Preferred Networks

2024/5/24 AI 人工知能EXPO 2024春

生成AIの進化と今後の展望

Preferred Networks 代表取締役 最高研究責任者
Preferred Computational Chemistry 代表取締役社長
Preferred Elements 代表取締役社長
岡野原 大輔
@hillbig

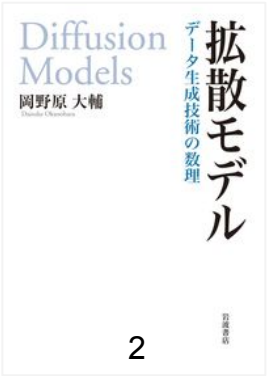
自己紹介：岡野原 大輔

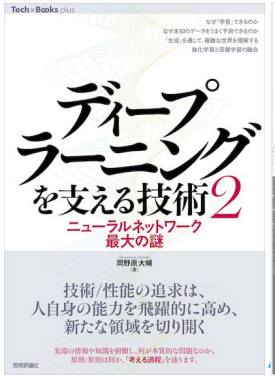
Preferred Networks 代表取締役 最高研究責任者
Preferred Computational Chemistry 代表取締役社長
Preferred Elements 代表取締役社長
Preferred Robotics 取締役　他

X (Twitter): @hillbig

AIに関する次世代リーダーとの車座対話メンバー
AI事業者ガイドライン 検討会委員

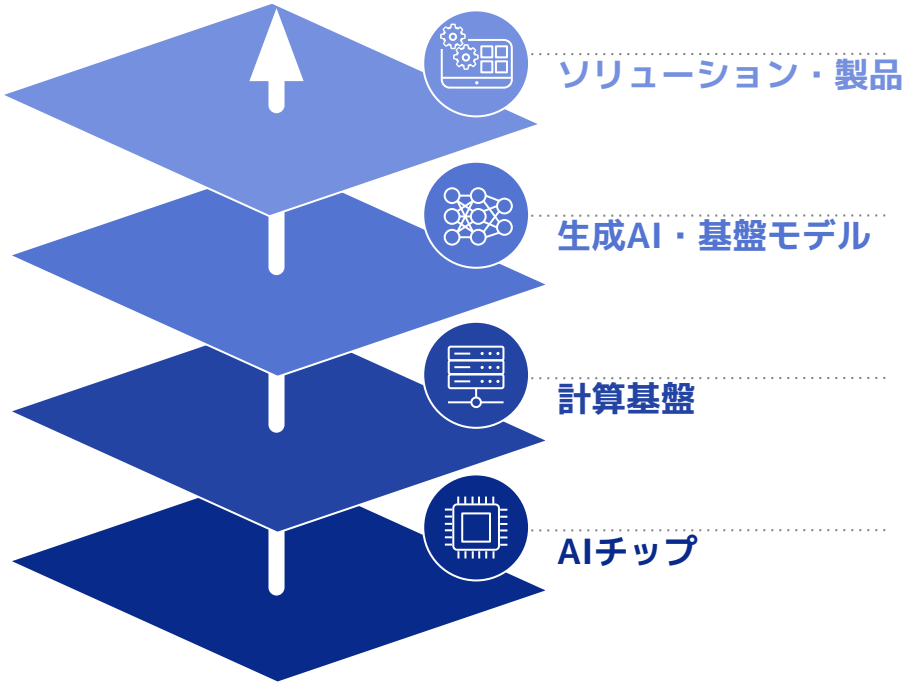
著書





PFNの事業: AI技術のバリューチェーンを垂直統合

Preferred Networks(PFN)は、チップ、計算基盤、生成AI・基盤モデル、ソリューション・製品までAI技術のバリューチェーンを垂直統合し、産業応用を進めている



様々な産業・消費者向けのソリューション・製品群

MATLANTIS / Preferred Networks Visual Inspection / PFN 3D Scan / rypko / kachaka / OPTUI

PLaMo 13B
大規模言語モデル

PLaMo
マルチモーダル基盤モデル
(2024年秋リリース予定)

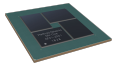
Preferred Potential (PFP)
物質の電子状態・
エネルギー計算モデル


GPUクラスタ


MN-3
(MN-Core™
クラスタ)

MN-4
(MN-Core™2
クラスタ


MN-Core™ 2による
計算能力のクラウド提供
(2024年開始予定)


MN-Core™

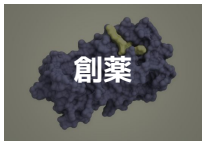

MN-Core™ 2

次世代機

Preferred Networks

PFNグループは生成AIとそれを支える計算資源を提供する

PFNグループは中長期的に社会基盤を担う大規模基盤モデル、それを支える計算力の提供に貢献していく



産業
生産性向上・品質改善
属人化回避・人手不足解消

社会
安心・安全な社会
高度な教育・医療

コンシューマー
人間の能力の拡張
新しい創作表現・娯楽体験

生成AI・基盤モデル

計算基盤

AIチップ

相互作用を引き起こし
イノベーションの連鎖へ

多くの仕事や社会生活が
AIによって支えられる

Preferred Networks

AIによる産業へのインパクト

AIは圧倒的に速く、大量の仕事をタフにこなすことが可能
例外なく多くの業界・職種に影響を及ぼすと予想されており、業界構造も激変する

AIができること

- 膨大な情報を瞬時に処理できる
  (例)数百万のAI同士が瞬時に複雑な条件下で交渉し契約を結ぶ
- 莫大な量のタスクをタフにこなせる
  (例)膨大な本や資料を分析、数万のレポートを数分で書上げる
  (例)24時間365日稼働可能。仕事を選ばない
- 高度な分析や判断、問題解決ができる
  (例) 大量のデータを元に改善
- ロボットと併用すれば実世界作業も自動化

AIがもたらすインパクト

- 2030年までに15.7兆ドルの経済効果[*1]
- 高い教育水準が必要な仕事にも影響
  (例) 営業・マーケティング・プログラマ・
  エンジニア・研究開発・士業・経営等
- 多くの仕事に影響を与える
  (例)・現在の50%の仕事が自動化される[*2]
  ・全世界で数億人が転職、業種を変える
- 人間の生活・常識・行動様式・
  価値観が変わりうる大きく変わる

*1 https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html

*2 https://www.zippia.com/advice/ai-job-loss-statistics/

AIによる知的労働の変化

AIは既に専門家並の知識を有している

評価対象	MMLU(*)正答率
[人] 一般人	34.50
[人] 専門家(分野毎)	89.80
[AI] OpenAI (GPT-4)	87.29
[AI] Google (Gemini Ultra)	90.04
[AI] Llama3 400B+	86.1
[AI]Phi-3 mini 3.8B	68.8

2023年時点では**クローズド**で**クラウド**でしか手に入らなかった能力が**オープン**または**エッジ**でも使えるようになる

(*) MMLU: Massive Multitask Language Understanding
AIの知識を評価するために設計されたベンチマークテスト。数学、生物学、歴史、計算機科学、法学など57の高度専門知識に関する総合力を評価。

MMLUの問題例

Find all c in $\mathbb{Z}_3$ such that $\mathbb{Z}_3[x]/(x^2+c)$ is a field.
(A) 0 **(B) 1** (C) 2 (D) 3

What is the embryological origin of the hyoid bone?
(A) The first pharyngeal arch
(B) The first and second pharyngeal arches
(C) The second pharyngeal arch
(D) The second and third pharyngeal arches

AIにより知的労働の制約は小さくなる

- 複数分野の専門家をツールとして利用可能
- これまで知的労働は教育された限られた人が行っていたがAIによる知的労働によって制約はなくなり生産総量は劇的に増える

人とAIはお互いの強みを活かし、協調する

- 既存の情報や知識を扱うのはAIが得意
- 人が新しい世界を開拓するのは不変
- 今後、人はAIを使いこなすスキルを身に着け、柔軟な発想力、想像力を活かしていく

人は新たな技術や知識を吸収し
柔軟に対応していく姿勢が求められる

Preferred Networks

Preferred Networks

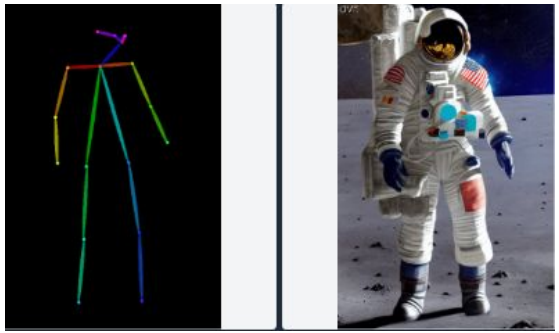
大規模言語モデルの仕組み

生成モデル

$x \sim p(X \mid C)$

X: 生成対象　C: 条件

ControlNet [Zhang+ 2023]



User: 2050年はどうなる？

System: 気候変動への対策が更に重要となります

- 生成モデル：対象ドメインのデータを生成できるようなモデル
  - テキスト、画像、動画、化合物、行動列 等
- **条件**を通じて、制約、指示、対象ドメインなどを指定する
  - 条件付した方が分布が単純になり学習や推論が簡単になる
- 大規模言語モデル、画像/音声生成、化合物生成など

Preferred Networks

生成能力の驚異的な進化速度

画像生成はわずか10年間で驚異的に進化した
他の分野（言語、化合物、制御など）の生成能力も今後急激に成長していく

(生成データ) (学習データ)

数字の生成

動物の画像生成
(2013年頃の生成モデル)

進化の要因

- 計算資源・学習データが**数百倍**に増加
- AI研究の進展
- 年間10万報の研究論文

群衆の画像生成 *
(技術：拡散モデル / 2023年)

10年前は、簡単な数字でさえ生成が難しく
複雑な構造を持った対象の生成は困難だった

わずか10年で非常に複雑なシーン（複数の人、光学現象、姿勢）を誰でも生成できるようになった

* Yuki Homma @y__homm
https://twitter.com/y__homm/status/1636186478899494912

大規模言語モデル ：AIが常識を理解して学習できるようになった

従来AIの学習方法（例：機械学習）

普遍的な事実・法則・常識

帰納的アプローチ

個別の事象　個別の事象

人の学習/勉強方法（例：教科書）

普遍的な事実・法則・常識

演繹的アプローチ

個別の事象　個別の事象

大規模言語モデルの学習方法

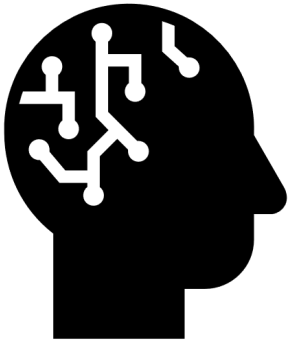
普遍的な事実・法則・常識
(学習対象の大量の言語データに含まれる)

演繹的

大規模言語モデル

帰納的

個別の事象　個別の事象



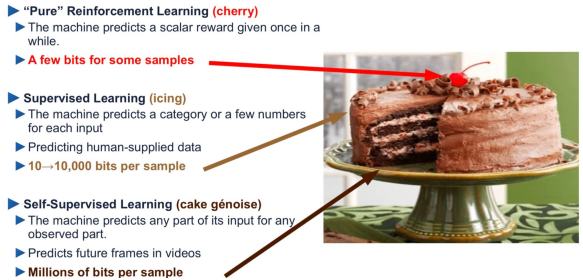
AIが現実世界、
そして「**常識**」を
理解し始めた

Preferred Networks

自己教師あり学習

正解がタダでいくらでも手に入る教師あり学習問題

- 予測：過去から未来を予測する
- 欠損補完：一部を欠損させ残りから欠損を予測
- 対比：意味が同じものと違うものを対比させる



[Lecun 2019]

そのタスクの達成自体が目標ではなく、そのタスクを達成するための副作用として様々な能力を獲得する

　教師あり学習と違って、膨大かつ多様なデータを利用でき、特定タスク向けでない理解ができるようになる

Preferred Networks

大規模言語モデル(LLM)の自己教師あり学習

- 次の単語をうまく予測するために文やその背後の情報を理解する必要
  - うまく予測できることと、理解することは多くの点で一致する

「こうしたことから、私は父と一緒に＊＊＊へ行き相談した」

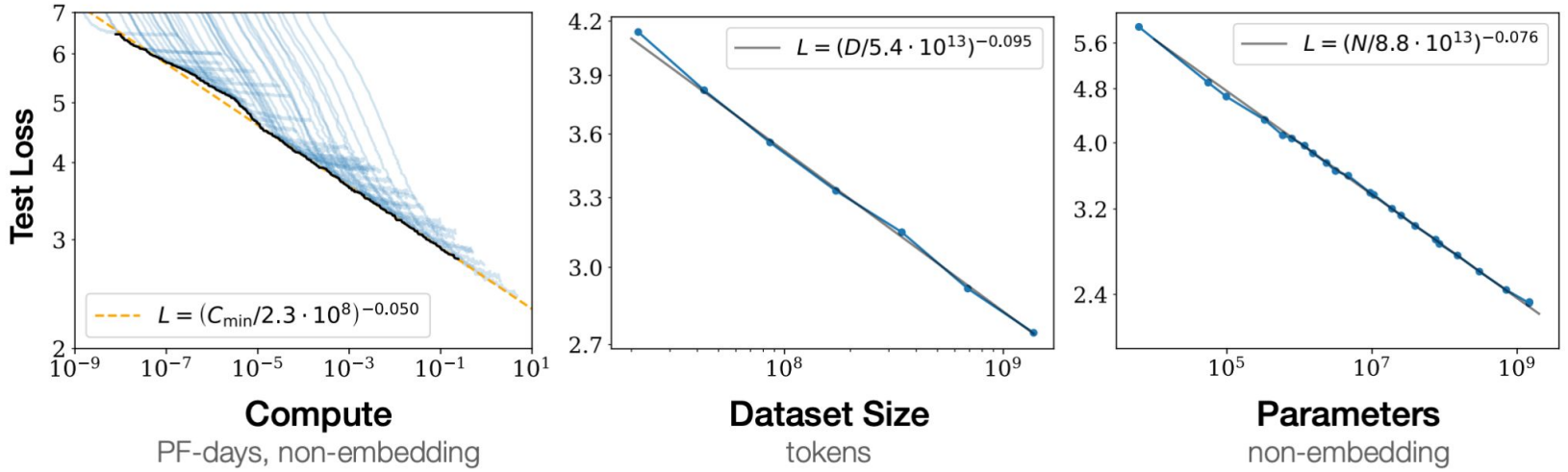
***に入る単語を予測するには？
(私や父はどういう人か、こうしたこととは何か
行って相談する場所はどこか　など)

言語モデルは次の単語を予測するために文章や概念を理解する必要に

駆られ、結果として言語やその背後の概念を理解している

Preferred Networks

言語モデルのべき乗則 [Kaplan+ 2020]



言語モデルでTransformerをモデルとして使った場合に
「学習時投入計算量」「学習データ量」「モデルサイズ」と、
「検証データのクロスエントロピー損失」との間にべき乗則が成り立つ

予測可能な形で予測性能や後続タスク性能が改善される

Preferred Networks

学習の大規模化による創発

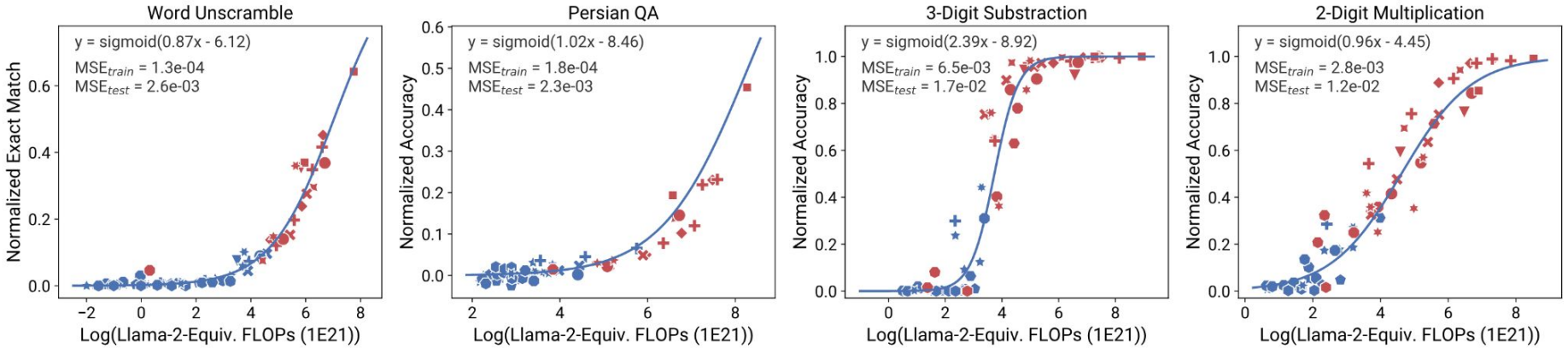
学習データ、モデルサイズ、学習時の投入計算量を大規模化すると様々な能力が創発する [Wei+ 2022][Ruan+ 2024]

創発：後続タスクの性能が急激に改善

　論理思考/質問応答/抽象的思考、本文中学習による事前知識の上書き能力

ツールを扱う能力

[Ruan+ 2024]

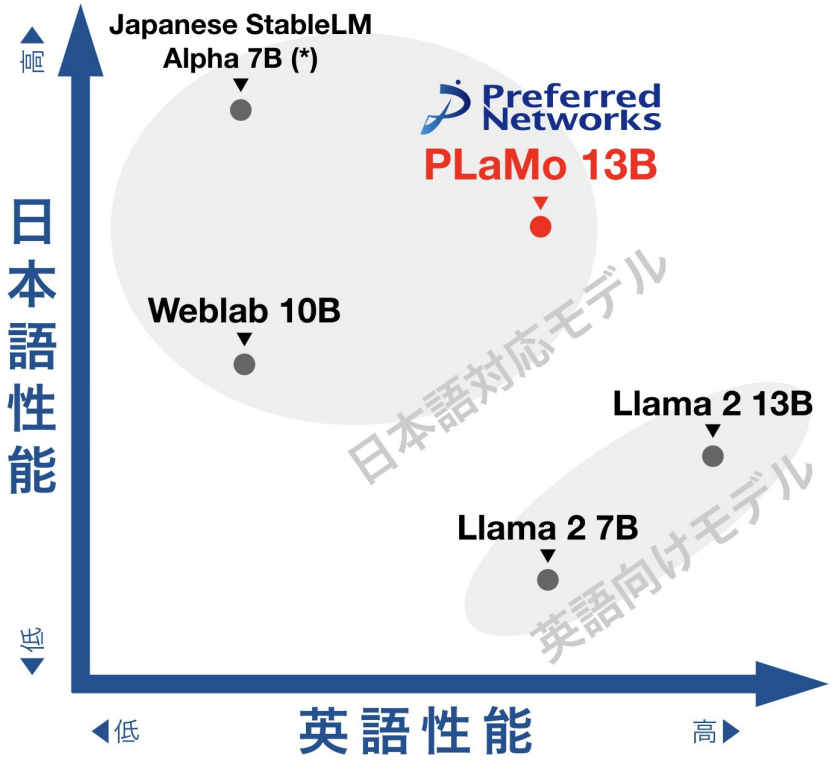


Preferred Networks

PFNの取り組み

PLaMo-13B PFNのマルチモーダル基盤モデル

PFNは研究/商用利用な13Bパラメータの大規模言語モデルを2023/9に公開



- 公開当時、日英2言語をあわせた能力で世界トップレベル
- 産総研のスーパーコンピュータABCIの A100 480GPUを1ヶ月弱利用。日本語、英語の1.4兆トークン（数兆文字）を使って学習
- 学習時の実行効率は41%で世界の他の学習基盤と比べても高く、効率的に学習資源を使える

- 開発実績をもとに基盤モデル開発・提供を行うPreferred Elementsを2023年11月に設立

Preferred Elements

Preferred Networks

100B/1Tパラメータからなる大規模マルチモーダル基盤モデルの構築

実施者	株式会社Preferred Elements
概要	本開発事業において2つのモデルの開発、検証を行う ・ 日本語性能に優れ、言語・画像・音声に対応したマルチモーダル100Bモデルの開発 ・ 1Tパラメータの言語モデルの事前学習の検証

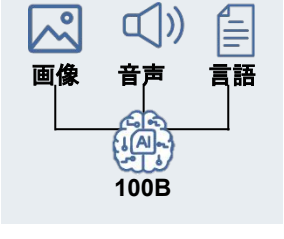
実施内容

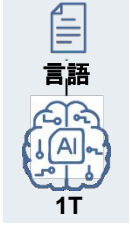
1. 100B モデルの事前学習・追加学習
2. 指示学習
3. 画像モーダル向けの事前学習・追加学習
4. 音声モーダル向け追加学習
5. 1Tモデルの事前学習の検証

開発・検証する基盤モデル

次の2モデルを開発

- 100Bのマルチモーダルモデル
  - 言語・画像・音声に対応
  - 一部タスクで世界最高レベルの性能
- 1Tの言語モデル
  - グローバルレベルでも大規模なモデルの学習を検証





社会実装の方法

自社ビジネスとしての展開

- 基盤モデルのAPI提供
- 基盤モデルのライセンス提供
- 基盤モデル及びAPIに付帯するビジネス（エンジニアリング･コンサルティング）
- 成果物の公開

ソースコード
（ファインチューニング用など）

モデル
事前学習済100Bモデルウェイト

開発ノウハウ
マルチモーダル化及び1Tモデル学習

GENIAC資料より

PLaMo-100Bの学習途中時点での性能

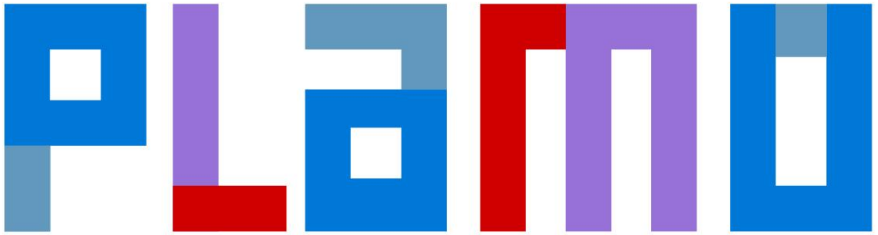
	Jaster 4shot
PLAMO-100B (5/13 暫定版)	**0.71**
以下参考値（GENIAC内での評価値）	
Swallow 70B	0.68
StableBeluga2 (70B)	0.67
GPT-4	0.77

Jaster
日本語LLMの性能を測るために標準的に使われるベンチマーク

1000億パラメータ
自社で構築した高品質学習データ 2Tトークンを使って学習

以上のPLAMO-100Bの結果は学習途中結果で、
今後残りの事前学習分と指示学習によって更に改善される見込み

PLaMo：PFNグループのマルチモーダル基盤モデル



- GENIACで開発中の100Bモデルを**2024年秋頃**にリリース予定
  - 日本語タスクや金融、医療など専門領域で高性能を目指す
  - マルチモーダルとして画像、音声に対応
- より軽量なモデル、特定タスクに強いモデルなども順次リリース予定
- 1Tパラメータ MoEモデルの検証・学習は**2024年夏頃**から開始

Preferred Networks

PreferredAI：PFNの生成AIを活用したプロダクト・サービス群

PreferredAI

powered by PLaMo

想定ユース毎にパッケージ化されたターンキー製品群を提供

- **PreferredAI Chat bot**
- **PreferredAI 対話シミュレータ**
- **PreferredAI Slide Review**
- **PreferredAI Notes**

など

これらを構成する技術要素を利用し、お客様の課題ニーズに合わせた独自ソリューションも提供します

Preferred Networks

生成AI事業

PFN/PFEは基盤モデル・パッケージ・ソリューションを提供し、事業パートナーと共に顧客の課題解決に向けて取り組んでいきたい

		ビジネス
個人での生成AI利用	個人向けの生成AIを使ったサービス・アプリの提供	SaaS
業務での生成AI利用	汎用SaaSやアプリを使いこなし業務で用いる	アプリ
インテグレーション	自社システム等に基盤モデルやアプリを連結する	API等
モデルのカスタマイズ	独自データなどを元にモデルを専用にカスタイマイズ	RAG 追加学習 最適化 事前学習
基盤モデル	基盤モデルやそれを使ったプロダクトを提供	モデル提供

どのレイヤーでも協業可能

例：

- 基盤モデルを社内で独自に大規模に使いたい
- 自社が持っている独自データを組み合わせて
- 専用モデルを作り事業化したい
- 他製品・アプリと組み合わせたい
- 機器に組み込みたい

Preferred Networks

LLMの利用事例：材料探索のデモ

人間の言葉による指示

Fe, Co, Ptについて、反応経路のindexが189(矢印で指した場所)の原子を置換した後に活性化エネルギーを計算してグラフにする

日本語の指示で必要なプログラムを生成

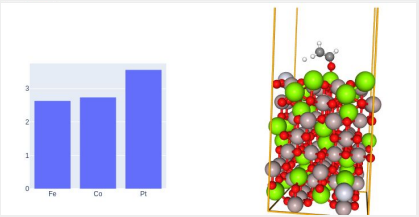
```
from typing import Dict, Optional, List, Literal
from ase import Atoms
import time
import ipywidgets as widgets
import plotly.graph_objects as go
from IPython.core.display import display
from IPython.display import clear_output
from atomsprog.utils.nglviewer import RawNGLViewer


class NGLManager:
                    def __init__(
                    self, title: str, width: str = "400px", height: str =
"400px", wrap: bool = True
                    ) -> None:
                    self.title = title
                    self.ngl_viewer = RawNGLViewer(width=width,
height=height)
                    self.wrap = wrap
                    self.recent_log = None
```

計算化学プログラムの専門知識をサポート

MATLANTIS

シミュレーションを実行

鉄が有望！



Preferred Networks

[8]:
```
interface = ProgInterface(ProgramGenerator(prompt="assets/prompt_minimum.txt"))
```

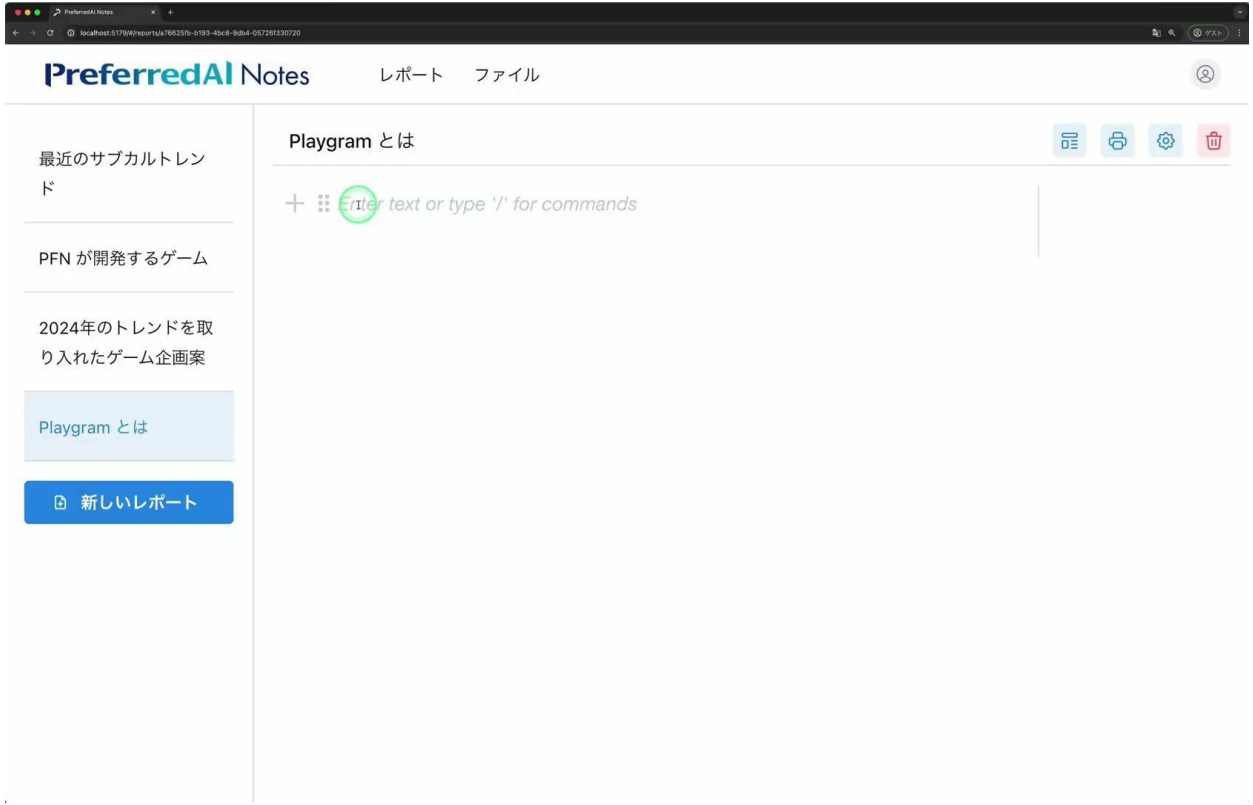
ここにテキストを入力してください

コード実行

[ ]:

PreferredAI Notes

社内外のデータを参照しノート作成



社内外のデータから数秒でノートを作成

PreferredAI Notesは、ビッグデータ活用を促進する**ストレージ統合型のAIドキュメントプラットフォーム。業務データ**をアップロードするだけで利用可能。要望を**伝えるだけ**で、蓄積された**あらゆるデータ**を元に検索も活用してノートやレポートを簡単生成。

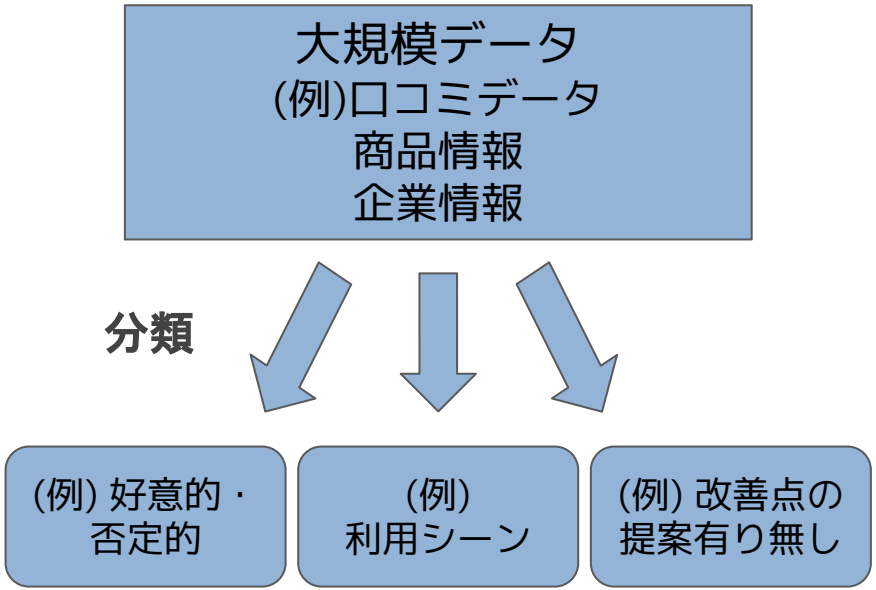
データが自然と集まるプラットフォーム

作成されたレポートも保存され、**データの好循環**が生まれる。時間のかかる文書作成を**圧倒的に効率化**でき、**データが自然と集まり、全社的な生産性アップ**に繋がる。

Preferred Networks

PLaMo Preference API

LLMを使った大量の分類タスクを簡単かつ高速に実行



LLMを使えば学習データ不要でに様々な**分類**が可能。しかし、従来LLMでは、分類**コストが大きすぎる**

- 独自技術でLLMの分類を**10~100倍高速化**
- コストが**安く**、大量のデータでも**速く**、学習不要で**柔軟な分類**の試行錯誤が可能！

Preferred Networks

開発中マルチモーダル基盤モデルによる画像認識例

入力画像

この画像は何が写っていますか

画像キャプション

ブロッコリと麺が入ったボウルです

A bowl of food with broccoli and noodles.

このボウルの食べ物を摂取することで
どのような良いことがありますか

VQA

この食べ物を摂取することで様々な良い点があります。麺から炭水化物を摂取でき、ブロッコリからはビタミンやミネラル、食物繊維を摂取できます。

The bowl contains a variety of food items, including noodles, broccoli, and meat. Consuming this meal can provide several health benefits. The noodles are a source of carbohydrates, which are an essential part of a balanced diet. Broccoli is a nutrient-rich vegetable that is high in vitamins, minerals, and fiber, promoting good digestion and overall health. The meat in the dish adds protein, which is crucial for muscle growth and repair. Overall, this meal offers a combination of essential nutrients, vitamins, and minerals that can contribute to a balanced and healthy diet.

野菜を検出

物体認識

日本語のゼロショットでの物体検出も実現

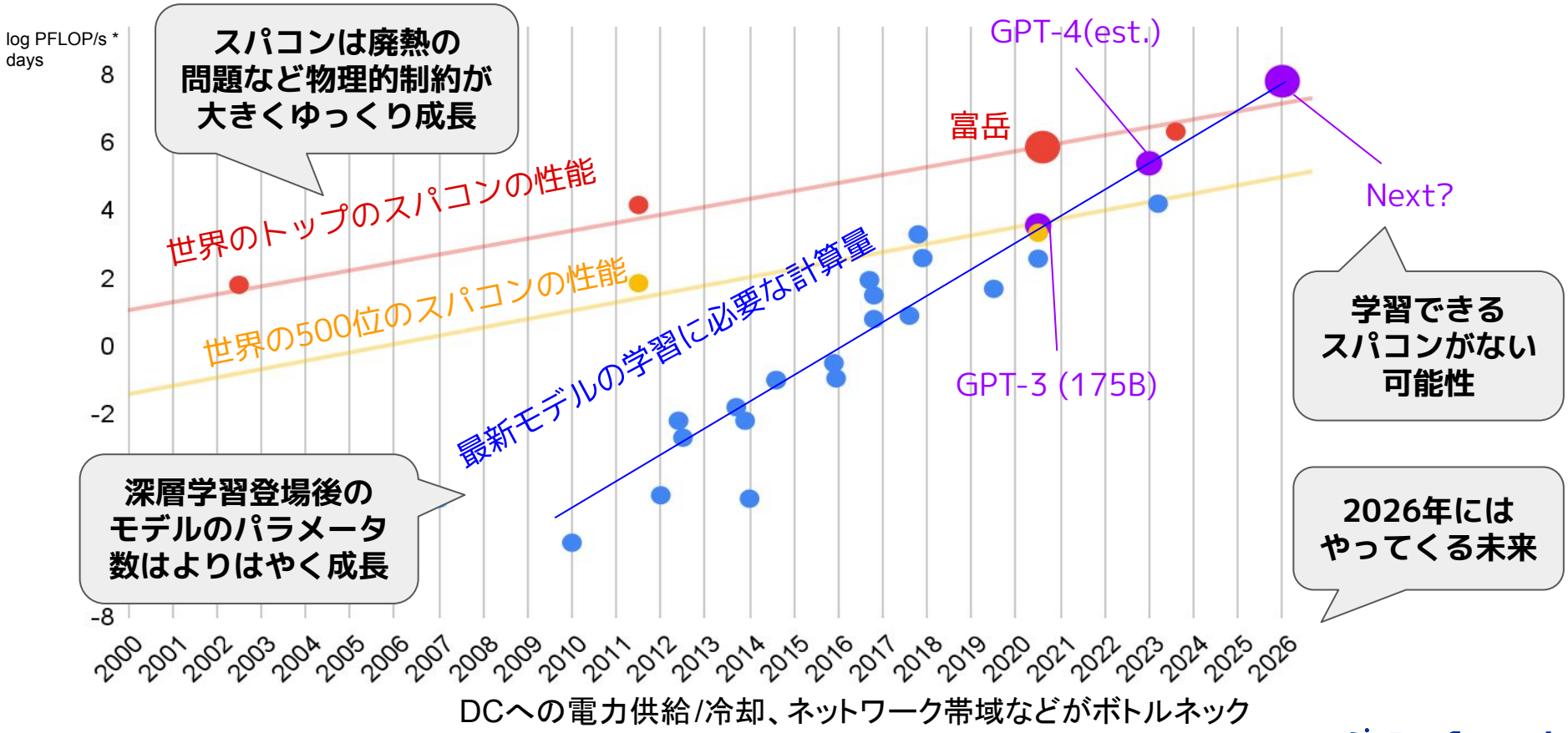
備考: prompt/結果を見やすくするため簡略化＆日本語化

Preferred Networks

Preferred Networks

LLM開発の今後

AIの最新モデルの学習に必要な計算資源の増加



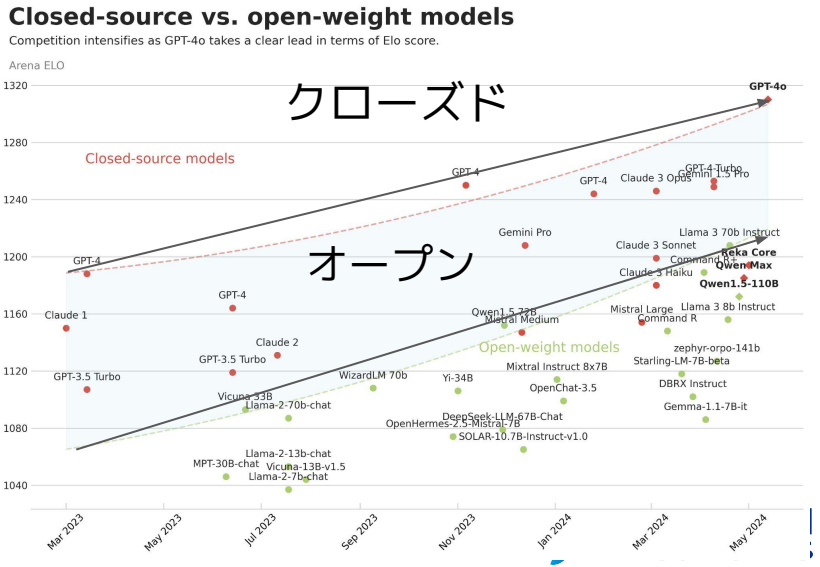
モデルの学習期間を半年とおいている

Preferred Networks

この1年のLLM開発の流れ

- “GPT-4レベル”に多くのLLMが追いつこうとしてきた
  - Gemini, Claude Opus, Llama3 400B+ 他
- 同じ能力を持つ推論に必要な計算資源は劇的に改善され続けてきた
  - あと1年でGPT-4レベルがスマホで動くようになってもおかしくない
- クローズドなモデルにオープンなモデルに追いつこうとしている

大きな非連続的変化でなく小さな改善が
積み重なることで淡々と賢くなり続ける



LMSys Chatbot Arena Leaderboardの比較
https://twitter.com/maximelabonne/status/1781265506772590855

スケール則による性能改善は続くのか？

- GPT-4レベルまでは可能だが次の規模は未知領域
  - チップあたり性能の改善とチップ数を増やし投入計算量自体は増加
  - 投資額が大きいためコスト面で他アクセラレータ利用も急速に進む

- スケール則の傾き改善は急速に進んでいる
  - 高品質データの利用、MoEにより投入計算あたり性能は劇的に改善
  - GPT-4並の性能が1/10のモデルサイズ、投入計算量で実現可能に

- 事後学習による改善も著しい
  - 新しい学習目標による改良が今後進む

Preferred Networks

学習データ改善の伸びしろが大きい：LLaMa3 Phi3 他モデル

- 収集データをLLMが学習しやすいようにLLMで加工することが重要
  - 強いLLMから小さいLLMに蒸留するだけでなく、
    「生成よりも識別の方が簡単」なためデータをフィルタリング、
    ラベリングし自分より強力なLLMの学習を助けることができる
- LLMで学習のためのデータを生成することは進む
  - 人が作った数百倍規模の「教科書」「Wikipedia」が作られていく
- 学習自体よりも学習データ作成の方に研究や計算投入が今後進む
  - データ素材、加工・選別技術、ユーザーフィードバックの反映
- 同じモデルサイズでも性能はまだ向上し続ける

Preferred Networks

知識の索引としてのLLM [Allen-Zhu 2024]

- LLMは様々な知識を後続タスクで自在に使えるような**索引**とみなせる
  - パラメータあたり2bitの情報（知識）を覚えられる
  - 例えば8bit量子化の場合、元の情報の4倍のオーバーヘッドで索引できる

- 一方、この記憶効率で覚えるには100回以上（理想的には1000回）その知識を様々な形で経験させる必要があり、学習には非常に時間がかかる
  - 同じ文章だと丸暗記するので言い換える必要がある
  - 例えば「PFNの本社は大手町です」という知識を覚えるにはこれを100回言い換えたテキストを学習しないとこの記憶効率は達成できない

Preferred Networks

プライベートデータの取り込み

どのように個人、組織、企業のプライベートデータをLLMに取り込むか？

- **ファインチューニング**
  - LoRA以外にも非常に多くの有望手法も登場している
  - モデルマージも登場。複数の学習結果をあわせることが可能である
  - 一方、新しく覚えるとハルシネーションが増加するとの報告も
- **RAG**
  - RAG自体の精度改善
  - NNS高速化に伴い学習時からRAGを使うことも現実的になってきた
- **長コンテキストによるIn-Context Learning**
  - 1千万トークン規模をコンテキストで扱うことが可能となっている
  - 説明書、指示書、制約などをその場で読み込むことも可能

まとめ

- LLMの進化は続く
  - 開発競争は単純なスケール則だけではなく、日本企業も勝機はでてくる
- LLMがより強力かつ使いやすくなる中でどのように使うのかが重要
  - 現在の能力や今後の能力進化に合わせて、使えるようになった分野で使っていく。解像度をあげてみていく必要がある

PFNではLLMモデル (PLaMo)の提供および、それを使ったアプリケーション・サービスの提供を通じ、生成AIの社会普及に取り組んでいきます