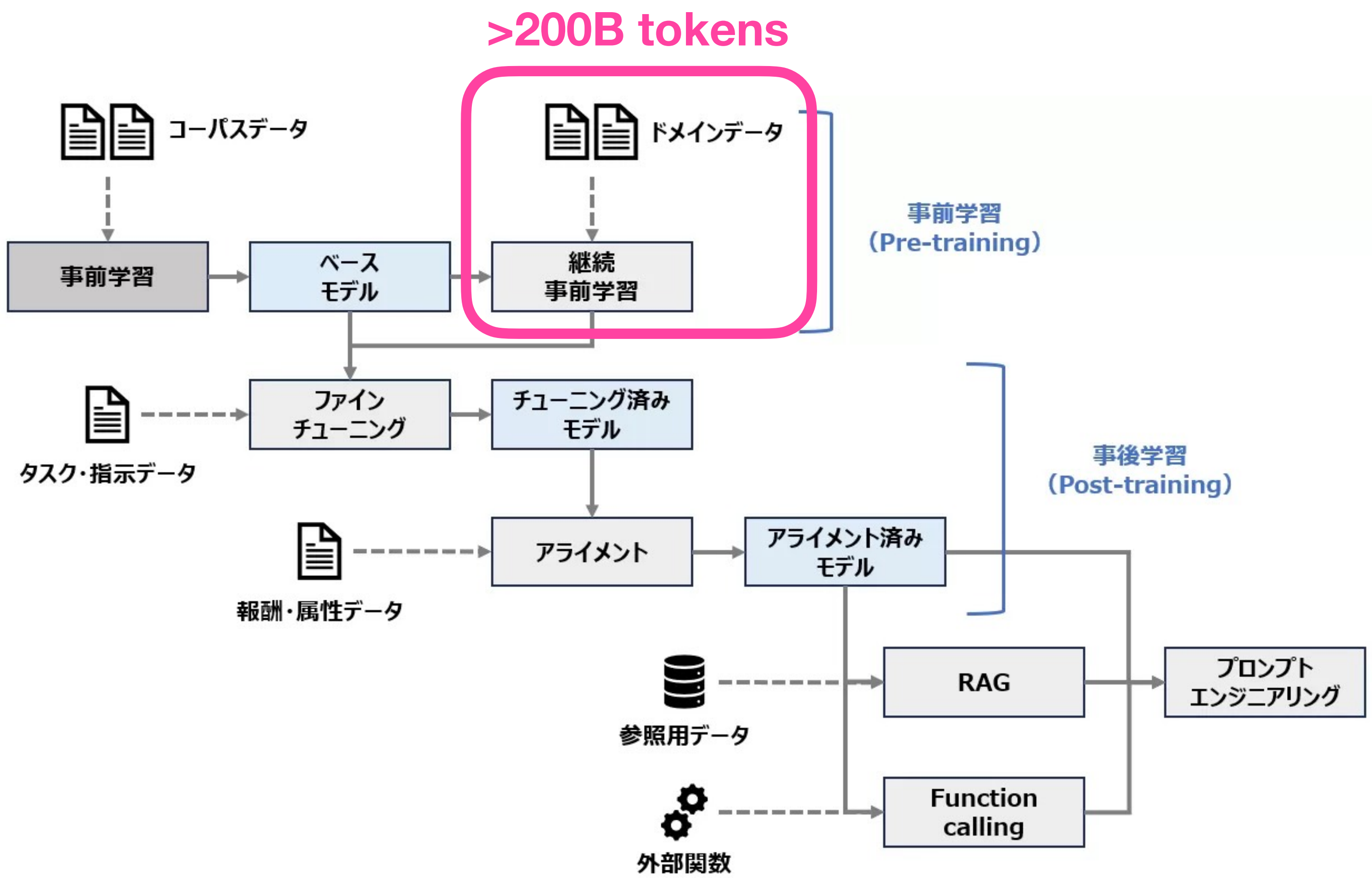
医療用日本語LLMの開発

FY2024
SIP第3期「統合型ヘルスケアシステムの構築における生成AIの活用」

医療用日本語LLMの開発

FY2024
SIP第3期「統合型ヘルスケアシステムの構築における生成AIの活用」

SAKURA internet　ELYZA　ABEJA　東京大学 THE UNIVERSITY OF TOKYO　理化学研究所　千葉工業大学 CHIBA INSTITUTE OF TECHNOLOGY　藤田医科大学 FUJITA HEALTH UNIVERSITY　国際医療福祉大学



医療用日本語LLMの開発

FY2024
SIP第3期「統合型ヘルスケアシステムの構築における生成AIの活用」

SAKURA internet ELYZA ABEJA 東京大学 THE UNIVERSITY OF TOKYO 理化学研究所 千葉工業大学 CHIBA INSTITUTE OF TECHNOLOGY 藤田医科大学 FUJITA HEALTH UNIVERSITY 国際医療福祉大学

2025年度の医師国家試験ベンチマークでのスコア比較

A. シンプルに問題を解かせた場合

モデル	分類	事業者	2022年得点（約500点満点）	2025年正答率
Weblab-Qwen-2.5-72B-Instruct（仮）	国内	ours	459	**93.3**
OpenAI-o1	国外	OpenAI	**461**	92.8
DeepSeek R1	国外	DeepSeek	439	91.5
Preferred-MedLM-Qwen-72B	国内	PFN	433	82.0
GPU-4o	国外	OpenAI	434	88.5
Qwen2.5-72B	国外	Alibaba	399	-
Llama3-Preferrerd-MedSwallow-70B	国内	PFN	395	-
GPT-4	国外	OpenAI	392	-

B. RAGやmajority votingなどの工夫を導入した場合（※）

※ ただし、図の参照を要する問題や計算を要する問題は、以下の実験からは除外

Weblab-Qwen-2.5-72B-Instruct（仮） ＋ RAG, majority voting	国内	提案者ら	-	**98%**

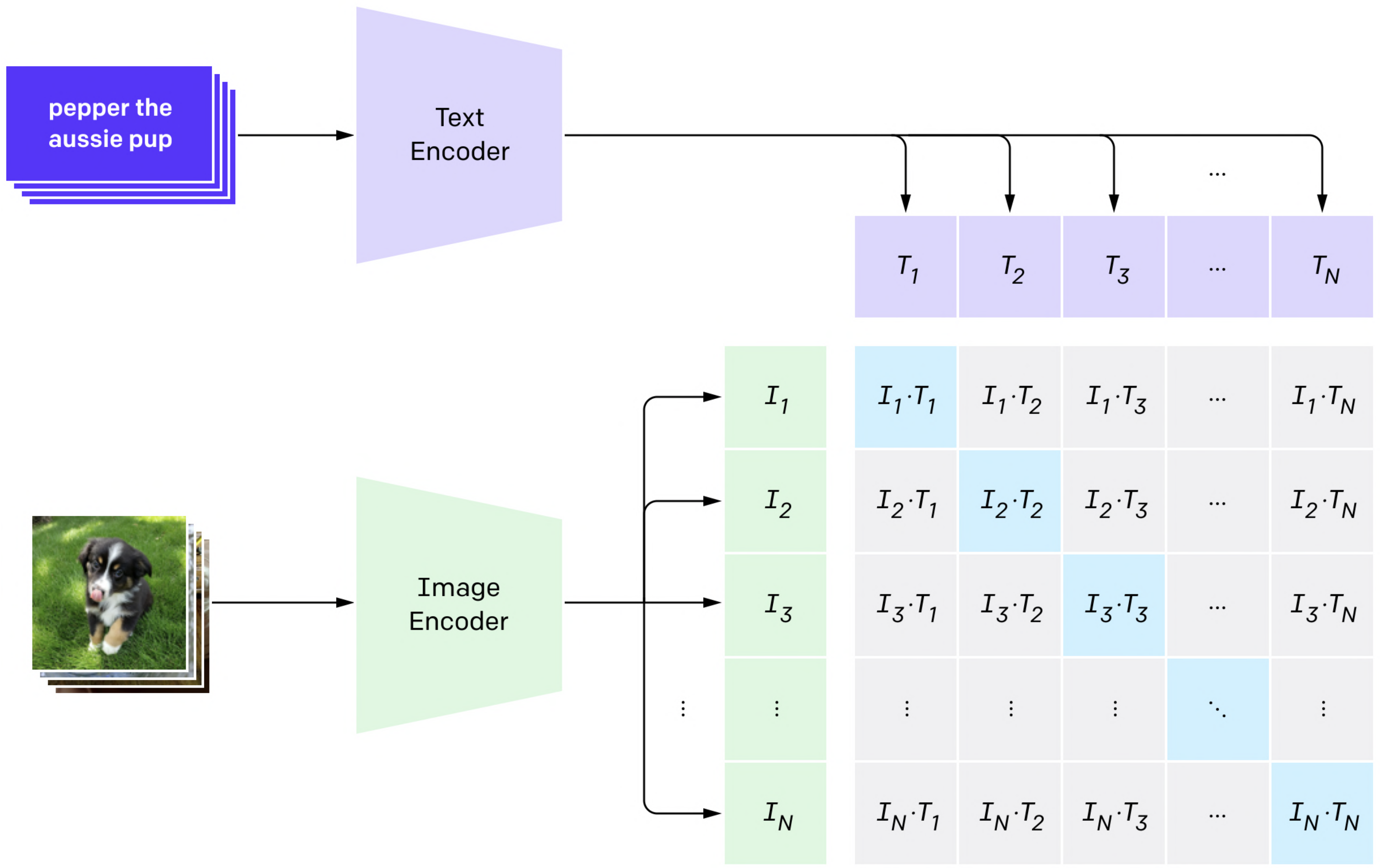
医療用日本語LLMの開発

FY2024
SIP第3期「統合型ヘルスケアシステムの構築における生成AIの活用」

↓

FY2025
NEDO「日本語版医療特化型LLMの社会実装に向けた安全性検証・実証」

CLIP (Contrastive Language–Image Pre-training)



https://openai.com/index/clip/

専門家に言語化させる

nature medicine

Article

https://doi.org/10.1038/s41591-023-02504-3

A visual–language foundation model for pathology image analysis using medical Twitter

Received: 26 March 2023

Accepted: 18 July 2023

Published online: 17 August 2023

Check for updates

Zhi Huang[1,2,4], Federico Bianchi[3,4], Mert Yuksekgonul[3], Thomas J. Montine[2] & James Zou[1,3]

The lack of annotated publicly available medical images is a major barrier for computational research and education innovations. At the same time, many de-identified images and much knowledge are shared by clinicians on public forums such as medical Twitter. Here we harness these crowd platforms to curate OpenPath, a large dataset of 208,414 pathology images paired with natural language descriptions. We demonstrate the value of this resource by developing pathology language–image pretraining (PLIP), a multimodal artificial intelligence with both image and text understanding, which is trained on OpenPath. PLIP achieves state-of-the-art performances for classifying new pathology images across four external datasets: for zero-shot classification, PLIP achieves F1 scores of 0.565–0.832 compared to F1 scores of 0.030–0.481 for previous contrastive language–image pretrained model. Training a simple supervised classifier on top of PLIP

専門家に言語化させる

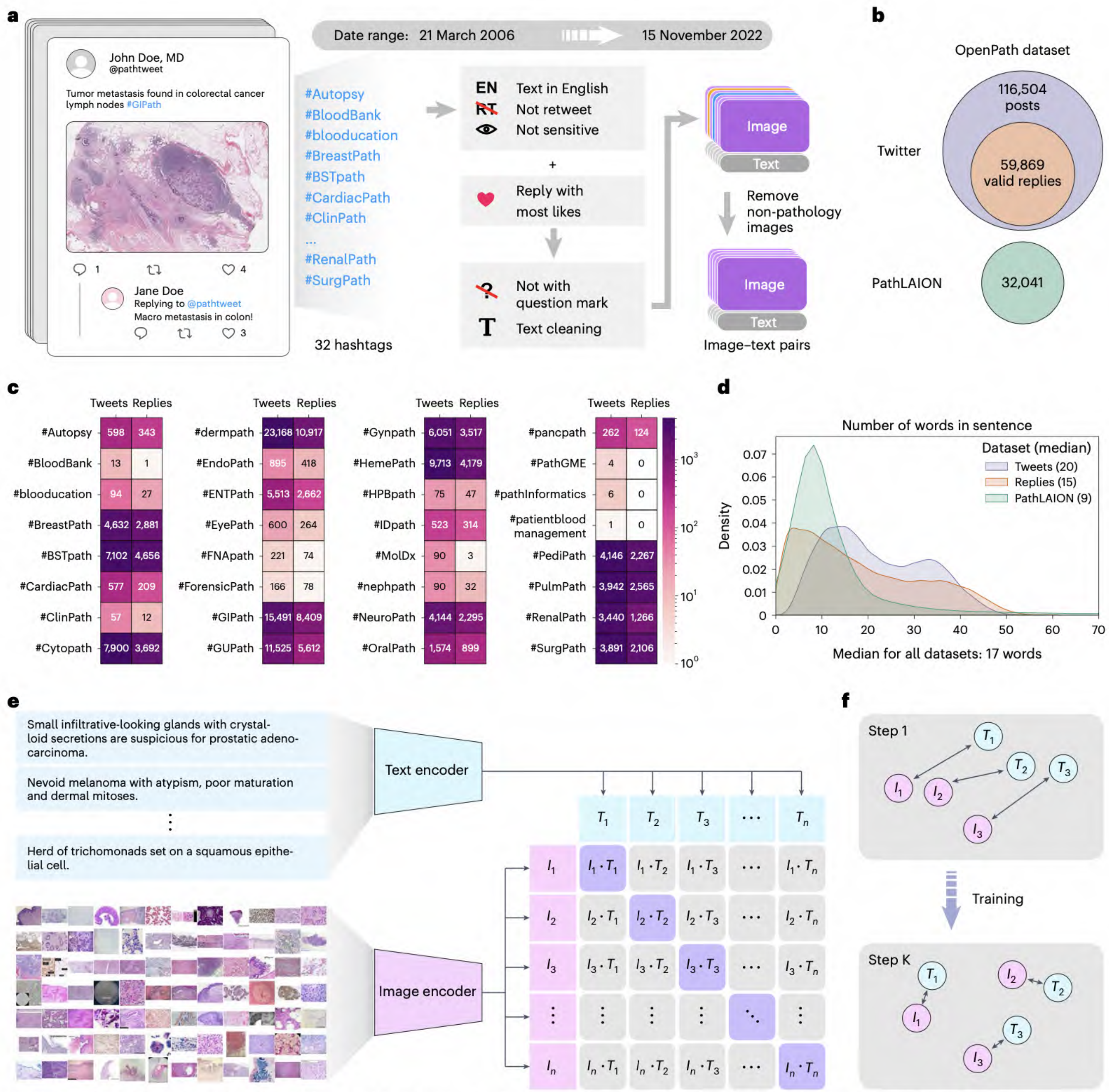
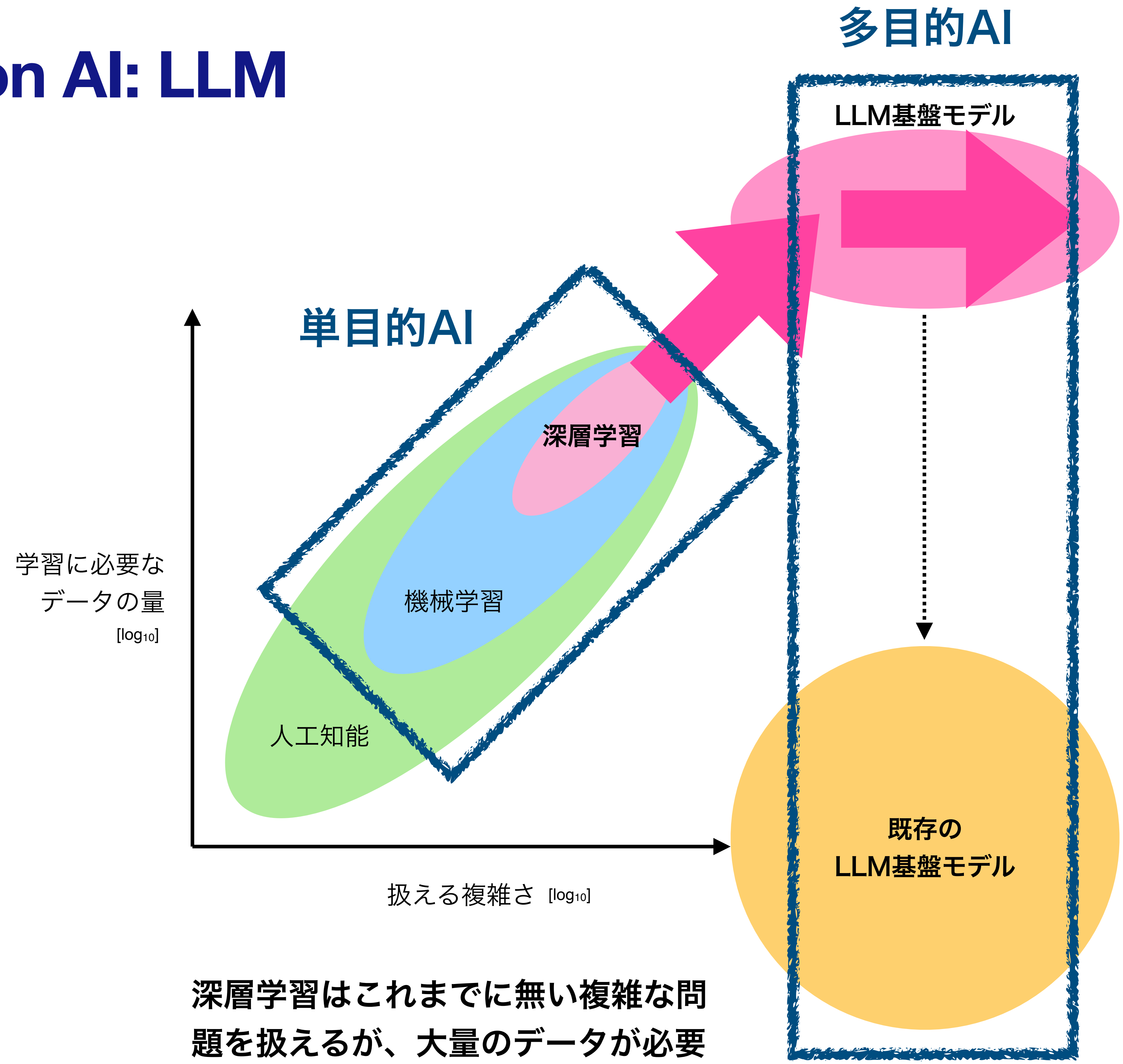


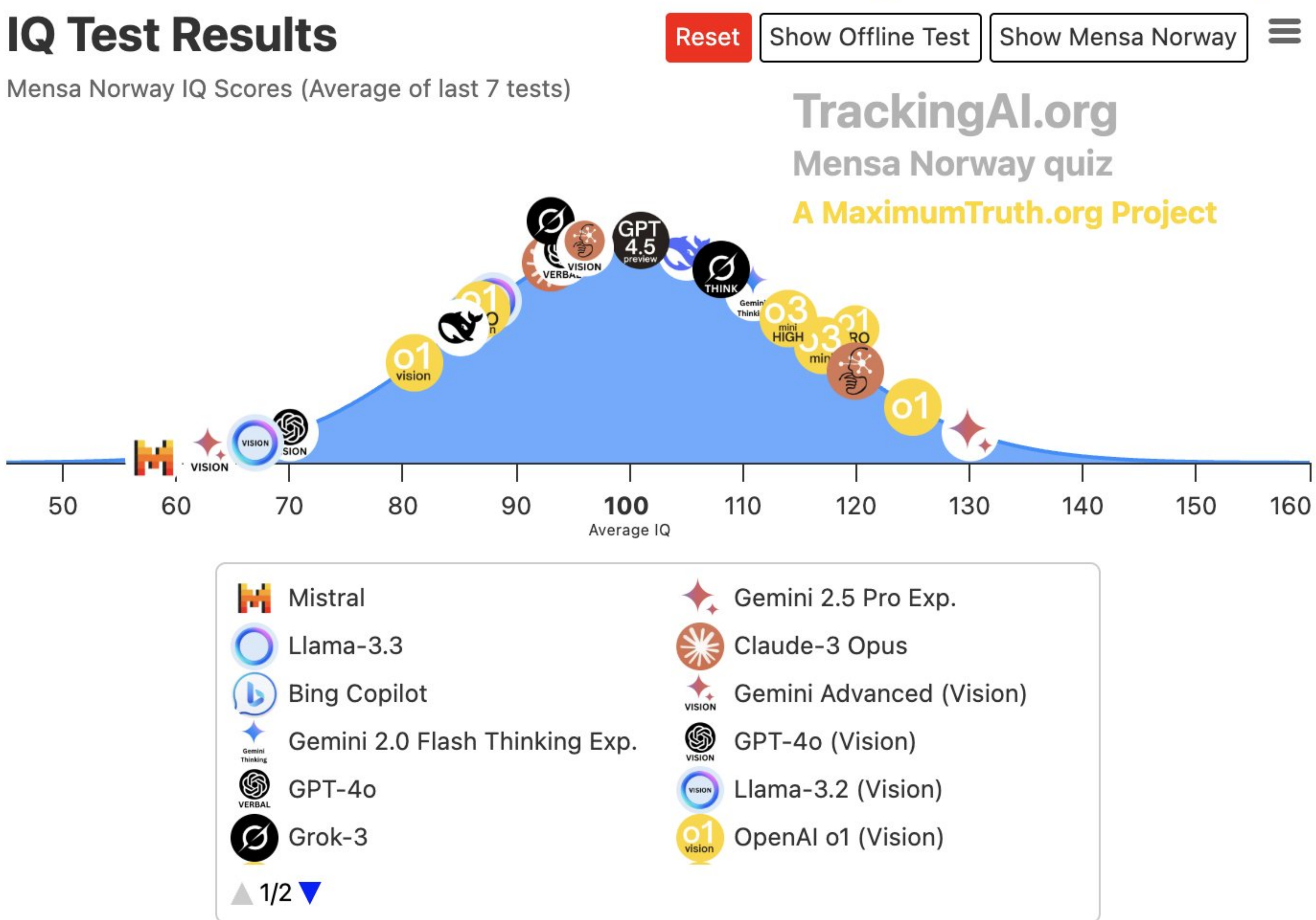
Fig. 1 | Overview of the study. a, Flowchart of data acquisition from medical Twitter. **b**, Overview of the OpenPath dataset. **c**, Total number of available image–text pairs from tweets and replies within each Twitter hashtag (sorted in alphabetical order). Replies are those that received the highest number of likes in Twitter posts, if applicable. **d**, Density plot of the number of words per sentence in the OpenPath dataset. **e**, The process of training the PLIP model with paired image–text dataset via contrastive learning. **f**, Graphical demonstration of the contrastive learning training process.

4th Generation AI: LLM



深層学習はこれまでに無い複雑な問題を扱えるが、大量のデータが必要になる。

Intelligence in Language Space



https://www.maximumtruth.org/p/massive-breakthrough-in-ai-intelligence

ようやく画像も基盤モデルに到達

Google DeepMind 2025-10-1

Video models are zero-shot learners and reasoners

Thaddäus Wiedemer[*1], Yuxuan Li[1], Paul Vicol[1], Shixiang Shane Gu[1], Nick Matarese[1], Kevin Swersky[1], Been Kim[1], Priyank Jaini[*1] and Robert Geirhos[*1]
[1]Google DeepMind

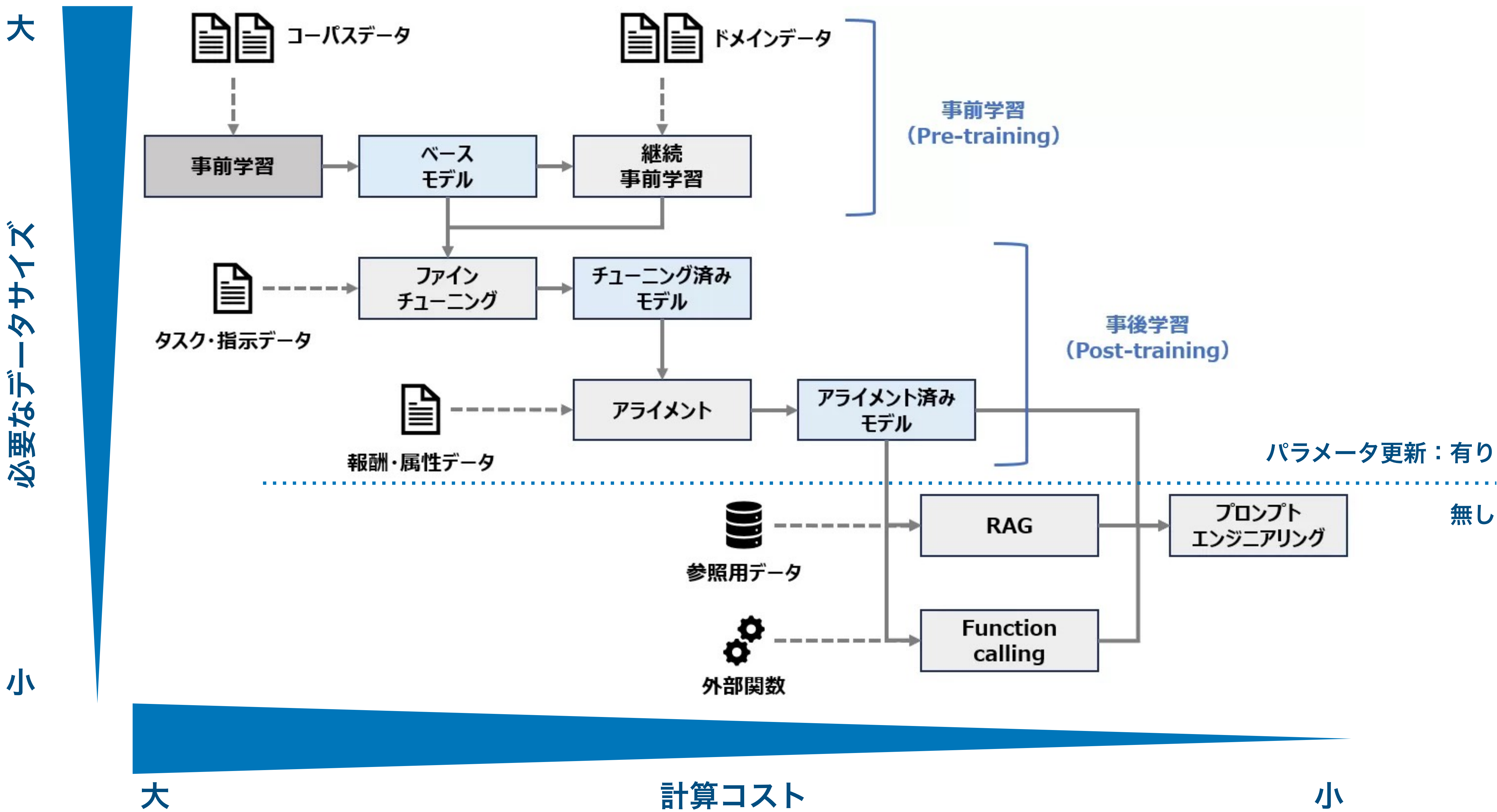
The remarkable zero-shot capabilities of Large Language Models (LLMs) have propelled natural language processing from task-specific models to unified, generalist foundation models. This transformation emerged from simple primitives: large, generative models trained on web-scale data. Curiously, the same primitives apply to today's generative video models. Could video models be on a trajectory towards general-purpose *vision* understanding, much like LLMs developed general-purpose *language* understanding? We demonstrate that Veo 3 can solve a broad variety of tasks it wasn't explicitly trained for: segmenting objects, detecting edges, editing images, understanding physical properties, recognizing object affordances, simulating tool use, and more. These abilities to perceive, model, and manipulate the visual world enable early forms of visual reasoning like maze and symmetry solving. Veo's emergent zero-shot capabilities indicate that video models are on a path to becoming unified, generalist vision foundation models.
Project page: https://video-zero-shot.github.io/

1. Introduction

We believe that video models will become unifying, general-purpose foundation models for machine vision just like large language models (LLMs) have become foundation models for natural language processing (NLP). Within the last few years, NLP underwent a radical transformation: from task-

https://arxiv.org/abs/2509.20328

LLMとデータ



https://dalab.jp/archives/journal/llm-finetuning-part1/

まとめ

- 深層学習を第3世代AIと第4世代AIに区別
- 「基盤モデル」を軽々しく使わない
- 第4世代AIを用いる場合
  - データだけでなくAIモデルを作る能力に投資必要
  - 自然言語以外のモーダルの言語化
  - 大規模モデル自体を作るなら
    - 圧倒的大規模なデータが必要
  - 既存大規模モデルの店子になるなら
    - 高品質で独自のデータが必要
  - データのオープン・クローズ戦略