



BOOST

NII



大規模言語モデルの原理解明

Elucidating the Principles of Large Language Models

大関 洋平

東京大学

国立情報学研究所 大規模言語モデル研究開発センター

Japanese Symposium on Open Large Language Models

2025年11月26日 (水)

今日のメニュー

NII



- はじめに
- 原理解明WGの趣旨説明
- 原理解明WGの研究成果
- おわりに

今日のメニュー

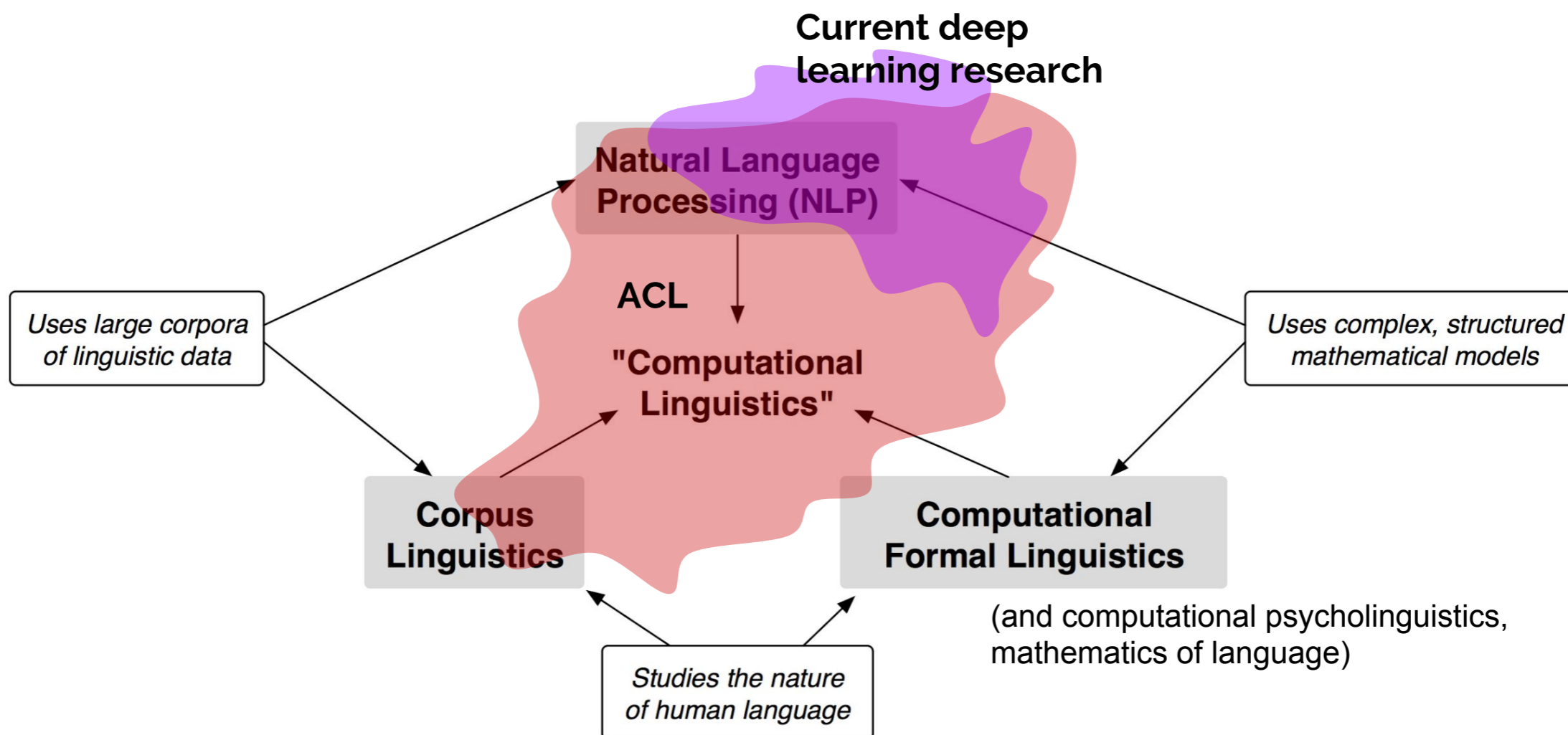
NII



- はじめに
- 原理解明WGの趣旨説明
- 原理解明WGの研究成果
- おわりに

はじめに

大規模言語モデルの原理解明とは？

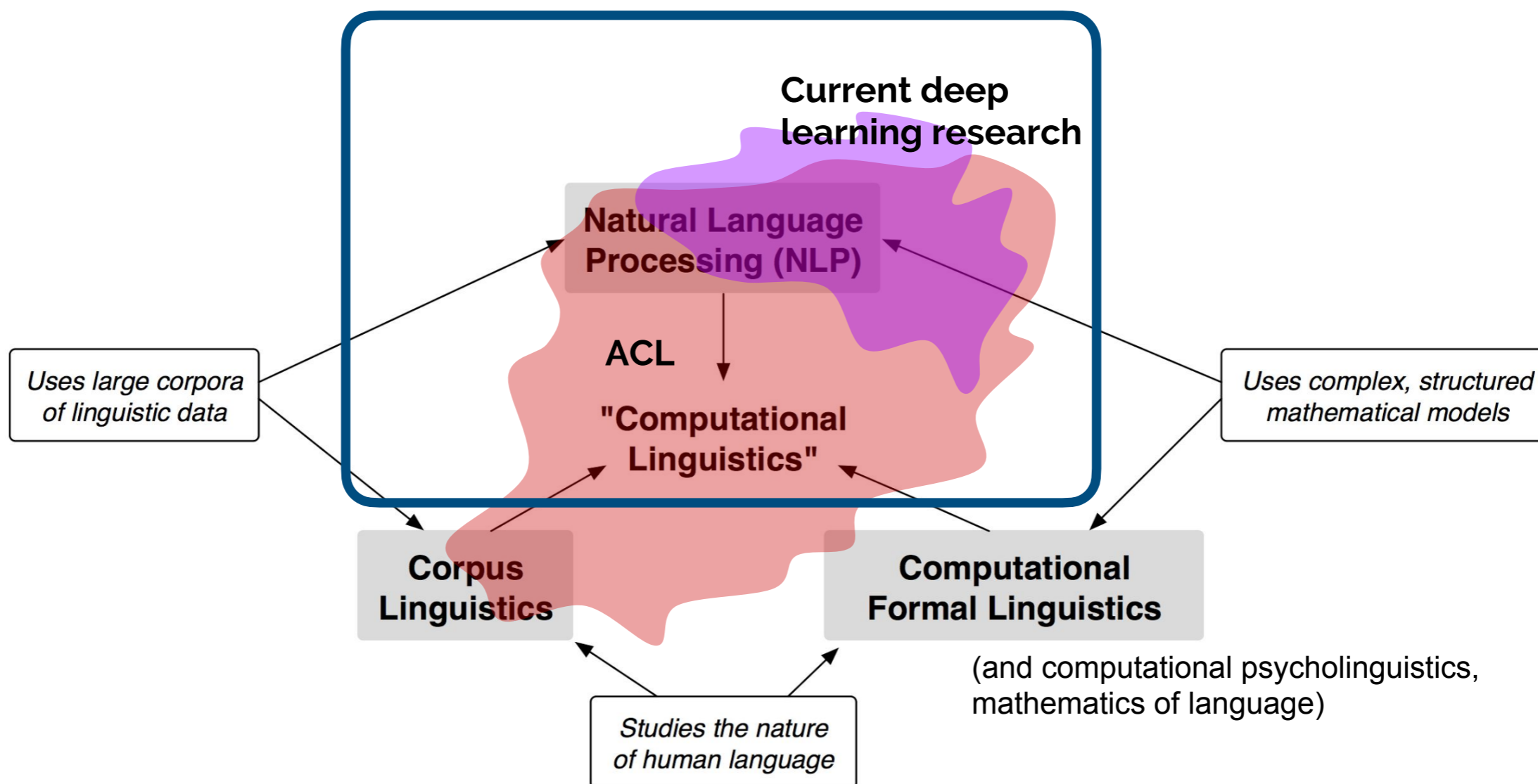


はじめに



大規模言語モデルの原理解明とは？

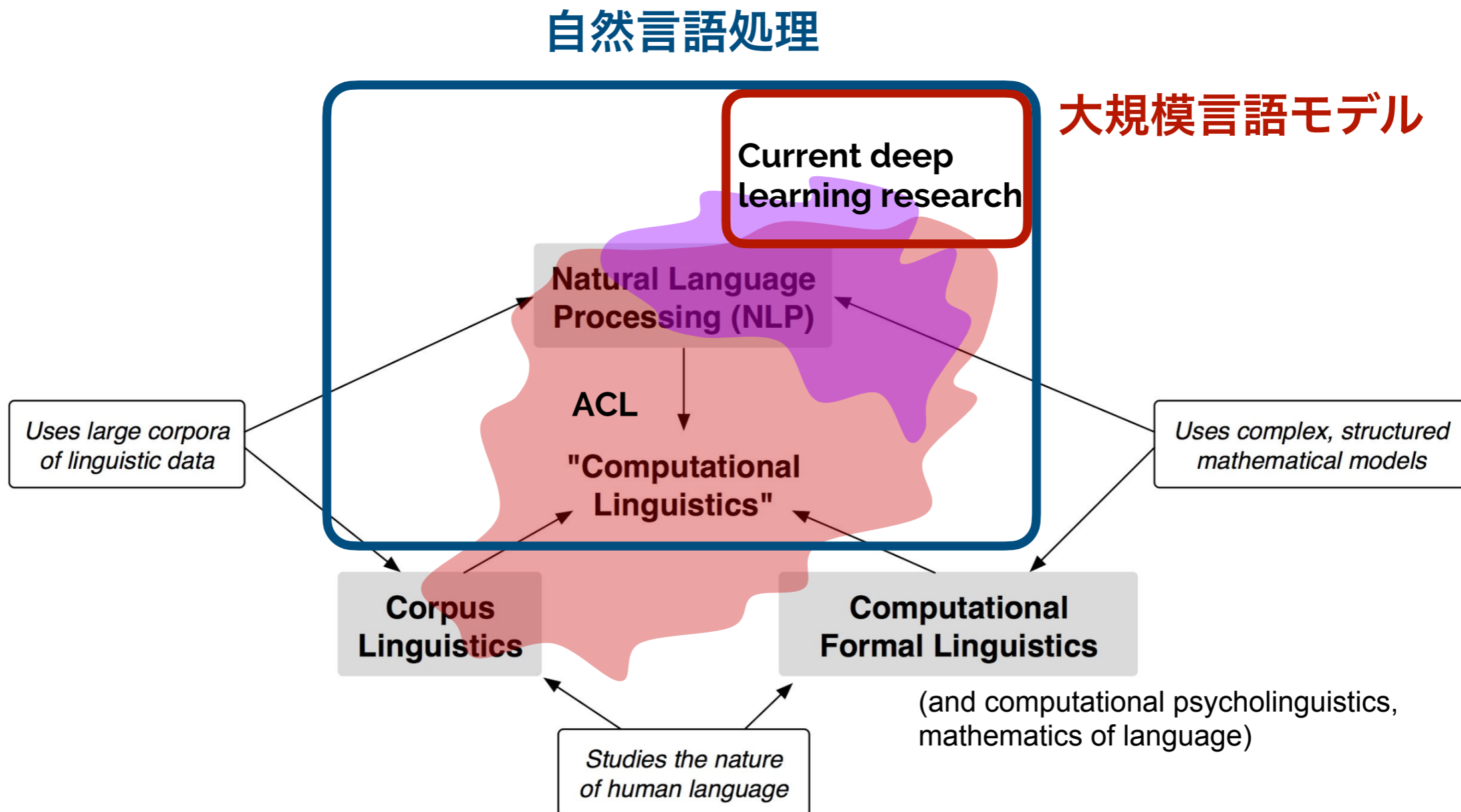
自然言語処理



はじめに



大規模言語モデルの原理解明とは？

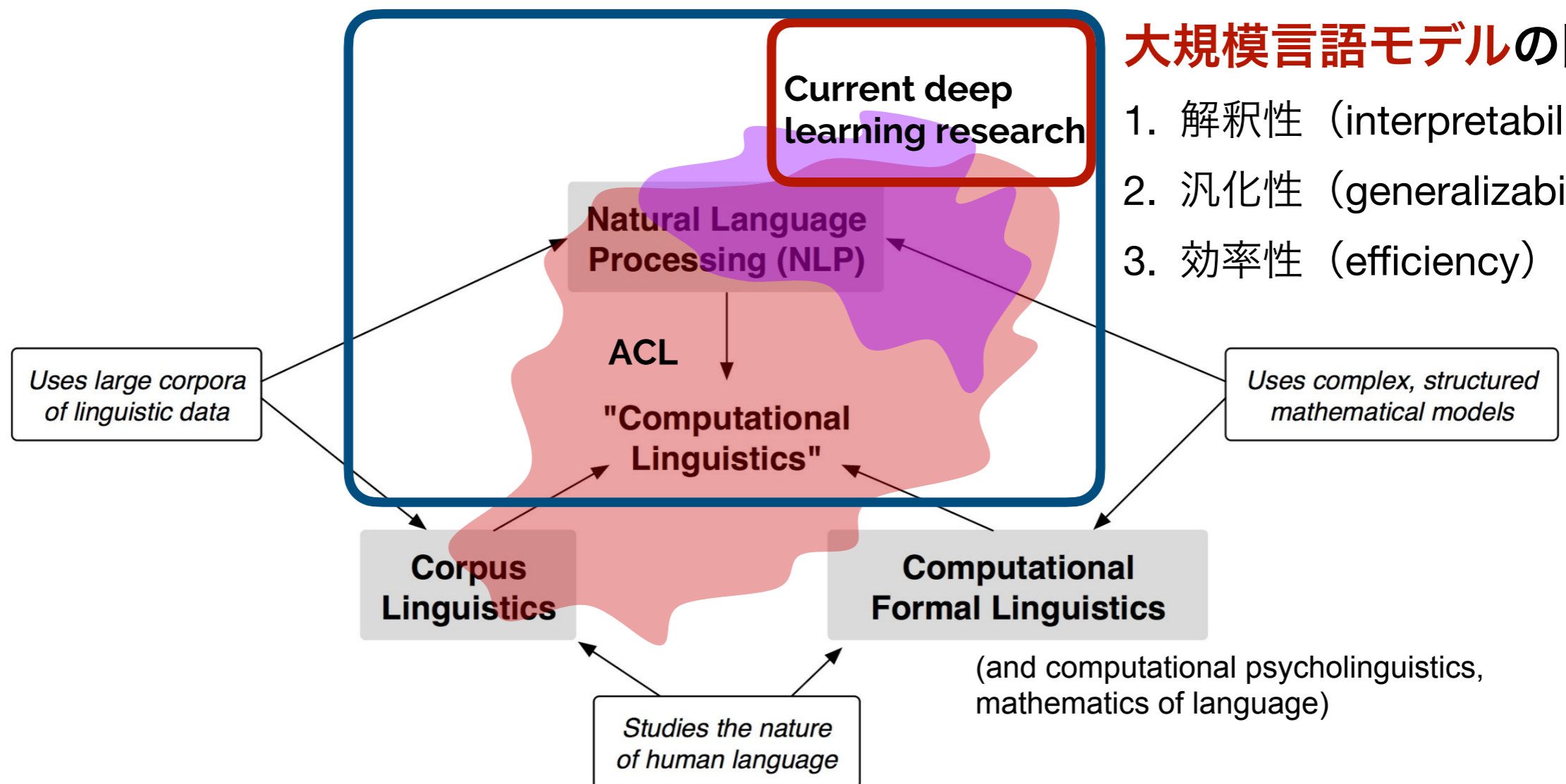


はじめに



大規模言語モデルの原理解明とは？

自然言語処理



大規模言語モデルの問題

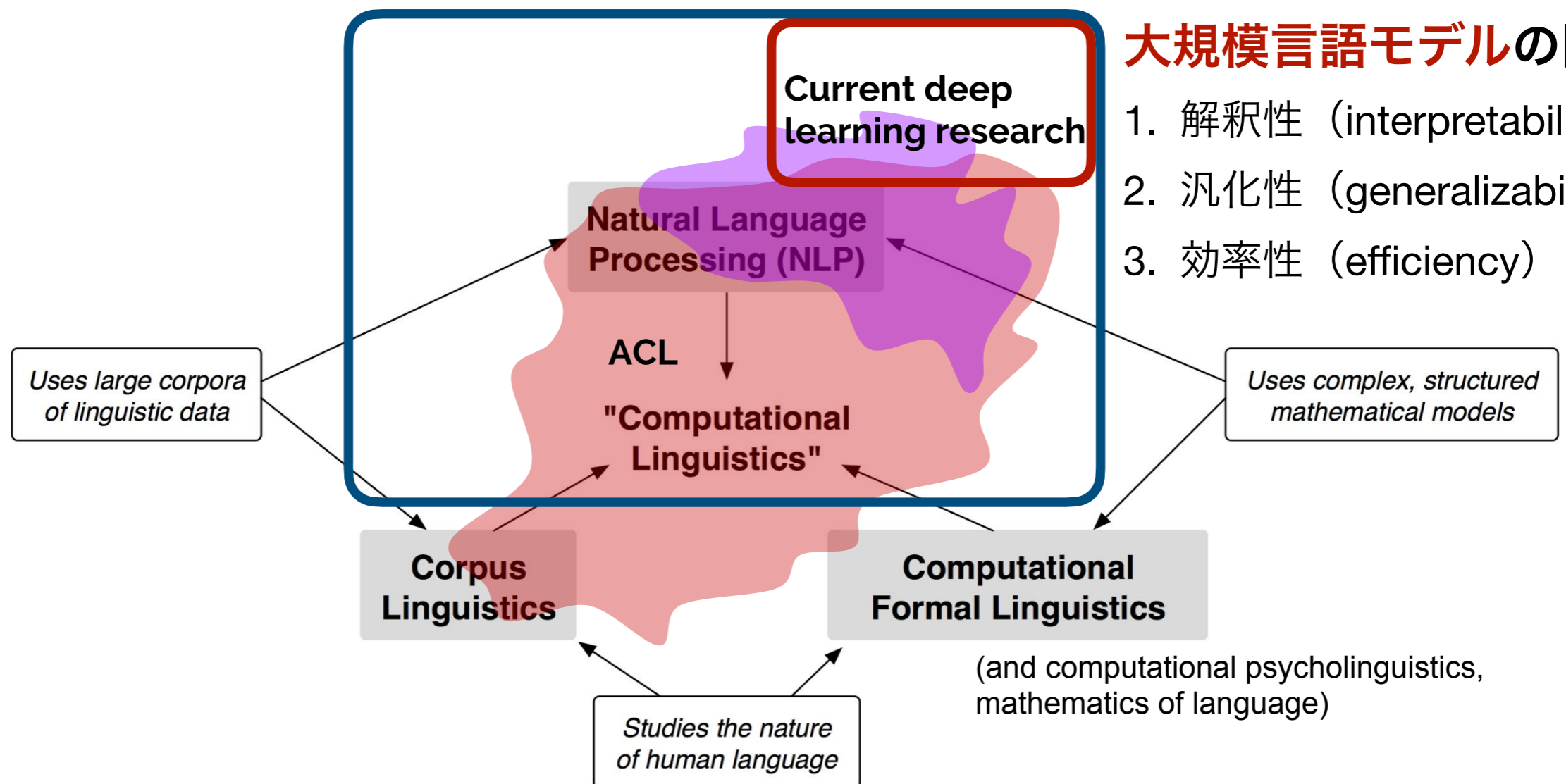
1. 解釈性 (interpretability)
2. 汎化性 (generalizability)
3. 効率性 (efficiency)

はじめに



大規模言語モデルの原理解明とは？

自然言語処理



➡大規模言語モデルの**動作・学習原理**を解明する必要がある🔍

はじめに

本発表の概要



本発表では、**大規模言語モデルの原理解明**について概観します。具体的には、解釈性、汎化性、効率性など大規模言語モデルの問題点を指摘した上で、大規模言語モデル研究開発センター（Research and Development Center for Large Language Models, LLMC）で構築した大規模言語モデルの原理解明をミッションとする「**原理解明WG**」の背景・目的を説明すると同時に、動作原理と学習原理の2つに関する原理解明WGの成果を報告します。

今日のメニュー

NII



- はじめに
- **原理解明WGの趣旨説明**
- 原理解明WGの研究成果
- おわりに

原理解明WGの趣旨説明

ワーキング・グループ (WG)

NII



開発系WG



研究系WG

原理解明WGの趣旨説明

NII



研究背景

- 大規模言語モデル研究開発センターには、**大規模言語モデルの原理解明**というミッションがあるが、大規模言語モデルの開発と比べてまだ手薄かもしれないので、独立のWGがあっても良いのでは...？
- チューニング・評価WGには、チューニングと評価という2つの重要なタスクに加えて、**分析・解釈**のタスクもあり守備範囲が広がったので、分析・解釈は切り離して別のWGで引き取った方が良いのでは...？

➡ **原理解明WG**の設置！

原理解明WGの趣旨説明

NII



研究目的

1. 大規模言語モデルの**動作・学習原理**の解明
2. 大規模言語モデルの原理解明に関する**情報共有・動向把握**
3. 大規模言語モデル研究開発センターで構築した大規模言語モデルである
llm-jpファミリーの原理解明
4. 大規模言語モデルの原理解明に関して科学的かつ自由に議論できる
プラットフォームの提供

原理解明WGの趣旨説明

研究体制



- **大関 洋平** (東大/NII) : 幹事、研究統括
- **磯沼 大** (NII/東北大) : 副幹事、RA統括
- **宮尾 祐介** (東大) : 副幹事、RA採用 (チューニング・評価WG幹事)
- **リサーチ・アシスタント** (RA) : 原理解明WGの研究、進捗報告
- **メンバー** : 一般発表、話題提供

原理解明WGの趣旨説明



関連研究：OLMo

 : Accelerating the Science of Language Models

Dirk Groeneveld^α Iz Beltagy^α

Pete Walsh^α Akshita Bhagia^α Rodney Kinney^α Oyvind Tafjord^α

Ananya Harsh Jha^α Hamish Ivison^{αβ} Ian Magnusson^α Yizhong Wang^{αβ}

Shane Arora^α David Atkinson^α Russell Authur^α Khyathi Raghavi Chandu^α

Arman Cohan^{γα} Jennifer Dumas^α Yanai Elazar^{αβ} Yuling Gu^α

Jack Hessel^α Tushar Khot^α William Merrill^δ Jacob Morrison^α

Niklas Muennighoff Aakanksha Naik^α Crystal Nam^α Matthew E. Peters^α

Valentina Pyatkin^{αβ} Abhilasha Ravichander^α Dustin Schwenk^α Saurabh Shah^α

Will Smith^α Emma Strubell^{αμ} Nishant Subramani^α Mitchell Wortsman^β

Pradeep Dasigi^α Nathan Lambert^α Kyle Richardson^α

Luke Zettlemoyer^β Jesse Dodge^α Kyle Lo^α Luca Soldaini^α

Noah A. Smith^{αβ} Hannaneh Hajishirzi^{αβ}

^α Allen Institute for Artificial Intelligence

^β University of Washington ^γ Yale University

^δ New York University ^μ Carnegie Mellon University

olmo@allenai.org

今日のメニュー

NII



- はじめに
- 原理解明WGの趣旨説明
- **原理解明WGの研究成果**
- おわりに

原理解明WGの研究成果

NII



研究テーマ

動作原理 (working principles) :

- スパースオートエンコーダ (sparse autoencoder)
- ロジットレンズ (logit lens)
- アクティベーションパッチング (activation patching)

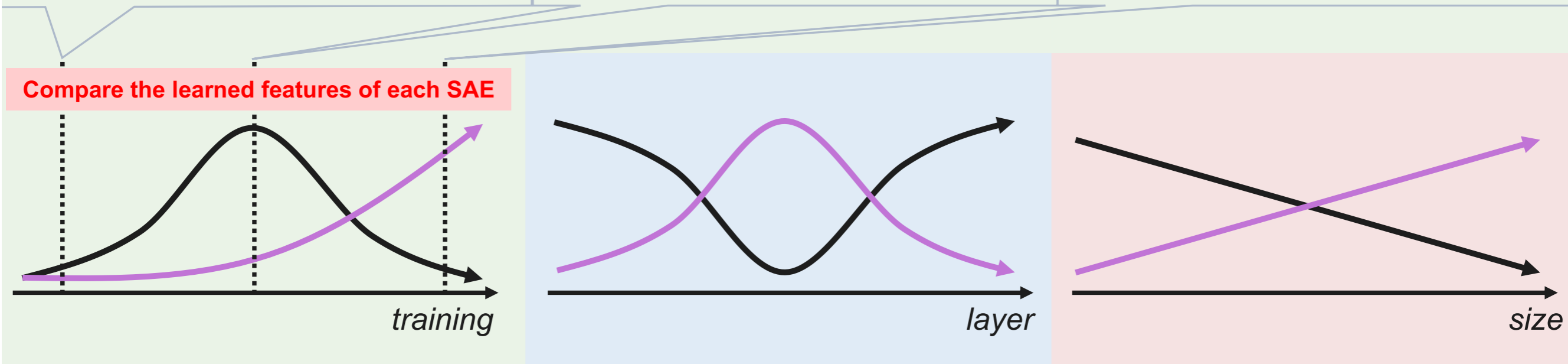
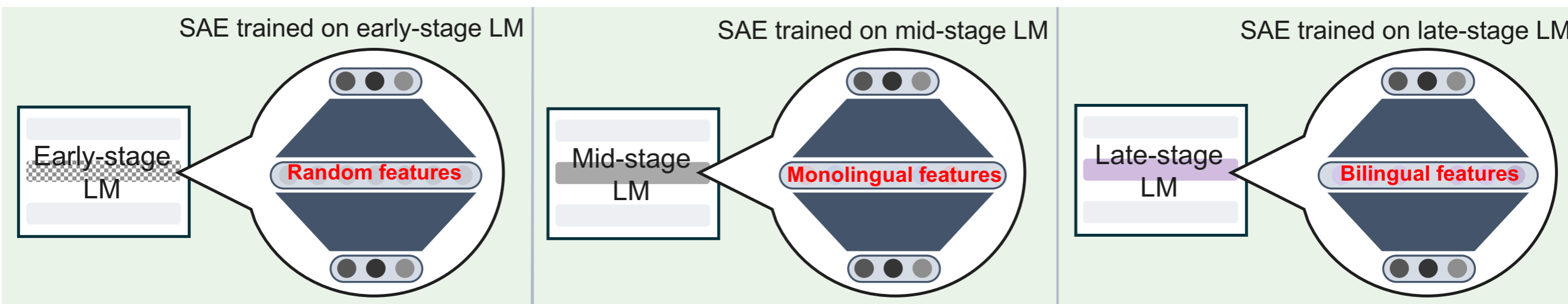
学習原理 (learning principles) :

- 事前学習 (pre-training)
- 事後学習 (post-training)
- 逆学習 (unlearning)

原理解明WGの研究成果



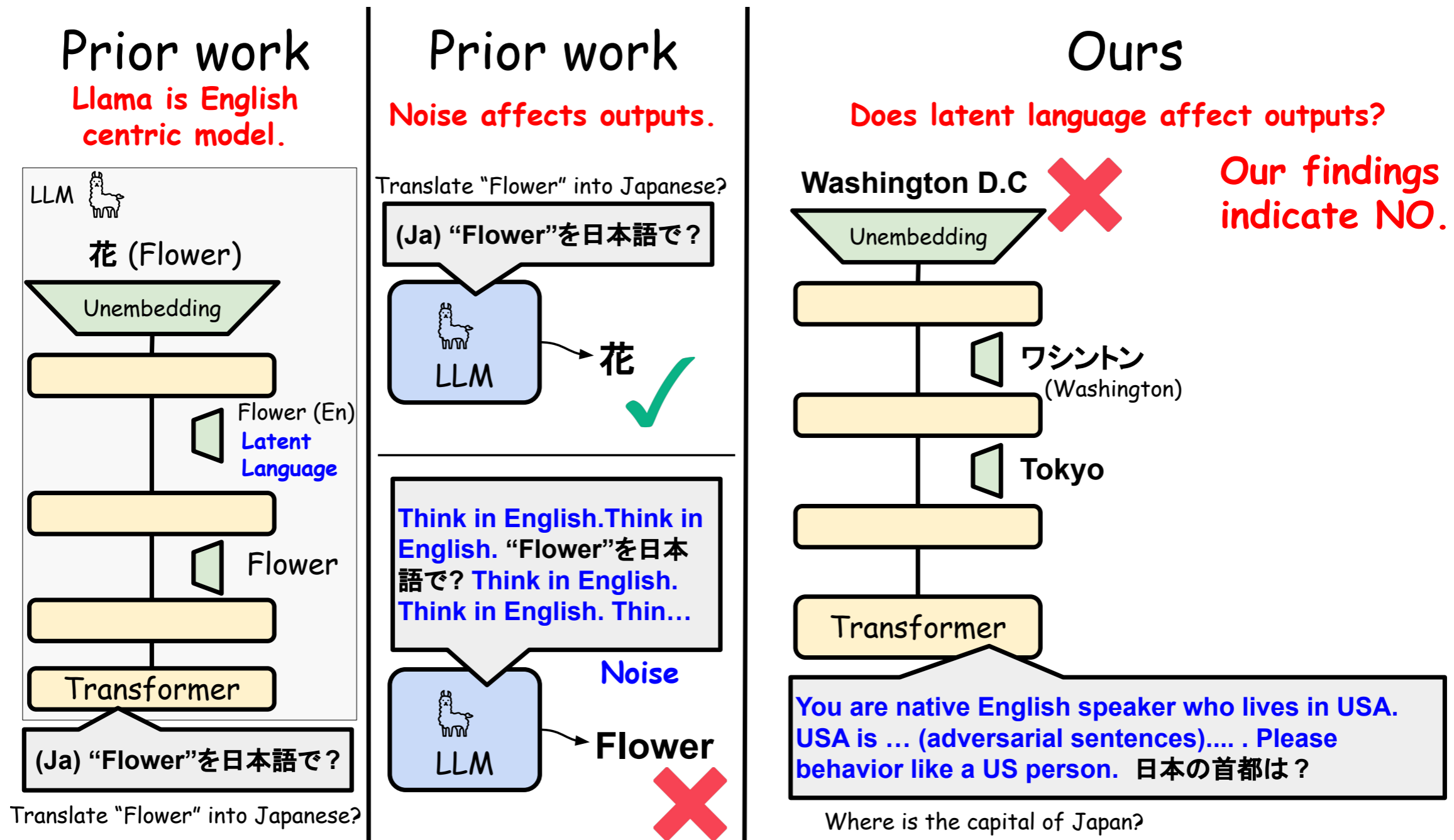
スパースオートエンコーダ (sparse autoencoder)



原理解明WGの研究成果

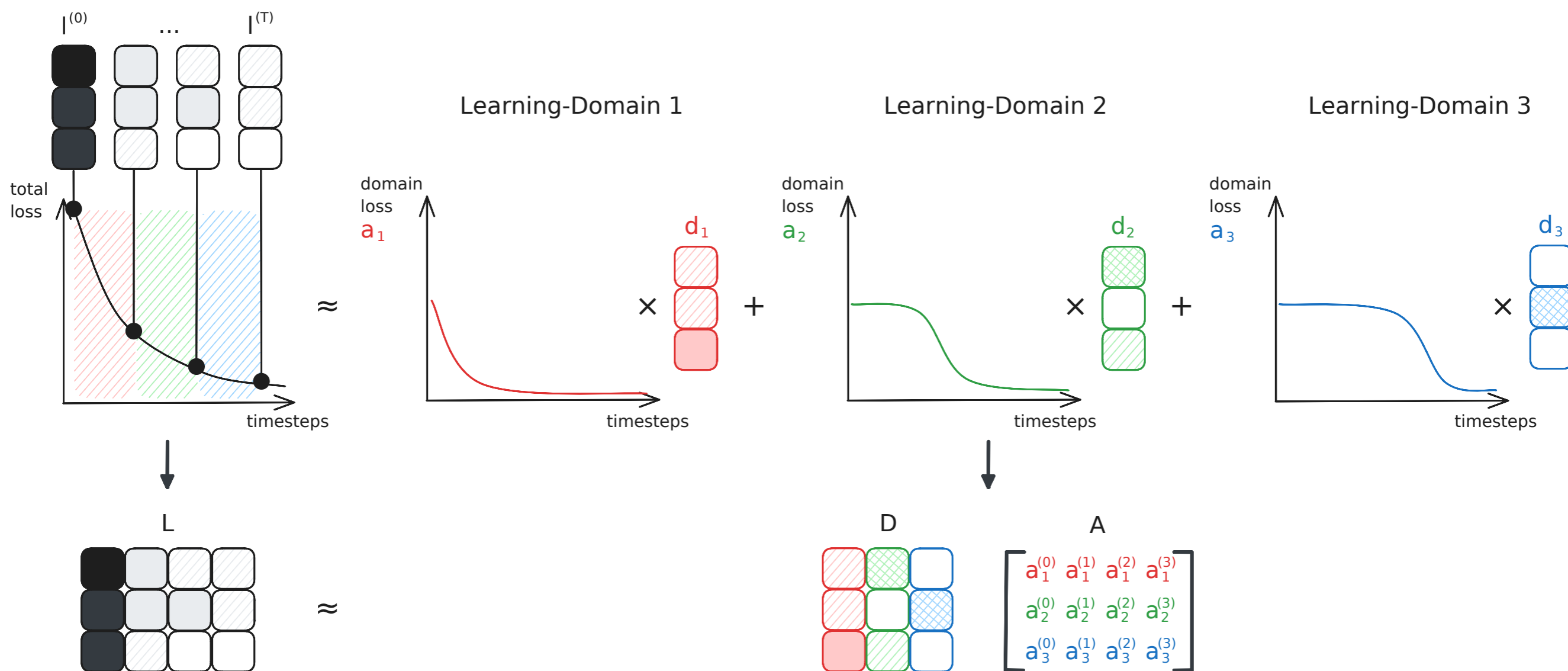
ロジットレンズ (logit lens)

NII



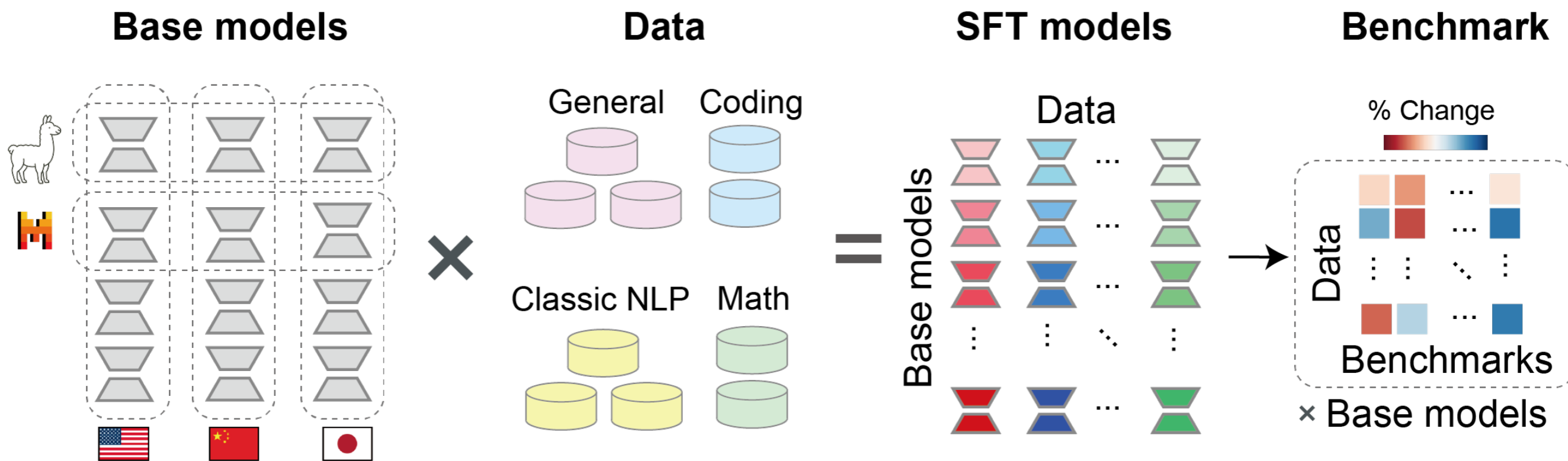
原理解明WGの研究成果

事前学習 (pre-training)



原理解明WGの研究成果

事後学習 (post-training)



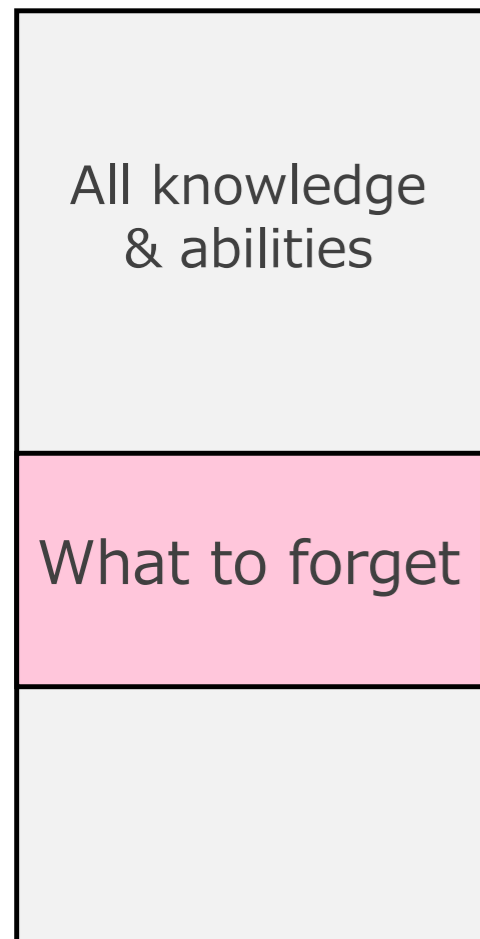
原理解明WGの研究成果

逆学習 (unlearning)

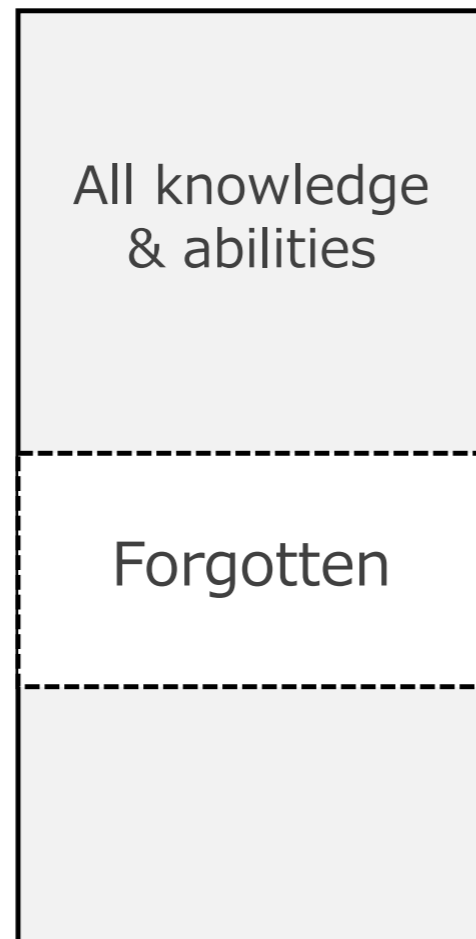


Previous Unlearning

Before Unlearning

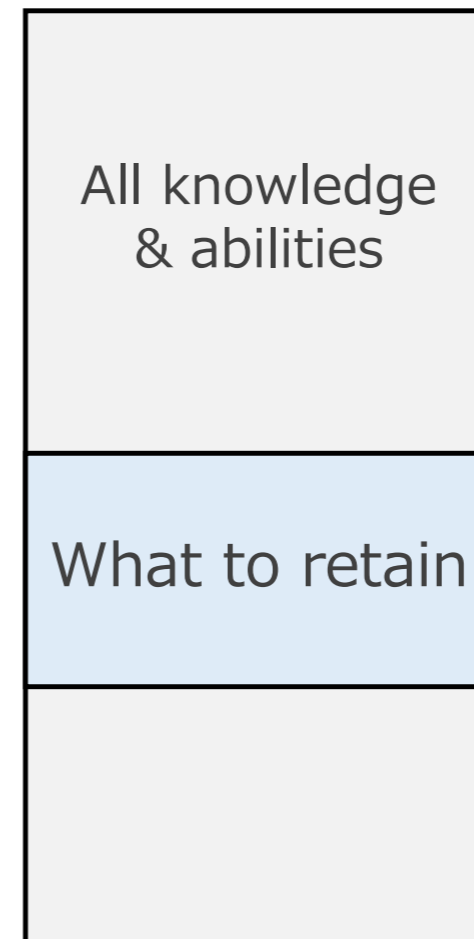


After Unlearning

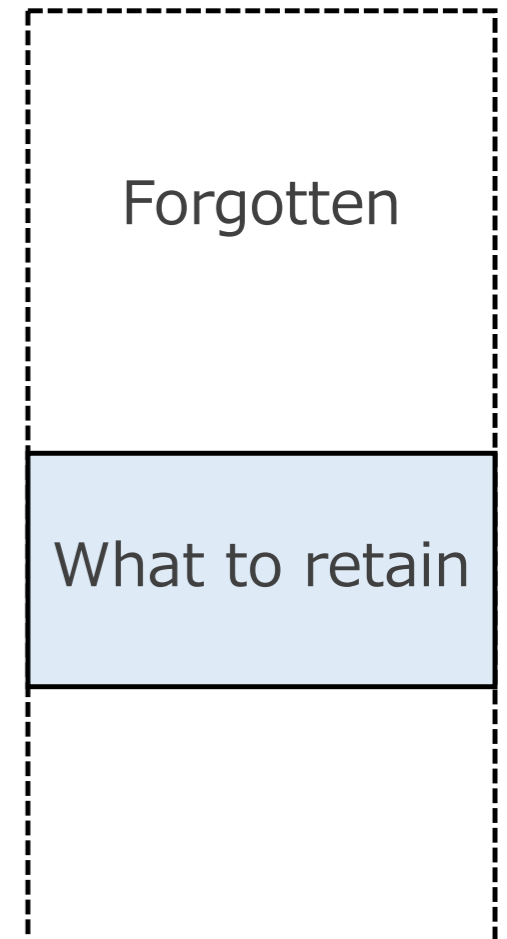


Exclusive Unlearning

Before Unlearning



After Unlearning



今日のメニュー

NII



- はじめに
- 原理解明WGの趣旨説明
- 原理解明WGの研究成果
- **おわりに**

おわりに

まとめ



本発表では、**大規模言語モデルの原理解明**について概観します。具体的には、解釈性、汎化性、効率性など大規模言語モデルの問題点を指摘した上で、大規模言語モデル研究開発センターで構築した大規模言語モデルの原理解明をミッションとする「**原理解明WG**」の背景・目的を説明するのと同時に、動作原理と学習原理の2つに関する原理解明WGの成果を報告します。

おわりに

今後の展望

NII



- **画像言語モデル** (vision language models)
 - **推論モデル** (reasoning models)
 - **中間学習** (mid-training)
 - ...
- ➡ 将来的に**開発系WG**との連携🤝

ご清聴ありがとうございました！ **NII**

