Preferred Networks

2024/3/15 NLP2024 生成AI時代の自然言語処理における産学官の役割と課題

大規模言語モデル開発の展望と今後の課題

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

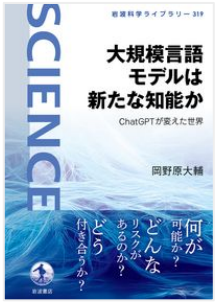
自己紹介：岡野原　大輔

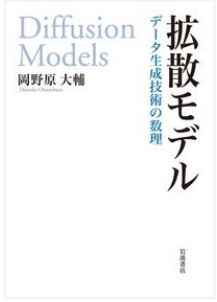
Preferred Networks 代表取締役 最高研究責任者 / 共同創業者
Preferred Computational Chemistry 代表取締役社長
Preferred Elements 代表取締役社長
Preferred Robotics 取締役　他

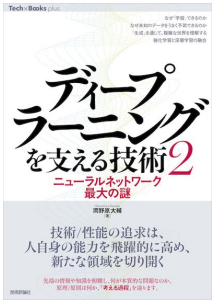
X(twitter): @hillbig

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



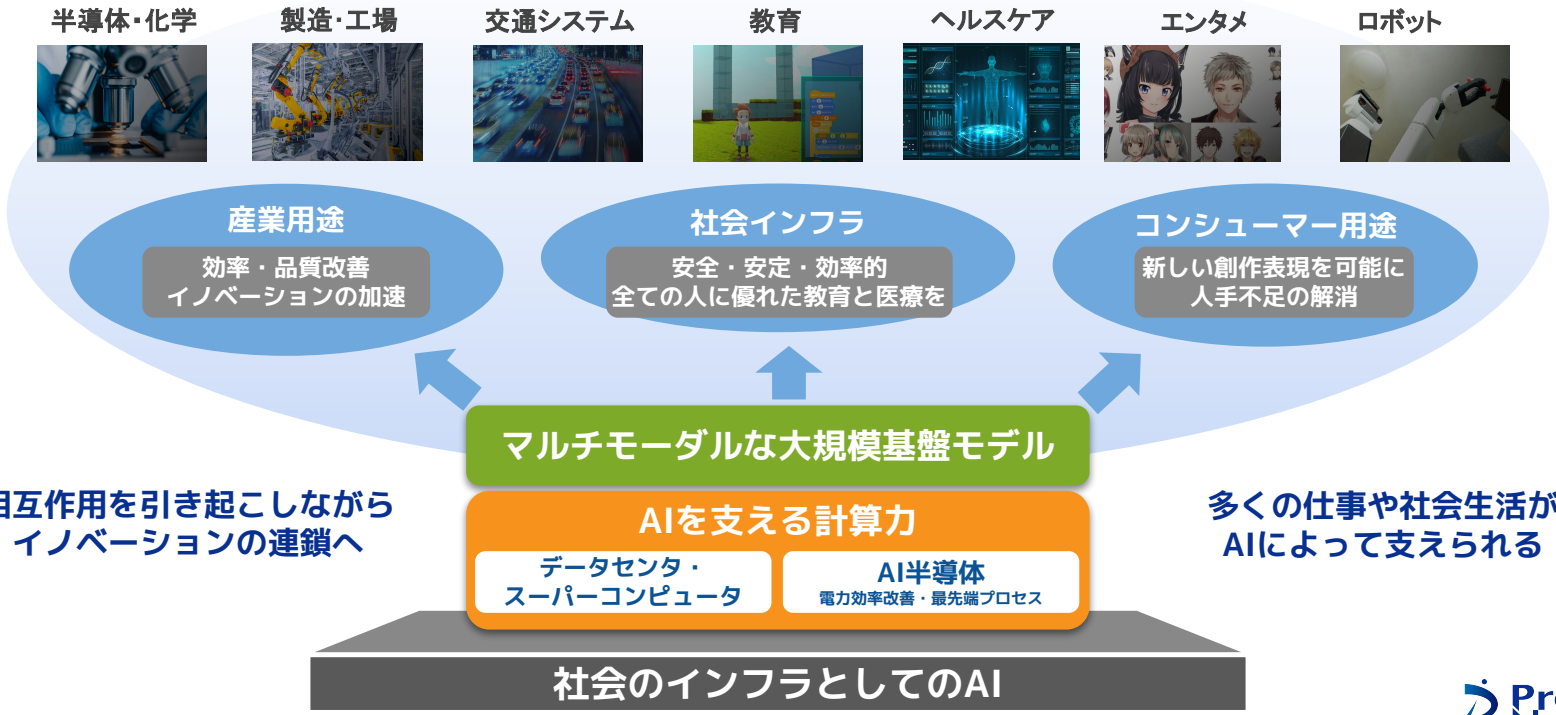




Preferred Networks

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

Preferred Networks（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



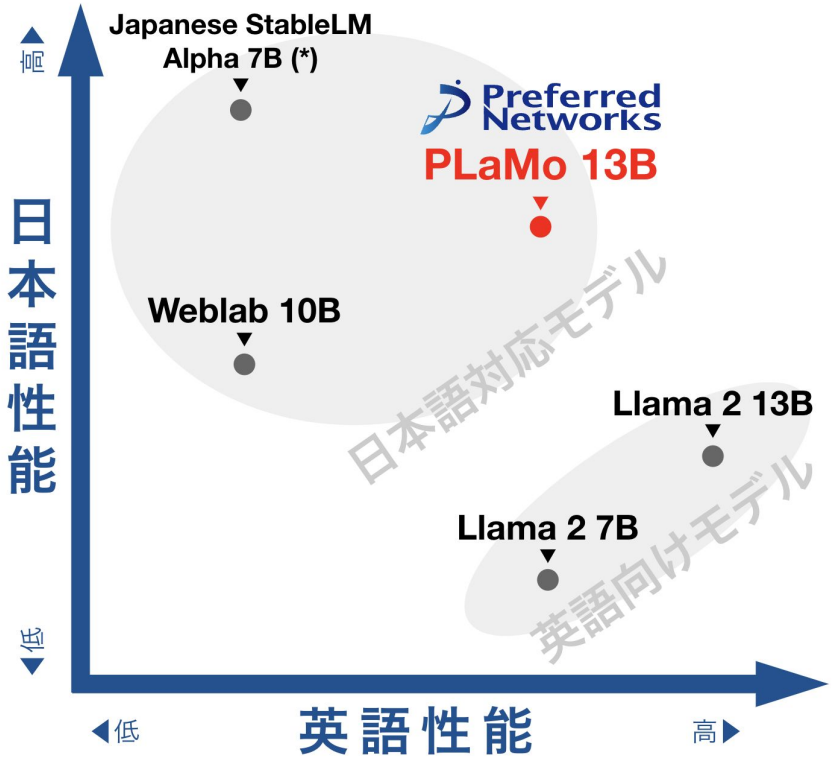
相互作用を引き起こしながら
イノベーションの連鎖へ

多くの仕事や社会生活が
AIによって支えられる

Preferred Networks

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
- 経産省ABCIのスーパーコンピュータ A100 480GPUで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

- 開発実績を元に基盤モデル開発・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 社会実装に向けたビジネス展開に加え、一部モデル・ノウハウ等の成果物も公開・提供 ・ 本開発事業で得られた成果物(モデル・開発ノウハウ)の一部を公開 ・ 自社ビジネスとして、国内外でのAPI・ライセンスの提供、付帯するビジネスを実施

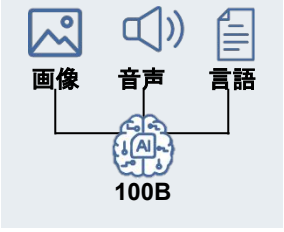
実施内容

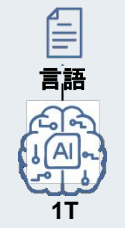
1. 100B モデルの事前学習・追加学習
2. 指示学習
3. 画像モーダル向けの事前学習・追加学習
4. 音声モーダル向け追加学習
5. 1Tモデルの事前学習の検証

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

今後の大規模基盤モデルの開発・リリース計画

- **100Bパラメータモデル**のマルチモーダル基盤モデルを**2024年秋頃**に投入予定
  - **言語・画像・音声**をサポート
  - 日本語性能や特定分野向けタスクでの最高性能を目指す
  - 安全性を検証した上で事前学習済みモデルや開発ノウハウは公開予定
- **1Tパラメータモデル**の基盤モデルの検証・学習を**2024年夏頃**から開始予定
  - MoEを利用予定。様々な研究・検証を実施予定
  - 得られた知見を元に自社リソースで大規模モデルの開発を進める
- また、データ生成、ファインチューニング、追加学習＋RAG、アプリケーションなどの知見をためている

- これらのモデルや知見を元に商用向けモデルを開発していく

Preferred Networks

LLMを使った材料探索のデモ

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

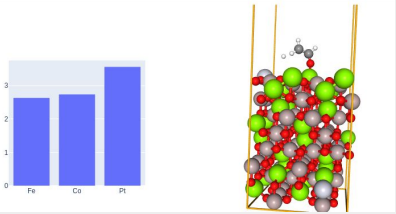
```
from typing import Dict, Optional, List, Literal
from ase import Atoms
import time
import ipywidgets as widgets
import plotly.graph_objects as go
from IPython.core.display import display
from IPython.display import clear_output
from atomsprog.utils.nglviewer import RawNGLViewer


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

[ ]:

Preferred Networks

PFNのマルチモーダル基盤モデルによる画像認識例 1)

入力画像

based on provided image, tell me what is happening in it

画像キャプション

A bowl of food with broccoli and noodles.

what benefits can one get from the food in the bowl?

VQA

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.

detect vegetables

物体認識

物体検出も統合されている日本語のゼロショットでの検出

備考: promptをみやすくするため簡略化

Preferred Networks

PFNのマルチモーダル基盤モデルによる画像認識例 2)

入力画像

what is going on in the image?

画像キャプション

A man riding a skateboard up the side of a ramp.

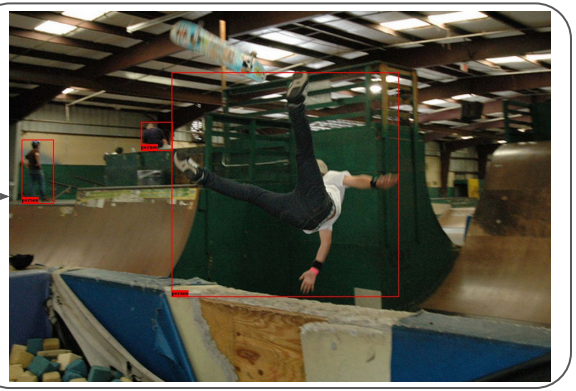
what went wrong with the skateboard trick and what could be the aftermath?

VQA

In the image, the skateboarder is attempting to perform a trick, but instead of successfully executing it, he loses his balance and falls off the skateboard. This could lead to various consequences, such as injuries, damage to the skateboard, or even embarrassment. The skateboarder might suffer from bruises, sprains, or even fractures, depending on the severity of the fall. Additionally, the skateboard itself could be damaged, requiring repairs or replacement. The skateboarder might also experience a loss of confidence or self-esteem, as they may have failed to perform the trick properly. This highlights the importance of practicing and mastering skateboarding tricks to avoid such mishaps and ensure safety while enjoying the sport.

detect people in the image

物体認識



備考: promptをみやすくするため多少簡略化

Preferred Networks

Preferred Networks

LLM開発における課題

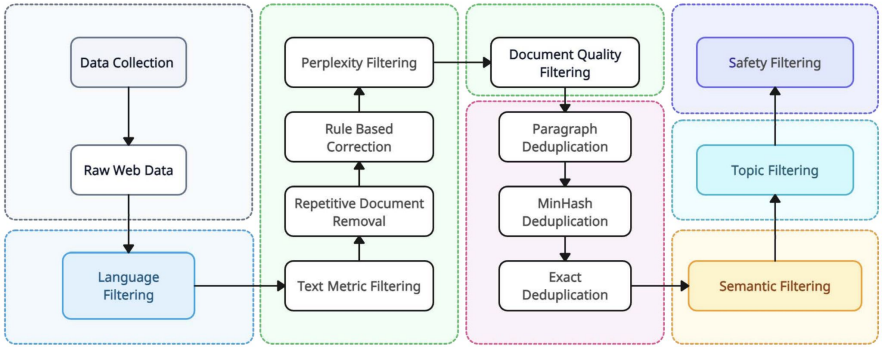
LLM開発における課題

- 学習データ
- アーキテクチャ
- Long Context
- 推論
- マルチモーダル

  （この他にもありますが時間の関係で話題を絞りました）

学習データ

- **どのように学習データを整備するか**
  - データや施策の良し悪しを後続タスク性能で評価するのは簡単でない
    - 良かれと思って入れた施策が悪影響を及ぼす場合も多い
    - 規模が大きくなると変わる場合も
  - Perplexity Filteringが有効だがなぜ良いかよく分かっていない



https://arxiv.org/abs/2403.04652
例：Yi のデータ処理パイプライン

- **どのように学習データを集めるか**
  - 良質でオープンな日本語データは限られている
  - 教材データ・書籍データなどは使えるか
    - 商用LLMでも、これらはかなり使われていると分析している
- **用途別データも足りない**
  - 例えば社内利用の場合は、社内データでFineTuningやRAGをするがその場合の標準的な学習データ・検証データがない

Preferred Networks

LLMを利用して学習データをどのように作るか

LLMを利用して学習データを整備する

- 学習データ自体を作る
  - 例：phiシリーズ　教科書（問題集が近い）を作る
    - プログラミング、数学などの分野で成功している　他も有望と思われる
- 蒸留
  - 強力な小さなモデルを作るために広く使われている

LLMを利用してデータをフィルタリング、タグ付けする

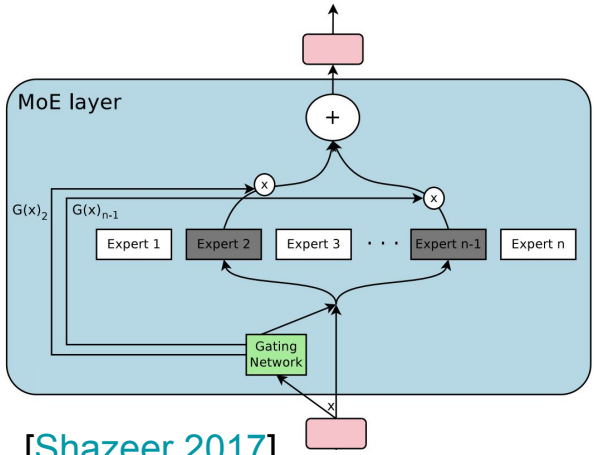
- 生成と違って評価は少ないコストで実行可能（1トークンだけ予測）
- RLHF, DPOなどにおいても利用できる
- あえて悪質なデータを学習データに含めた上で、悪いとタグ付けした上で生成の時にはそれを外して生成する（例：PaLMなど）など様々な工夫がありうる

学習データ作成の方法論、LLMを利用したデータ整備についてはかなりの伸びしろがある
一方、小規模実験、再現性などで研究として難しい部分が多々ある

Preferred Networks

アーキテクチャ 1 MoE

- MoE (Mixture of Expert） [Shazeer 2017] (GPT-4, Mixtral, Gemini-1.5等）
  - MLPブロックを複数のExpertに分割し、学習可能なルーターが選択する
  - 学習や推論コストを抑えつつパラメータ数を増やすことができ、同じ性能を達成するため必要な学習資源や推論時コストが数十分の一に
- MoEでも、べき乗則が成り立つ
  - 大きくなるにつれ、各Expertのパラメータ数を減らす必要がある [Karajewski 2024]
  - チンチラ則よりも最適なパラメータ数の増加割合は速い
- エキスパート側がトークンを選択する
  Expert Choice [Zhou 2022]も広く使われ始めている

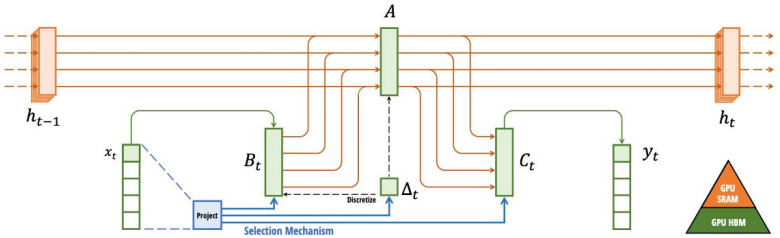


[Shazeer 2017]

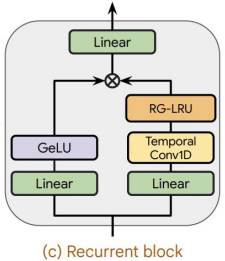
Preferred Networks

アーキテクチャ 2 状態空間モデル：注意機構の置き換え

- 推論時コストはkvキャッシュの読み込みが支配的（特に小さなモデルの場合）
  - kvキャッシュ：文脈中のkey/value列
- 文脈を定数サイズに圧縮するためにはRNN、状態空間モデルが有効
  - 改善は3~5倍。大きいモデルはMLPブロックの影響が大きく効果は限定的
- Mamba [Gu 2023]
  - 線形状態空間モデルではじめてTransfomerに匹敵する性能を達成
  - 入力に依存した重みを使えるようにし、Selective CopyなどICL発動に必要な能力を実現。ただ、ICL特性はTransformerと異なる [Grazzi 2024]
- Griffin [De 2024]
  - 回帰結合モデル(RG-LRU)とTransformerを組み合わせたモデル



Mambaの図



RG-LRUの図

Preferred Networks

Long Context

- 最近では扱える文脈長が数百万トークンになってきた (Gemini1.5, list)
  - ソースコード全体や説明書全体を文脈に入れて使うといったことが可能
- 実現例：
  - 状態空間モデル（前項）
    - 文脈情報を定数サイズの状態にオンラインで圧縮
  - Ring Attention [Liu 2023]
    - kv-cacheがデバイス上で順に転送させる
  - 線形注意機構 BASED [Arora 2024]
    - softmax注意機構を線形近似し長距離は線形、近距離は非線形で扱う
  - RoPEの改良
    - 位置符号を学習時にみたことない相対位置も対応できるように工夫
- また、Context内で複数の異なるモーダル（文書、画像、音声、動画）を統一的に扱うことができるようになっている
- FineTuningせずにContextに入れてICLする例も増えてきている

Preferred Networks

推論コストの削減 (1/2)

- 推論に必要なコストを減らせるかが重要
  - 提供コストを下げ、大きく強力なモデルを使える
  - ユーザー端末、エッジ端末（スマホ、コントローラー）で使えるように
- TransformerのLLMは学習に比べて推論の効率が非常に悪い
  - トークン毎に、kvキャッシュを読み取る必要があり、メモリ帯域ネック
  - 工夫無しでは実行効率が数%（学習の場合は限界近くの40%まで可能）
- ユーザー体験としてもTime to first token (TTFT)とスループットの改善
  - バッチ化することでTTFTは悪化しスループットは改善される場合が多い

Preferred Networks

推論コストの削減 (2/2)

- より小さなモデルの利用
  - 小さなモデルの能力を大きなモデルによる蒸留などで引き上げる
- 量子化
  - AWQ [Lin 2023]が広く使われる。bit数毎量子化されたモデルが提供
  - BitNet B 1.58 [Ma 2024] 重みを3値(-1/0/1)に。性能はほぼ落ちない
    - 現状はメモリ帯域減少分だけ有効。
- 疎化
  - MoEを除いて、一般に量子化の方が効果が高い [Kuzmin 2023]
- kvキャッシュの構造化（prefix-tree) [Ye 2024]
- 投機的ストリーミング [Bhendawade 2024]
- フラッシュメモリの利用 [Alizadeh2024]

Preferred Networks

マルチモーダル対応

- 画像、音声、動画などへの対応が進んでいる
- 事前学習時から複数モーダルをまとめて学習する [Flamingo, Gemini]
- 言語データを使って画像、音声、動画の認識をブーストすることは進んでいるがその逆は今後おきてくるか？
    - 画像、音声、動画による学習が言語の学習を助けるのか

今後の展望 (1/2)

- モデル大規模化は2023-2024年年から1~2T程度で頭打ちになっている
  - 学習→検証が短くならないと大規模モデル側の進化速度が遅い
  - MoEが進展すればパラメータ数が更に大きくなることも考えられる
    - 今後 10T規模ぐらいまではいくかもしれない
- 大きなモデルを作りそれを使って強力な小さなモデルを作る流れは加速する
  - LLMを使って作られた大量の高品質な学習データを使った学習が進む
    - 学習しやすくなるだけで新しい知識が増えるわけではないことに注意
  - 難しい問題セットを作るなどして推論能力などはまだまだ上げられる
- その場で学習する能力が改善されていった場合、モデル自体は小さくできる
  - RAGやin-contextは必要な情報をよびだし、相対的にモデルは小さく可能

Preferred Networks

今後の展望 (2/2)

- 心理モデルをどのように導入するのか
  - 人の能力と大きなギャップがあり、言語運用にも大きな支障
  - 今の学習をスケールするだけでは獲得できないと思われる。
- 他モーダルとのさらに自然な融合
  - 画像・動画・音声などは必ずしも最適な表現で獲得されていない
  - 事前学習時から融合させていく
- Transformerに代わるブレークスルーはおきるか
  - 状態空間モデルは有望だが、記憶容量が固定でありrecallに優れない
  - 大容量記憶かつ効率的な書き込み/読み込みが可能なアーキテクチャ
- ICLの学習効率は勾配法より良いが、これを事前学習にも使えるか

Preferred Networks

まとめ

産業とアカデミックとの連携が必要

- データセットの整備
- 知見の共有
- 課題の共有
- 異なる強みを持ったリソースを連携させる

PFN/PFEではLLM Open Houseを行います
・3/29 学生向け（満席になりました）
・4/24 一般向け（近日中に公開予定）

ご興味のある方はご参加ください！

Preferred Networks