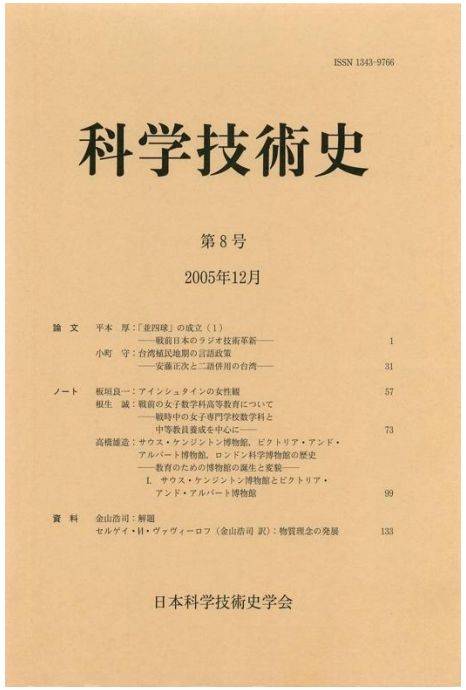
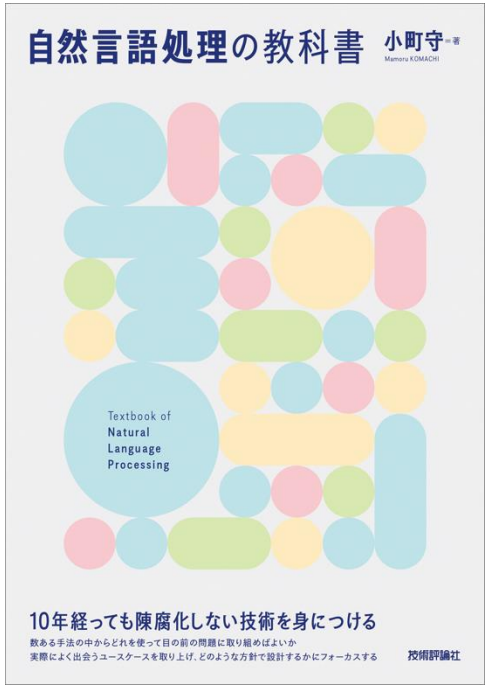
大規模言語モデルの知らない世界

一橋大学大学院
ソーシャル・データサイエンス研究科
小町守
<mamoru.komachi@r.hit-u.ac.jp>

自己紹介: 小町守（こまちまもる）

- 2005.03 東京大学教養学部基礎科学科
  科学史・科学哲学分科卒業
- 2010.03 奈良先端科学技術大学院大学
  博士（工学）
- 2010.04～2013.03 奈良先端大
  助教（自然言語処理研究室）
- 2013.04～2023.03 首都大（現都立大）
  准教授～教授（自然言語処理研究室）
- 2023.04～ 一橋大学大学院
  教授（計算言語学研究室）





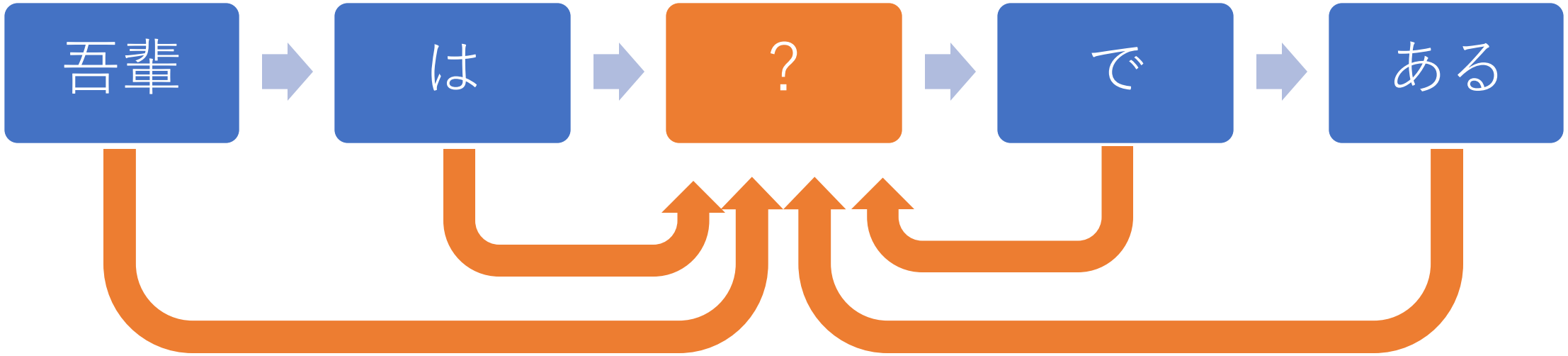
言語モデルが自然言語処理の基礎

吾輩 → は → 猫 → で → ある

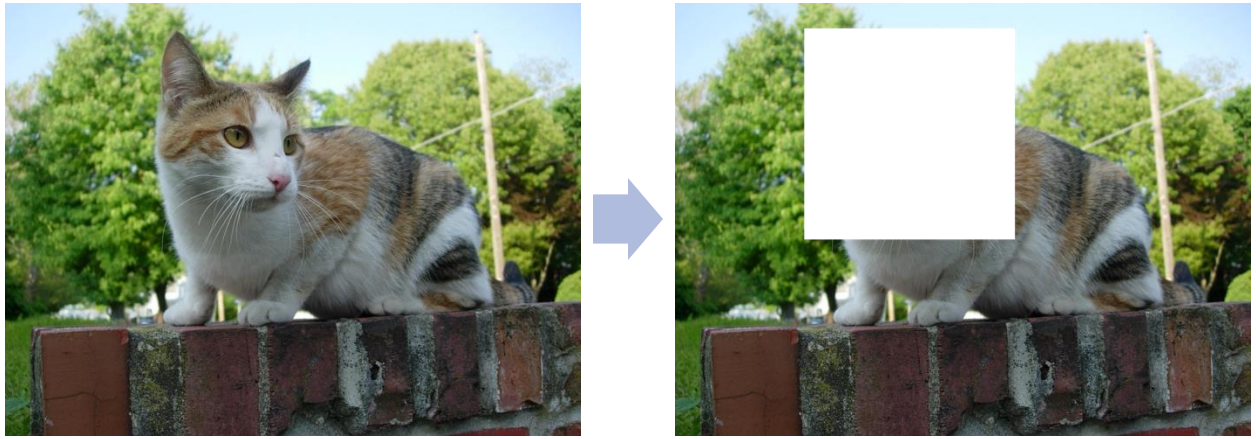
P(吾輩は猫である)　・・・文の生成確率
=P(吾輩)
×P(は|吾輩)
× P(猫|吾輩は)
× P(で|吾輩は猫)
× P(ある|吾輩は猫で)

周辺文脈から単語を予測

マスク言語モデルで自己教師あり学習



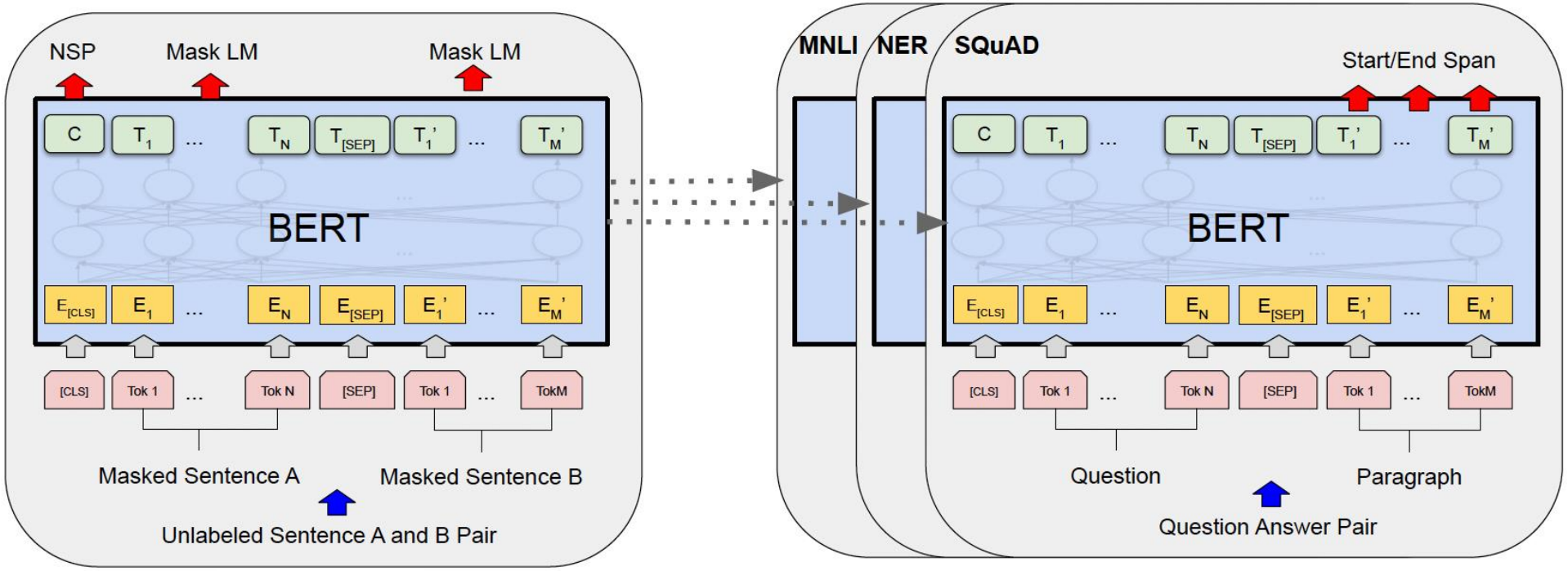
- 単語を隠して（マスク）正解データを作成し、周辺文脈から予測する言語モデルを学習（自己教師あり学習）



深層学習モデルを用いた文のモデリング

- マスク言語モデルで自己教師あり学習した言語モデル

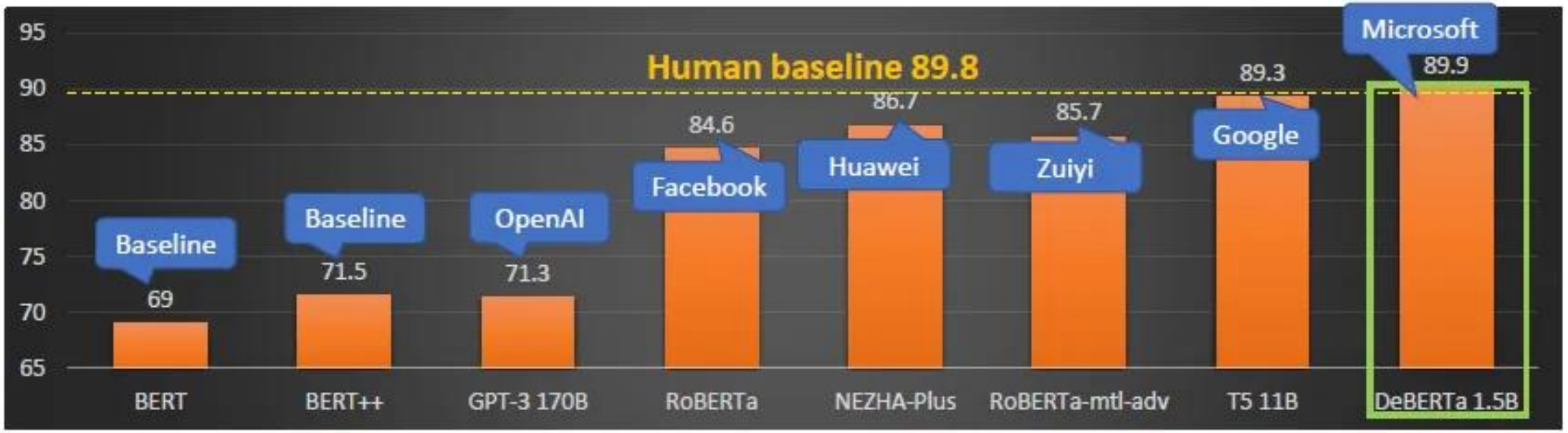
BERT: Bidirectional Encoder Representations from Transformers [Devlin+, 2019]



事前学習　　微調整

少ないデータで微調整するだけで多くのタスクに転用可能

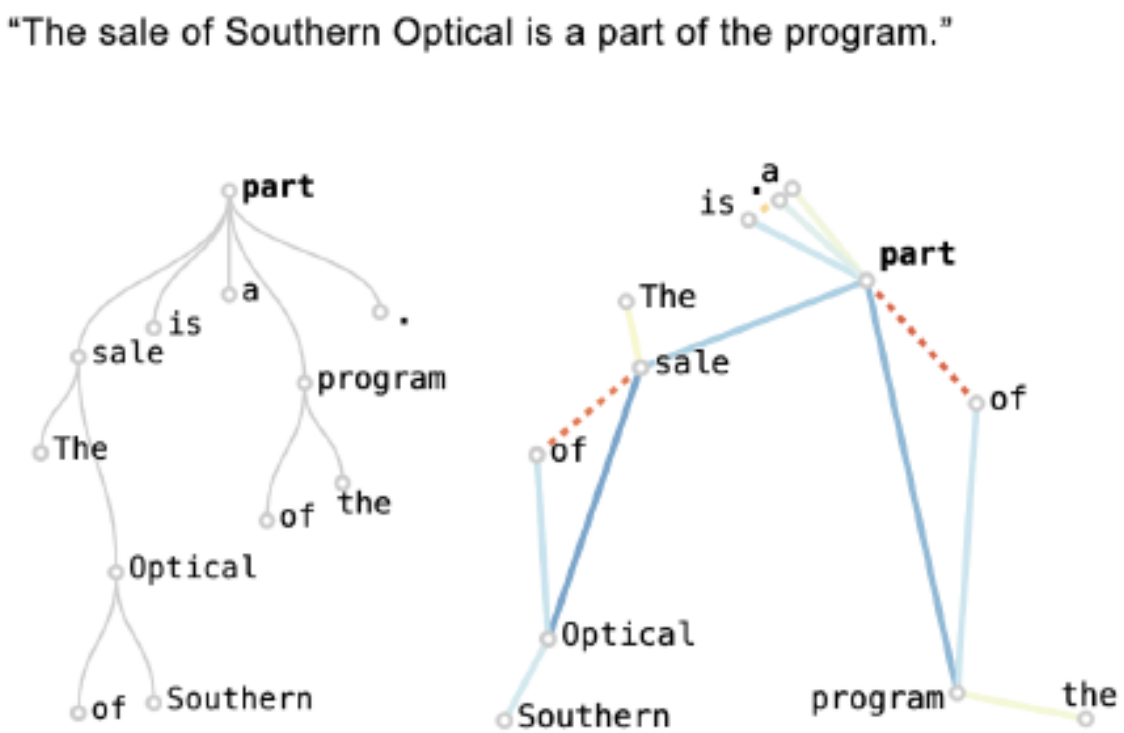
言語理解ベンチマーク (SuperGLUE) で 2021 には人間を超えるパフォーマンス



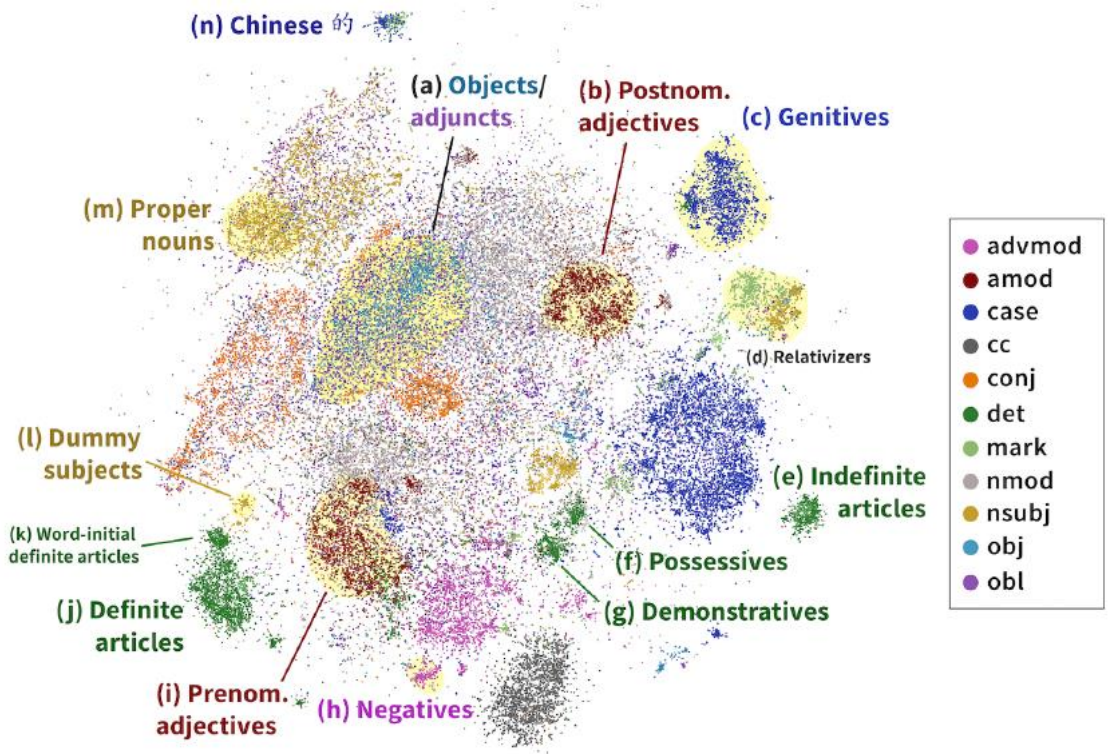
https://sh-tsang.medium.com/brief-review-deberta-decoding-enhanced-bert-with-disentangled-attention-f5cdb9a8bf0b

マスク言語モデルによる文のモデリング: BERT は何をエンコードしているか？

- 文法や意味をエンコード [**Coenen+, 2019**]



- 多言語 BERT でも文法をエンコード [**Chi+, 2020**]



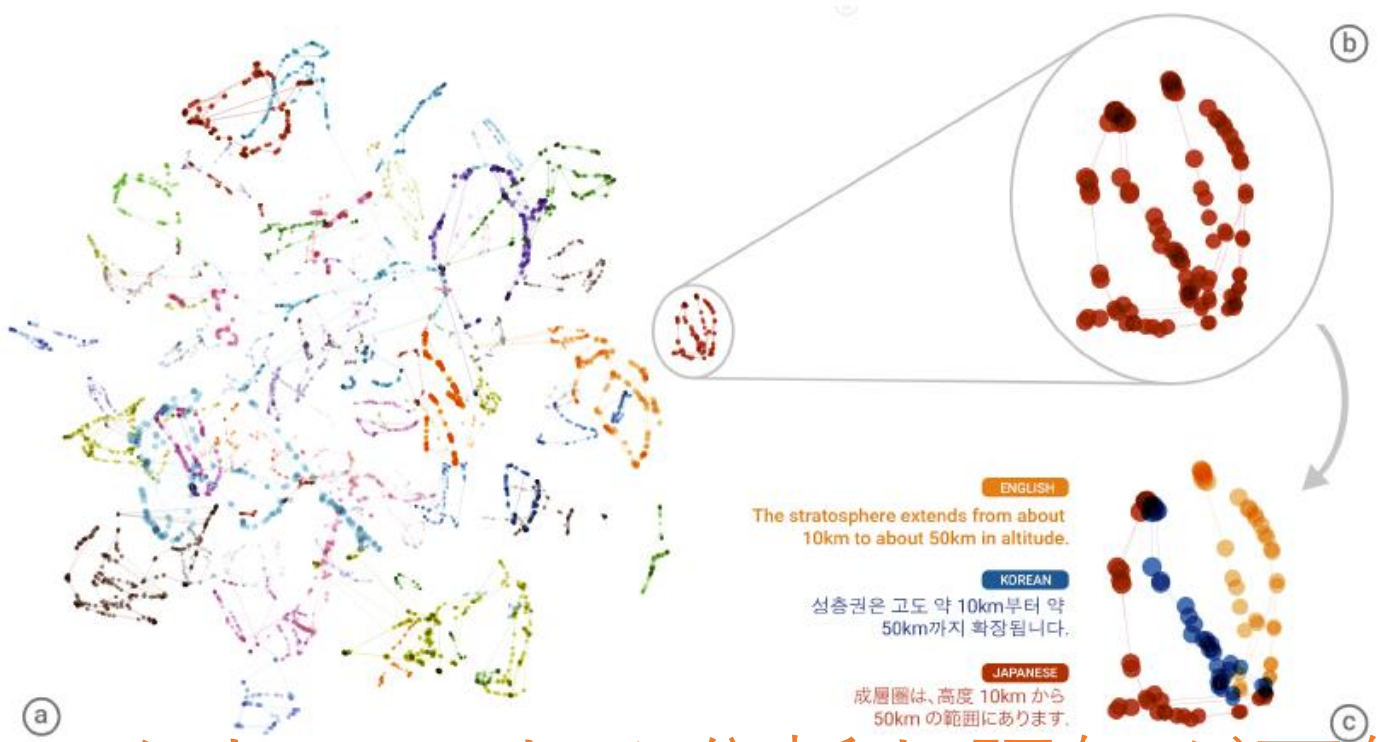
単語ベクトルによる用例の分析が可能

クロスリンガル言語モデルを用いた文の意味の分析と可視化

- 複数コーパスで同時学習した言語モデルによる言語表現の分析

Google's Multilingual Neural Machine Translation System [Johnson+, 2017]

複数言語対での同じ意味の文ベクトルの可視化（同じ色が同じ意味の文）



- **Japanese:** 私は東京大学の学生です。 → I am a student at Tokyo University.
- **Korean:** 나는 도쿄 대학의 학생입니다. → I am a student at Tokyo University.
- **Japanese/Korean:** 私は東京大学학생입니다. → I am a student of Tokyo University.

w_{ko}	I must be getting somewhere near the centre of the earth.
0.00	私は地球の中心の近くにとこかに行っている に違いない。
0.40	私は地球の中心近くのとこかに着いているに 違いない。
0.56	私は地球の中心の近くのとこかになっている に違いない。
0.58	私は지구の中心의가까이에어딘가에도착하고있 어야한다。
0.60	나는지구의센터의가까이에어딘가에도착하고있 어야한다。
0.70	나는지구의중심근처어딘가에도착해야합니다。
0.90	나는어딘가지구의중심근처에도착해야합니다。
1.00	나는어딘가지구의중심근처에도착해야합니다。

文ベクトルによる分析と評価が可能

大規模言語モデル (LLM) を支える技術

多言語データで
大規模言語モデルを
自己教師あり学習

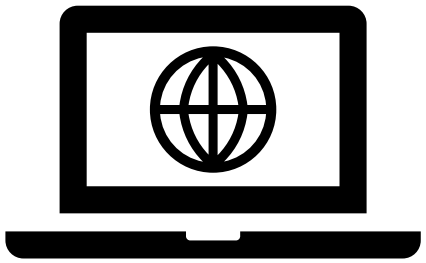
どういう出力が良くて
どういう出力がダメか
指示に従わせる

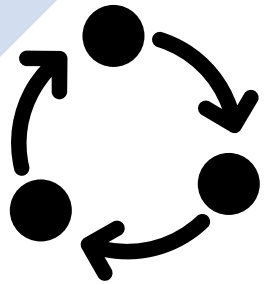
どう生成すれば人間に
とって好ましい出力に
なるのか学習する

事前学習

教師あり学習

強化学習





プロンプトを用いた LLM の操作

- プロンプト（テキストによる指示）でタスクを指示可能

Language Models are Few-Shot Learners [Brown+, 2020]

Few-shot

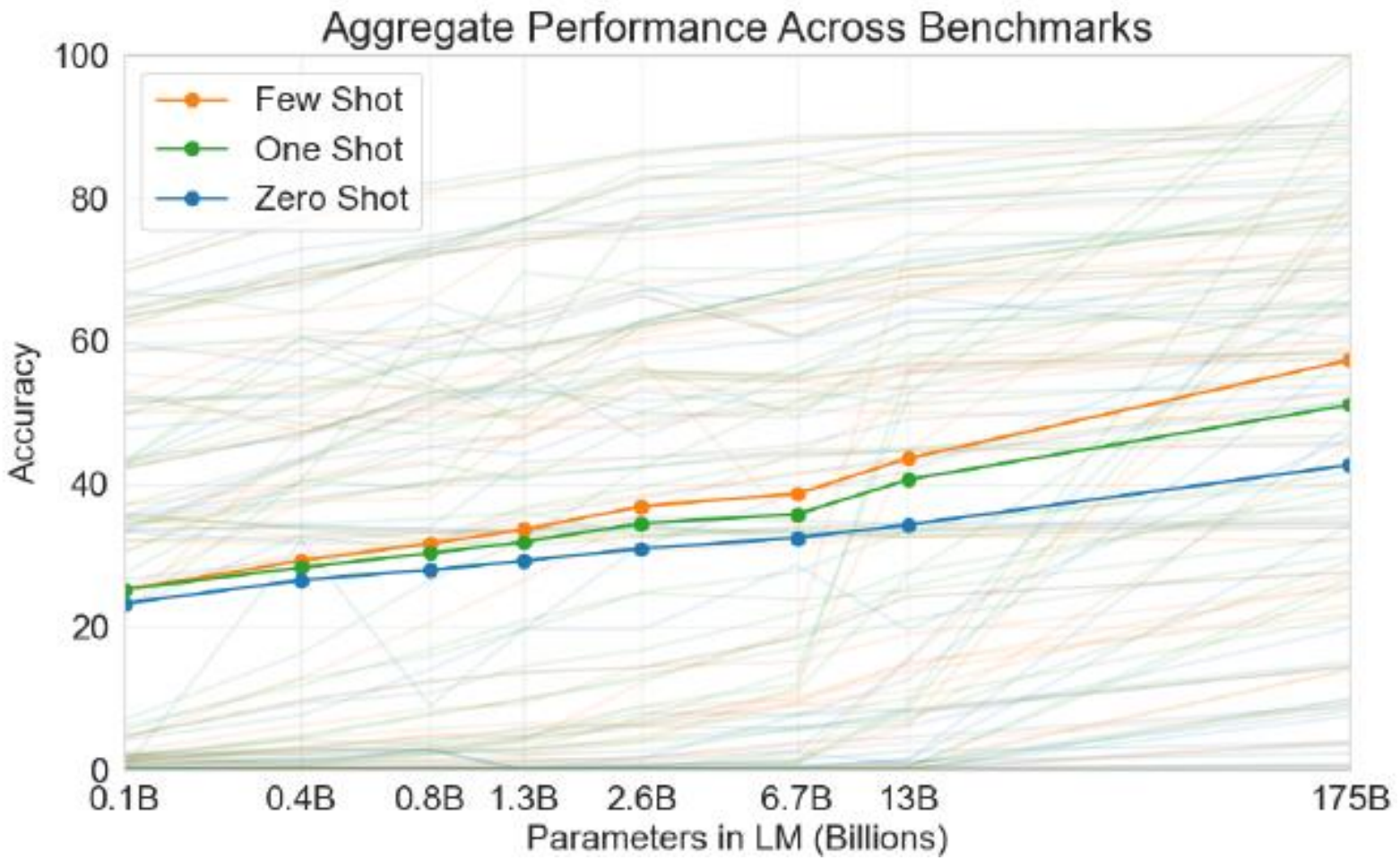
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ← examples
4  plush girafe => girafe peluche      ← examples
5  cheese =>                           ← prompt
```

言語による説明と事例があれば、微調整不要

```
Title:  United Methodists Agree to Historic Split
Subtitle:  Those who oppose gay marriage will form their own denomination
Article:  After two days of intense debate, the United Methodist Church
has agreed to a historic split - one that is expected to end in the
creation of a new denomination, one that will be "theologically and
socially conservative," according to The Washington Post.  The majority of
delegates attending the church's annual General Conference in May voted to
strengthen a ban on the ordination of LGBTQ clergy and to write new rules
that will "discipline" clergy who officiate at same-sex weddings.  But
those who opposed these measures have a new plan:  They say they will form a
separate denomination by 2020, calling their church the Christian Methodist
denomination.
The Post notes that the denomination, which claims 12.5 million members, was
in the early 20th century the "largest Protestant denomination in the U.S.,"
but that it has been shrinking in recent decades.  The new split will be the
second in the church's history.  The first occurred in 1968, when roughly
10 percent of the denomination left to form the Evangelical United Brethren
Church.  The Post notes that the proposed split "comes at a critical time
for the church, which has been losing members for years," which has been
"pushed toward the brink of a schism over the role of LGBTQ people in the
church." Gay marriage is not the only issue that has divided the church.  In
2016, the denomination was split over ordination of transgender clergy, with
the North Pacific regional conference voting to ban them from serving as
clergy, and the South Pacific regional conference voting to allow them.
```

GPT: 大規模言語モデル時代の新常識



モデルサイズ（横軸）を大きくすればするほど性能（縦軸）が向上

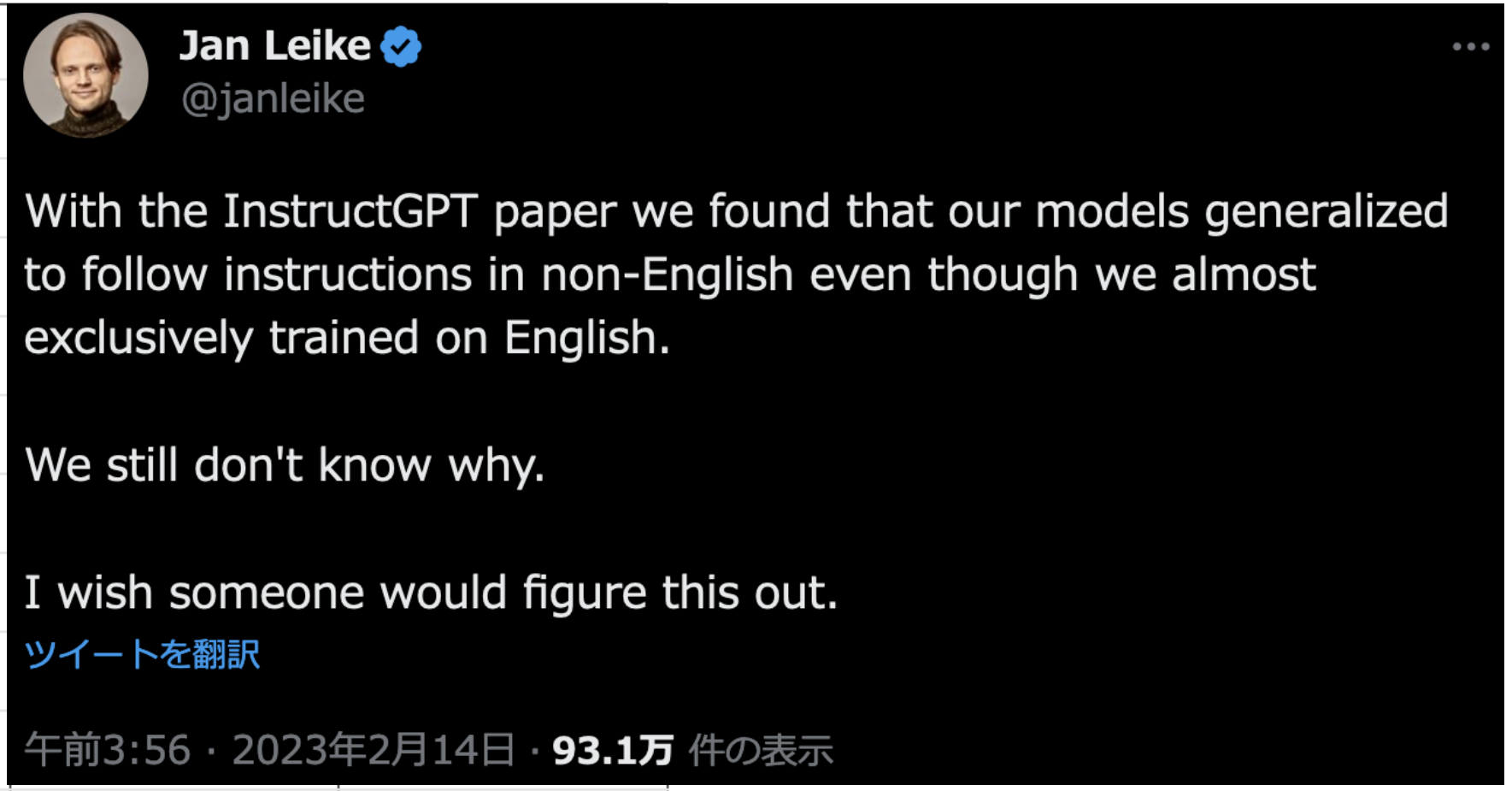
突然できるようになることがある（創発）

Language Models are Few-Shot Learners [Brown+, 2020]

学習データの大半は英語だがなぜか動く

crawl	CC-MAIN-2022-49
language	%
eng	46.3044
deu	5.8640
rus	5.6647
fra	4.7768
zho	4.5969
spa	4.5435
jpn	4.4586
<unknown>	2.7037

ウェブ中に
日本語は4%



https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

https://twitter.com/janleike/status/1625207251630960640

LLM に関する3つのリサーチクエスチョン

多言語大規模言語モデルは知らない言語でもなぜ動く？

→Pruning Multilingual Large Language Models for Multilingual Inference (EMNLP 2024 Findings)

多言語大規模言語モデルは英語の方が性能が高いって本当？

→多言語大規模言語モデルにおける英語指示文と対象言語指示文の公平な比較（言語処理学会年次大会 NLP2025 発表予定）

大規模言語モデルの指示チューニングって、本当にいいの？

→アライメントが大規模言語モデルの数値バイアスに与える影響（言語処理学会年次大会 NLP2025 発表予定）

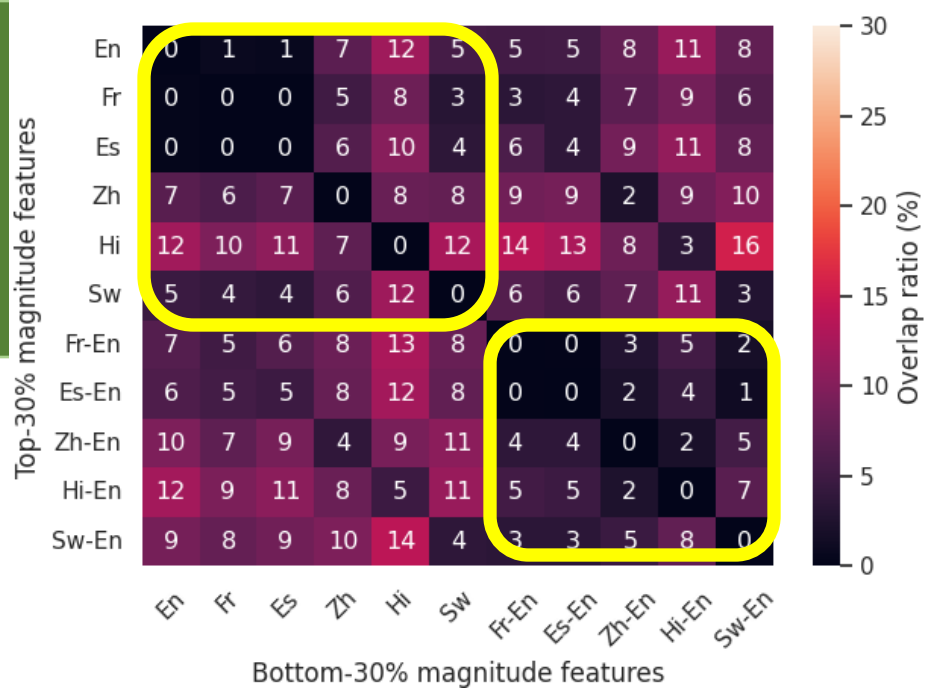
多言語 LLM は知らない言語でもなぜ動く？

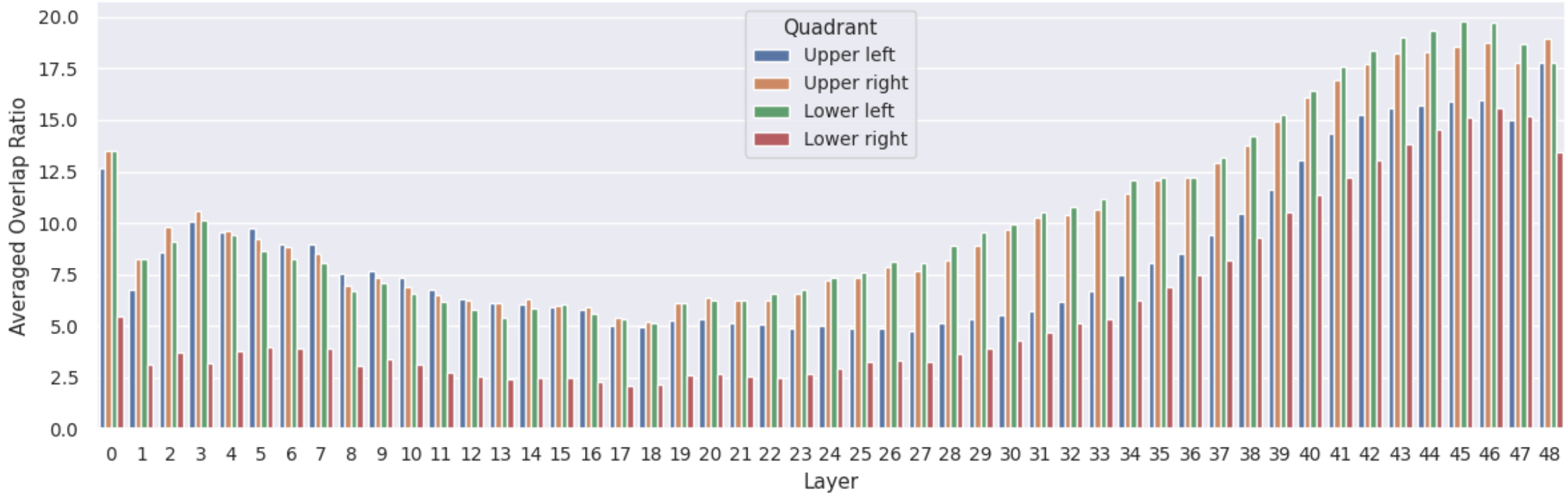
1. LLM の内部で知らない言語から知っている言語への翻訳をしているのではないか？
   - 翻訳タスクでのみ活性化される特徴量が存在する？
   - 翻訳タスクでの活性度合いが翻訳性能と関係している？

2. LLM の翻訳能力を活用して、知らない言語での下流タスクでの性能を伸ばすことができないか？

翻訳タスクでのみ重要度が高くなる特徴量がある

単言語（第2象限）多言語（第4象限）同士では活性度が重なっていない（言語固有ニューロンの存在も知られている）

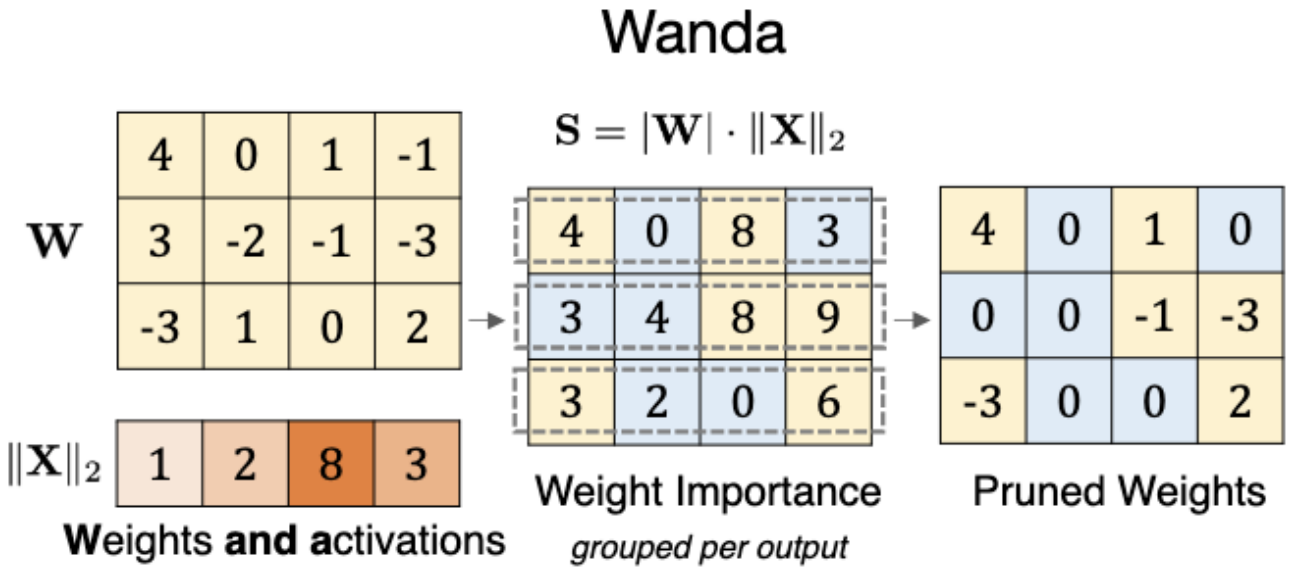




機械翻訳タスクでの特徴量の重要度を考慮した多言語 LLM の重みの枝刈り

機械翻訳タスクのプロンプトを入れた際の活性度を考慮して重みを枝刈り（右図で青く塗ったセルの重みを0に潰す）

→翻訳タスクでの活性度と翻訳性能の関係、活性度を用いた枝刈り後の下流タスクでの性能を見る



重み行列 W
隠れ層の出力 X

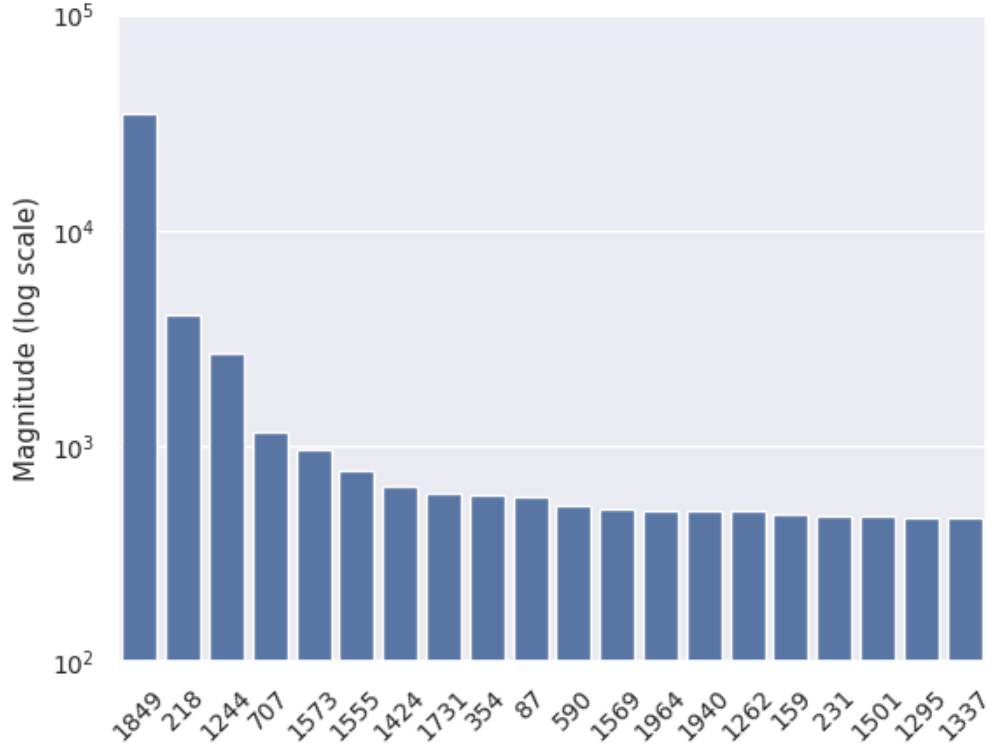
特徴量の重要度 S
出力ごとに枝刈り
（活性度を考慮）

最終的な重み W

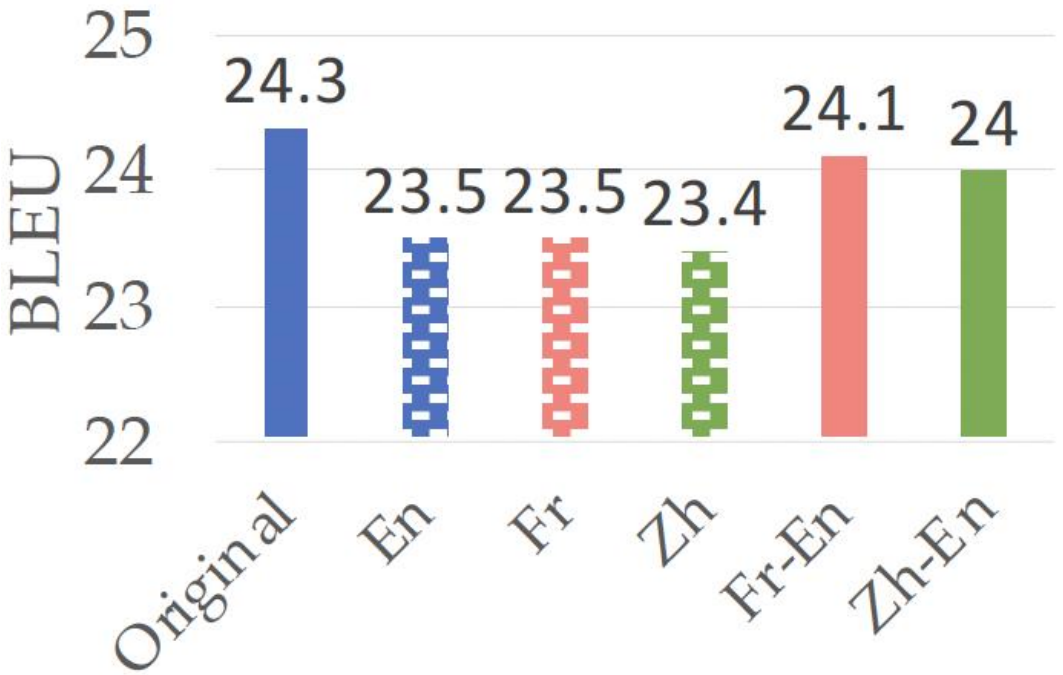
A SIMPLE AND EFFECTIVE PRUNING APPROACH FOR LARGE LANGUAGE MODELS [Sun+, 2024]

翻訳タスクでの重要度の高い特徴量が多言語 LLMの翻訳性能と関係している

- 中英翻訳タスクで重要度が高い特徴量



- 重要度の低い特徴量を枝刈りしても翻訳性能は落ちない



翻訳タスクで枝刈りした多言語 LLM はゼロショット転移能力が向上する

翻訳タスクで重みを枝刈りした LLM $\Theta_{src-tgt}$の評判分析実験

実験設定

- XGLM-2.9B, mGPT-1.7B, BLOOM-3B（後述）を比較
  →GPT と似たオープン LLM
- Θ_{Rand}はランダムに重みを落とす弱いベースライン
- LRP2 (Language Represenation Projection) は目的言語を使って重みを調整する強いベースライン

Model	Weight	De	Ja	Fr	Es	Zh	AVg.
XGLM	θ	64.0	64.6	60.2	60.8	67.6	63.4
	LRP2	-	-	60.4	60.1	66.3	-
	θ_{Rand}	49.3	52.3	55.4	53.4	51.1	52.3
	$\theta_{\text{Fr-En}}$	64.4†	**66.9**†	**60.8**†	60.8	69.2†	**64.4**
	$\theta_{\text{Es-En}}$	64.1	66.3†	**60.8**†	**61.0**†	68.7†	64.2
	$\theta_{\text{Zh-En}}$	**64.5**†	64.4	60.5†	60.0	68.0†	63.5
	$\theta_{\text{Hi-En}}$	63.8	66.1†	60.0	60.4	**69.5**†	63.9
	$\theta_{\text{Sw-En}}$	63.9	64.8	59.8	60.3	68.5†	63.5
mGPT	θ	65.9	56.3	64.1	65.6	53.9	61.2
	LRP2	-	-	51.7	52.4	50.7	-
	θ_{Rand}	52.8	54.4	50.3	50.8	51.9	52.0
	$\theta_{\text{Fr-En}}$	**66.4**†	57.1†	63.6	65.7	55.2†	61.6
	$\theta_{\text{Es-En}}$	**66.4**†	**57.2**†	64.6†	**65.9**†	55.3†	**61.9**
	$\theta_{\text{Zh-En}}$	66.1	56.9†	**64.7**†	**65.9**†	**55.5**†	61.8
	$\theta_{\text{Hi-En}}$	66.2†	57.7†	63.7	65.6	55.4†	61.7
	$\theta_{\text{Sw-En}}$	66.3†	57.4	64.2	65.8	55.2†	61.7

翻訳タスクで枝刈りした多言語 LLM はゼロショット転移能力が向上する

枝刈りした LLM $\Theta_{src-tgt}$ はクロスリンガル推論の性能が向上

Model	Weight	Ar	Bg	De	El	Hi	Ru	Sw	Th	Tr	Ur	Vi	Fr	Es	Zh	Avg.
XGLM	θ	44.0	41.8	41.6	44.0	44.7	39.5	42.5	44.2	41.0	42.7	45.2	45.0	35.2	43.9	42.5
	LRP2	-	-	-	-	44.6	-	42.4	-	-	-	-	46.4	36.0	45.1	-
	θ_{Rand}	32.4	33.5	34.6	33.9	33.2	33.5	34.3	34	33.2	34.8	33.5	36.2	34.4	33.9	33.9
	$\theta_{\text{Fr-En}}$	**44.9**†	45.8†	42.9†	**46.2**†	44.9	43.0†	**43.5**†	45.4†	42.5†	42.8	**47.7**†	47.2†	39.7†	**47.2**†	**44.5**
	$\theta_{\text{Es-En}}$	44.3	**45.9**†	42.5†	45.6†	44.7	43.2†	43.3†	**45.9**†	**42.8**†	42.5	47.5†	47.0†	36.3†	46.8†	44.1
	$\theta_{\text{Zh-En}}$	44.1	**45.9**†	**43.0**†	45.9†	**45.0**	**43.6**†	42.5	45.5†	42.2†	**43.0**	47.5†	**47.7**†	**39.8**†	46.7†	44.4
	$\theta_{\text{Hi-En}}$	43.7	42.8†	40.3	45.2†	44.7	42.5†	42.7	45.5†	41.0	42.2	45.5	46.2	37.2†	46.0†	43.3
	$\theta_{\text{Sw-En}}$	43.8	44.5†	40.3	46.1†	43.7	41.7†	42.0	45.6†	41.7	42.0	45.4	46.4†	38.9†	46.1†	43.4
mGPT	θ	39.2	39.7	35.0	41.0	38.9	39.2	34.0	41.6	39.9	39.9	42.2	42.3	39.4	**41.8**	39.6
	LRP2	-	-	-	-	35.2	-	34.4	-	-	-	-	34.2	33.1	34.1	-
	θ_{Rand}	33.3	33.2	32.9	33.2	33.3	33.3	33.4	33.0	33.0	33.2	33.5	33.6	34.0	33.1	33.2
	$\theta_{\text{Fr-En}}$	39.3	39.5	36.3†	41.9†	**40.2**†	39.5†	34.7†	42.6†	**40.2**†	40.0	42.8†	42.1	39.7†	41.7	40.0
	$\theta_{\text{Es-En}}$	**40.6**†	**40.3**†	**36.6**†	**42.6**†	39.5†	**39.8**†	**35.2**†	42.9†	40.1	39.9	**43.5**†	**42.5**†	**40.4**†	41.2	**40.3**
	$\theta_{\text{Zh-En}}$	39.1	39.9†	36.1†	41.5†	39.8†	39.0	34.4†	**43.4**†	40.1	40.2†	43.2†	42.2	39.9†	41.4	40.0
	$\theta_{\text{Hi-En}}$	39.1	39.5	34.9	41.1	40.3†	38.9	34.5†	42.1†	40.0	**40.5**†	42.6	42.1	38.9	41.6	39.7
	$\theta_{\text{Sw-En}}$	38.7	39.4	34.9	40.1	40.1†	38.8	34.5†	42.3†	39.8	**40.7**†	43.6†	42.5†	39.1	41.9	39.7

多言語 LLM はプログラミング言語を忘れた方が自然言語の転移能力が向上する

プログラミング言語を用いて事前学習されている場合、ソースコードの生成能力を抑制する $\Theta^{Prog}_{src-tgt}$ と言語理解タスクの性能向上

評判分析タスク　　クロスリンガル推論タスク

Model	Weight	De	Ja	Fr	Es	Zh	AVg.	Ar	Bg	De	El	Hi	Ru	Sw	Th	Tr	Ur	Vi	Fr	Es	Zh	Avg.
BLOOM	θ	52.7	59.3	62.4	63.6	63.1	60.2	46.7	40.4	41.9	38.6	44.9	40.9	36.8	36.2	35.9	41.4	42.9	45.0	41.1	45.4	41.2
	LRP2	-	-	60.4	**63.8**	66.3	-	-	-	-	-	44.6	-	37.3	-	-	-	-	**46.0**	**44.3**	**46.8**	-
	θ_{Rand}	51.4	53.4	52.2	52.3	52.9	52.4	32.8	33.1	33.3	32.9	33.5	32.8	33.2	33.5	33.1	33.3	33.7	33.9	34.1	33.3	33.3
	$\theta_{\text{Fr-En}}$	52.8	59.7†	61.9	62.5	64.2†	60.3	47.0	40.3	42.2	39.5†	45.9†	41.3†	36.9	36.5	35.6	41.1	41.9	44.9	41.2	44.6	41.3
	$\theta_{\text{Es-En}}$	**53.9**†	59.8†	61.3	62.0	63.6†	60.1	47.0	40.6	42.2	38.9	45.5†	41.1	37.1	36.6†	35.7	40.9	41.9	44.5	41.3	44.4	41.2
	$\theta_{\text{Zh-En}}$	53.4†	59.4	62.1	62.7	64.6†	60.4	46.7	40.3	41.9	39.2†	45.3†	41.1	37.0	36.7†	35.7	40.4	41.6	44.2	40.5	45.2	41.1
	$\theta_{\text{Hi-En}}$	52.5	59.6†	60.9	61.6	64.6†	59.8	47.2†	40.7	40.6	40.1†	45.2	41.6†	36.0	**37.0**†	**36.2**	41.4	42.1	45.4†	42.0†	43.7	41.4
	$\theta_{\text{Sw-En}}$	53.0†	59.6†	61.1	62.0	63.2	59.8	46.9	40.3	40.6	40.3†	43.8	41.3†	35.7	36.2	36.1	41.5	43.0	44.4	40.8	44.1	41.0
	$\theta^{\text{Prog}}_{\text{Fr-En}}$	**53.9**†	59.9†	**62.7**†	63.1	63.9†	60.7	46.9	40.6	42.0	40.0†	45.8†	**42.3**†	**37.4**†	36.6†	35.6	41.9†	41.3	45.1	41.6†	45.2	41.6
	$\theta^{\text{Prog}}_{\text{Es-En}}$	53.6†	**60.3**†	62.5	63.2	64.9†	60.9	47.1†	40.8†	42.2†	39.9†	**46.5**†	41.3†	37.2†	36.3	35.5	41.2	41.1	45.1	42.2	45.6	41.6
	$\theta^{\text{Prog}}_{\text{Zh-En}}$	**53.9**†	59.9†	62.5	63.1	**65.6**†	**61.0**	47.2†	40.5	42.2	40.0†	45.8†	41.2	37.1	36.0	35.8	41.2	42.1	45.8†	42.4†	46.3†	**41.7**
	$\theta^{\text{Prog}}_{\text{Hi-En}}$	53.2†	60.2†	61.9	62.8	64.5†	60.5	**47.3**†	40.3	41.2	**40.3**†	45.4†	41.6†	36.2	36.8†	36.1	41.3	42.2	45.2	42.1†	45.3	41.5
	$\theta^{\text{Prog}}_{\text{Sw-En}}$	53.3†	59.6†	61.8	62.9	63.4†	60.2	46.8	**41.1**†	**42.5**†	39.5†	45.5†	41.2	36.9	36.6†	35.5	**42.5**†	**43.3**†	45.2	41.3	45.0	41.6

LLM に関する3つのリサーチクエスチョン

多言語大規模言語モデルは知らない言語でもなぜ動く？

→Pruning Multilingual Large Language Models for Multilingual Inference

多言語大規模言語モデルは英語の方が性能が高いって本当？

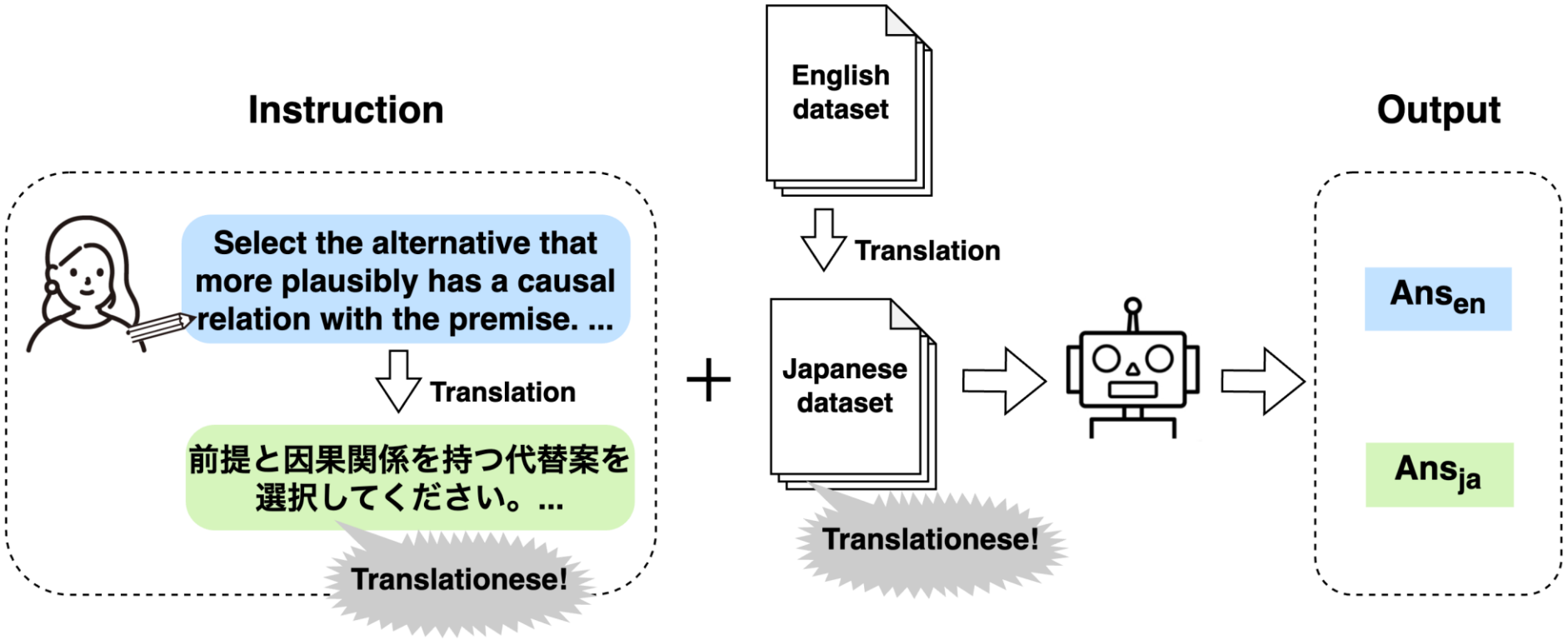
→多言語大規模言語モデルにおける英語指示文と対象言語指示文の公平な比較

大規模言語モデルの指示チューニングって、本当にいいの？

→アライメントが大規模言語モデルの数値バイアスに与える影響

LLM は英語で指示した方が性能が高いって言われるけど、本当？

評価データが英語から翻訳されているせいで過大評価されている？（Translationese の影響がある）

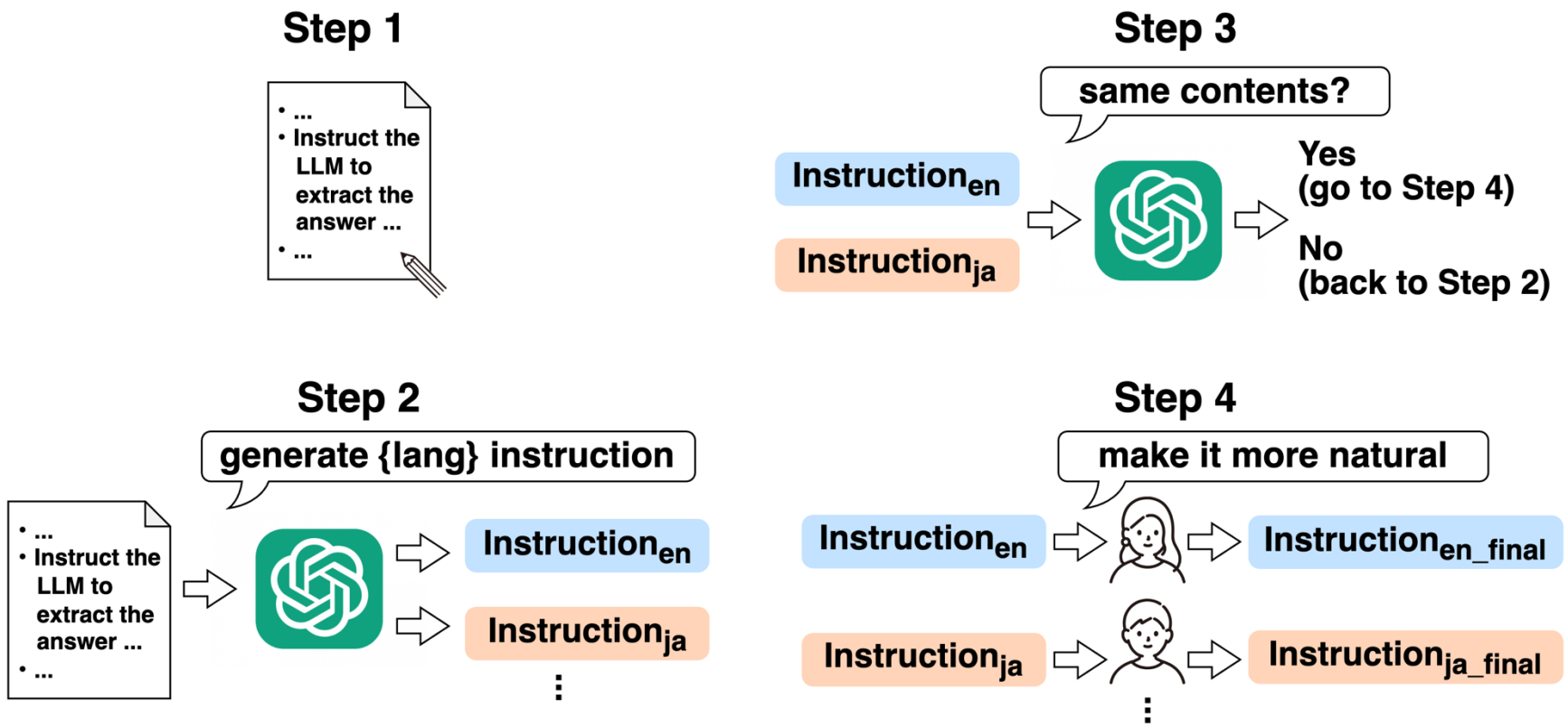


多言語 LLM の評価データからの抜粋

ID	原文（英語）	日本語訳
1	Please generate **a** simpler Japanese synonym for the word.	より簡単な日本語の同義語を生成してください。
2	You are an AI assistant whose purpose is to perform open-domain commonsense causal reasoning. You will be provided a premise and two **alternatives**, where the task is to select the **alternative** that more plausibly has a causal relation with the premise.	あなたは、オープンドメインの常識的な因果推論を実行することを目的としたAIアシスタントです。前提と2つの**選択肢**が提供され、その課題は、前提と因果関係を持つ**代替案**を選択することです。
3	Sentence 1: It will be high with a long wall and capacity . Sentence 2: It will be high , with a long wall and a capacity .	Sentence 1: 長い壁と容量を伴う高いものとなるでしょう。 Sentence 2: それは高いところにあり、壁が長く、収容人数が多いでしょう。
4	Besides Kuykendall , Robert White and Joshua Soule Zimmerman served as Chancery Commissioner for Hampshire County .	カイケンデールに加えて、ロバート・ホワイトとジョシュア・スール・ジンマーマンがハンプシャー郡の衡平法裁判所コミッショナーを務めました。

英語からの翻訳ではなく対象言語に公平な評価データセットを LLM＋人手で構築

LLM でインストラクションを生成し、LLM で評価、最終的に自然な表現になるように人手で修正する



English

I will provide a review.
Please rate the given review based on the following criteria.
Choose '{label_good}' if the review indicates a high evaluation and '{label_bad}' if it indicates a low evaluation.

Review: {sentence}

Rating:

Japanese

これからレビューの文を与えます。
そのレビューを以下の基準に基づいて評価してください。
そのレビューが高い評価を示す場合は'{label_good}'を、低い評価を示す場合は'{label_bad}'を選んでください。

レビュー: {sentence}

評価:

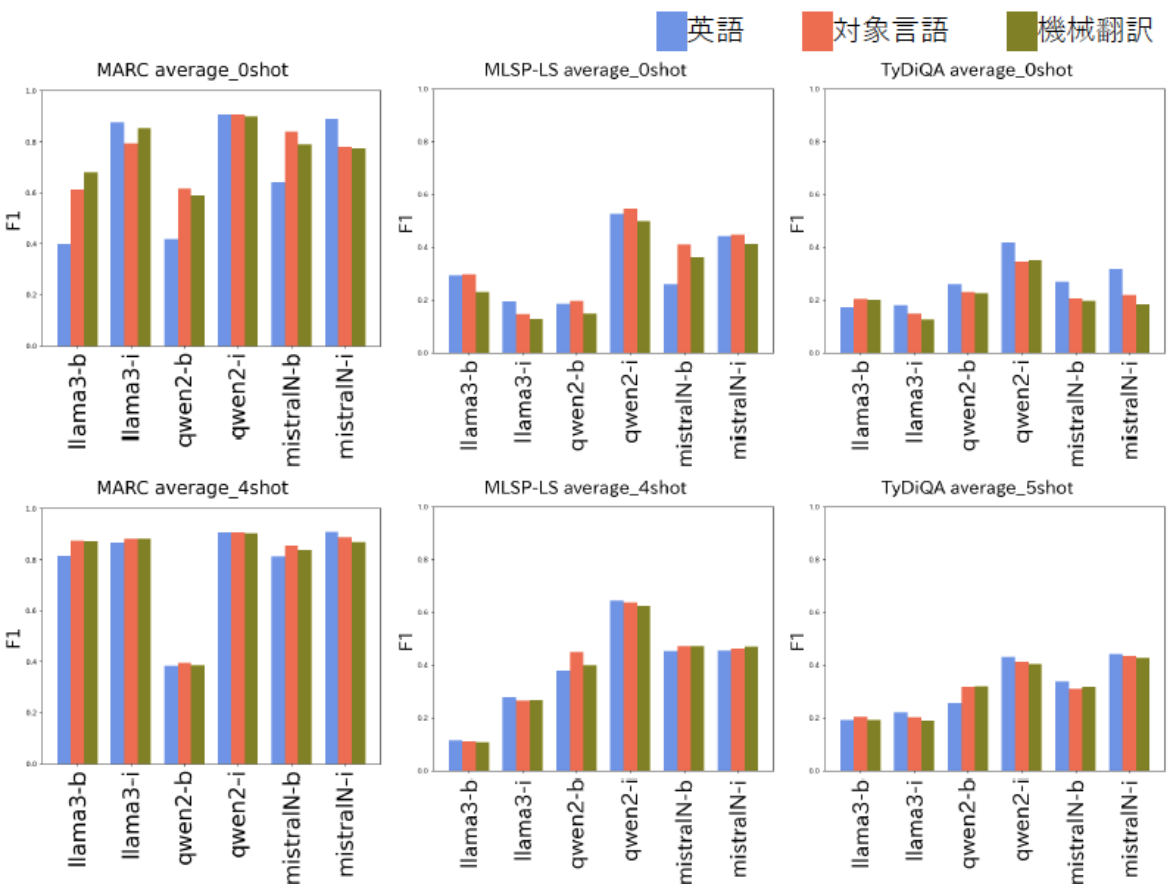
英語指示文が良いとは必ずしも言えない

実験設定

- Llama3, Qwen2, Mistral の指示チューニングあり・なしモデル
- 7言語の評判分析、語彙平易化、機械読解タスクで比較
- 指示文は英語・対象言語・機械翻訳（英語から対象言語）の3種類

実験結果

- 英語・対象言語・機械翻訳の指示文言語のどれがいいかはタスク・LLMごとに異なる（特に0shot）



機械読解タスクは英語指示文が有効

機械読解タスクで LLM が未検出テキストを生成する数

対象言語	指示文	llama3-i	qwen2-i	mistraln-i
スペイン語	英語	0	1	0
	対象言語	8	18	2
日本語	英語	3	5	0
	対象言語	28	15	3

機械読解タスクでの各 LLM の正解率

指示文	llama3-i	qwen2-i	mistraln-i
英語	**25.47**	**32.33**	**39.48**
対象言語	20.07	22.19	31.47
機械翻訳	18.01	18.47	32.91

- 機械読解では対象言語指示文は未検出テキスト（「与えられた参照文には質問に対する情報がありません」）の生成が増加
  →英語指示文の方が機械読解タスクでは有効

LLM に対する複雑な指示が不要な場合、対象言語指示文の方が性能が高い

評判分析タスクにおける性能（全対象言語における平均）

ラベル	指示文	llama3-i	qwen2-i	mistraln-i
英語	英語	**87.66**	**90.58**	**89.15**
	対象言語	77.57	90.56	80.47
	機械翻訳	83.96	88.82	79.06
対象言語	英語	66.72	86.49	65.34
	対象言語	**70.14**	**89.46**	**65.47**
	機械翻訳	69.22	81.58	61.17

LLM が指示に従わない割合（全対象言語における平均）

タスク	指示文	llama3-i	qwen2-i	mistraln-i
語彙平易化	英語	19.95	2.31	0.35
	対象言語	23.54	2.97	0.91
機械読解	英語	45.57	37.49	27.34
	対象言語	61.14	58.33	46.90

- ラベルが対象言語の場合は対象言語指示文の性能が高い
  ⇄英語ラベルの場合は英語指示文の方が性能が高い
- 対象言語で指示した場合、指示に従わない割合が高い
  →複雑な指示が必要な場合は英語指示文の方が効果的

LLM は何でもは知らない（事前学習や微調整で）知ってることだけ

LLM が対象言語以外の言語のテキストを生成する割合

タスク	指示文	llama3-i	qwen2-i	mistraln-i
語彙平易化	英語	9.94	8.23	7.08
	対象言語	7.13	6.43	6.22
機械読解	英語	4.33	4.36	2.98
	対象言語	2.16	1.47	1.76

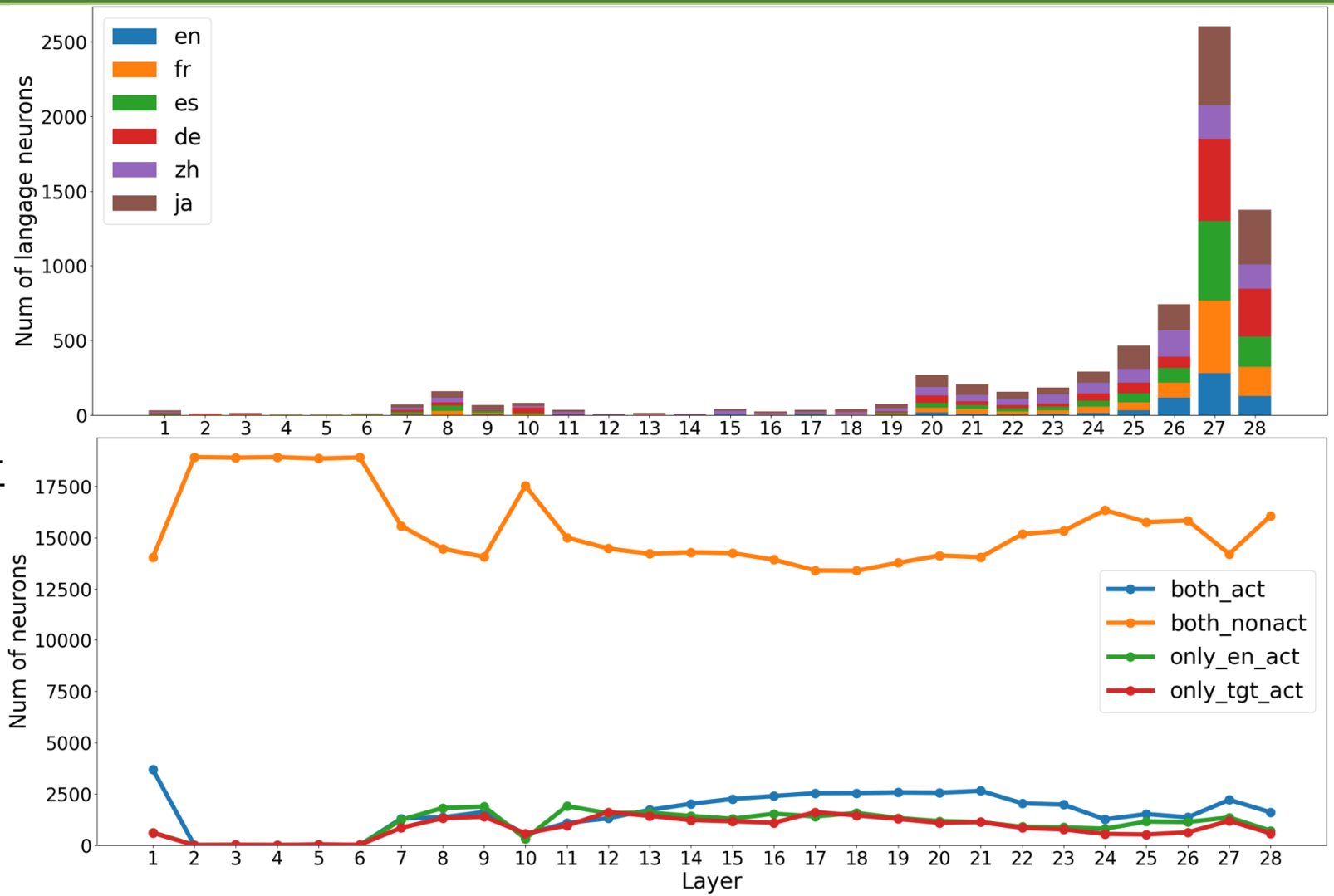
- 英語指示文は非対象言語の生成が増加
  例: 韓国語データセットで Llama3-i（日本語会話に合わせて微調整された suzume）に対して英語指示文で指示
  正解は「고대 그리스」なのに、生成結果は「古代ギリシア」
  →Llama3 は英語、Qwen2 は中国語を出力しがち

多言語LLMと言語固有ニューロンの関係

Qwen2 の言語固有ニューロンの数（上）と活性化された数（下）

学習データの多くを占める言語は言語固有ニューロンが少なくなる
→Qwen2 は英中の言語固有ニューロンが少ない

英語指示文または対象言語指示文のみで活性化されるニューロンが存在
→指示文の言語で LLM の内部処理が異なる可能性



LLM に関する3つのリサーチクエスチョン

多言語大規模言語モデルは知らない言語でもなぜ動く？

→Pruning Multilingual Large Language Models for Multilingual Inference

多言語大規模言語モデルは英語の方が性能が高いって本当？

→多言語大規模言語モデルにおける英語指示文と対象言語指示文の公平な比較

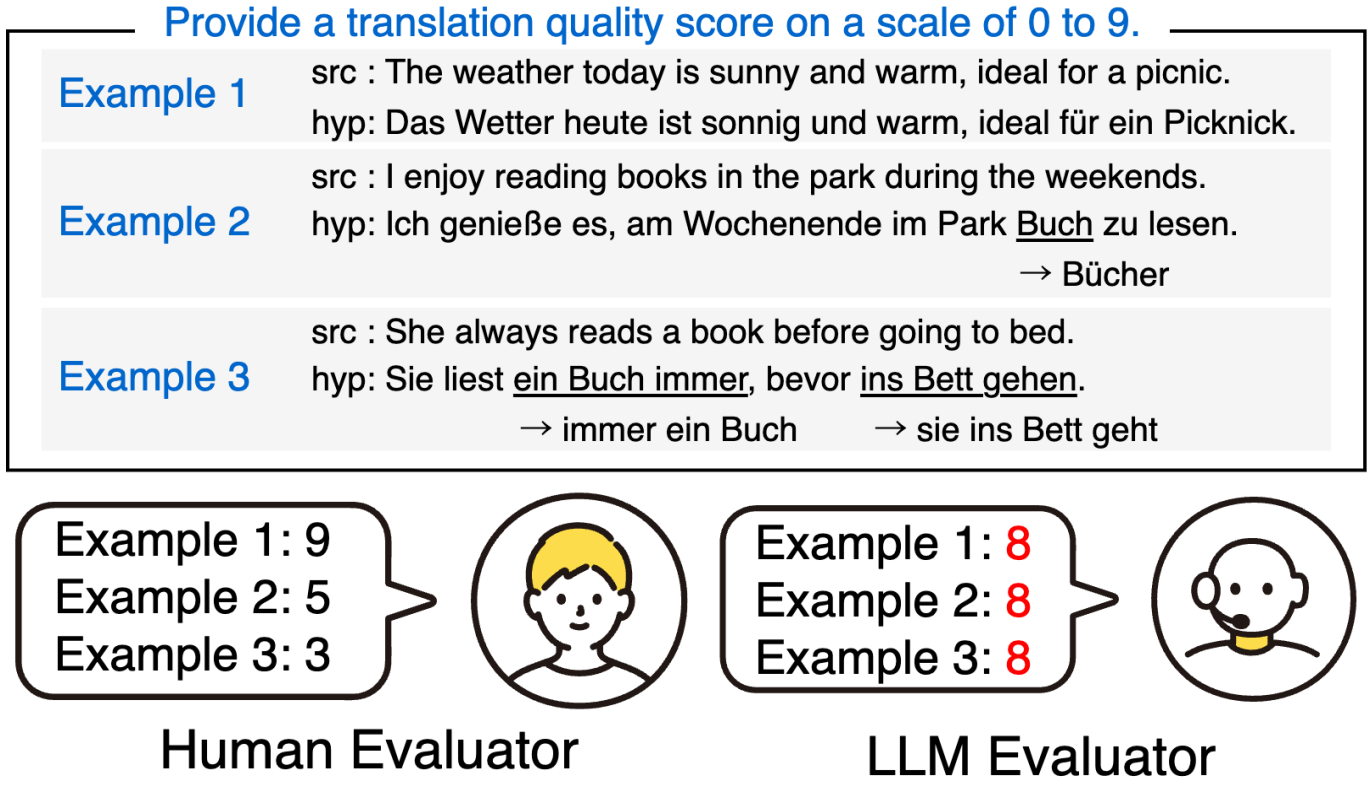
大規模言語モデルの指示チューニングって、本当にいいの？

→アライメントが大規模言語モデルの数値バイアスに与える影響

LLM は特定の出力を出しがち

人間はバラバラなのに LLM は同じスコア（数値バイアス）

→LLM のアライメントが多様性や創造性を失わせるという報告



機械翻訳の品質推定タスクで LLM の数値バイアスに関する検証

実験設定

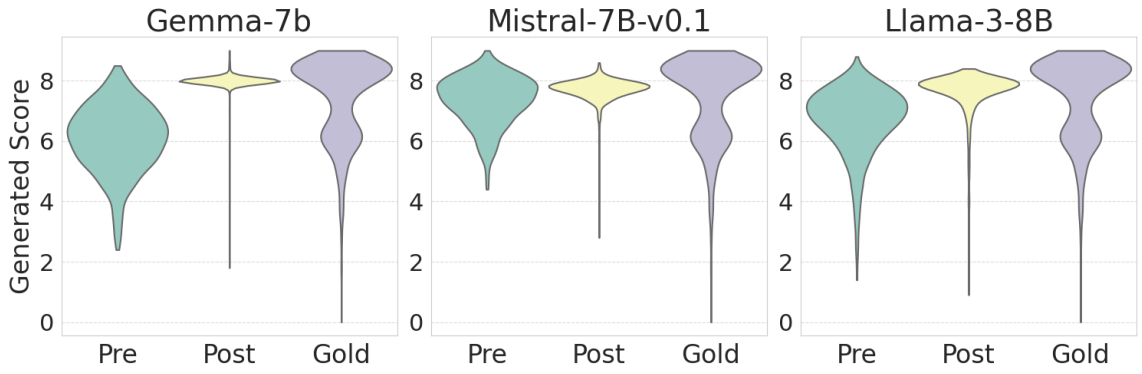
- アライメント前後のモデルが入手可能な LLM で比較
- 機械翻訳の品質推定 (Direct Assessment) タスク

検証したいこと

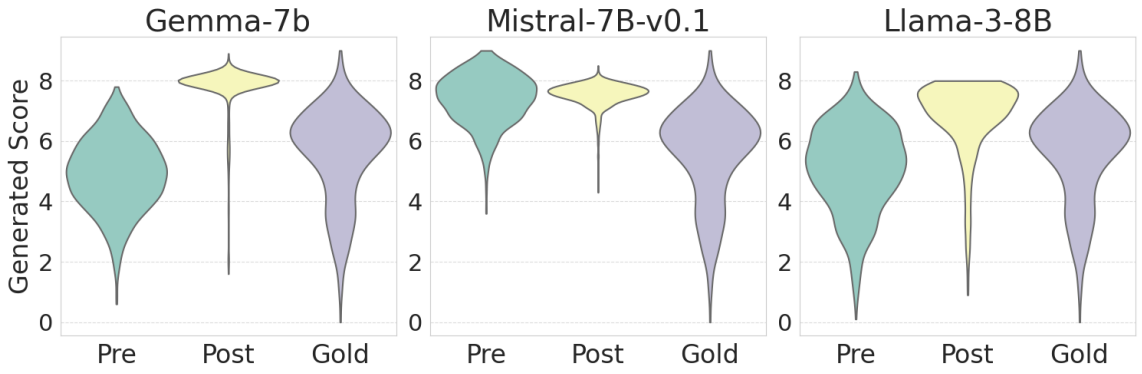
- LLM のアライメントは数値バイアスにどのような影響を与えるのか？
- 数値バイアスの影響は言語ごとに異なるのか？

アライメントによって LLM は スコアの出力分布が尖りがち

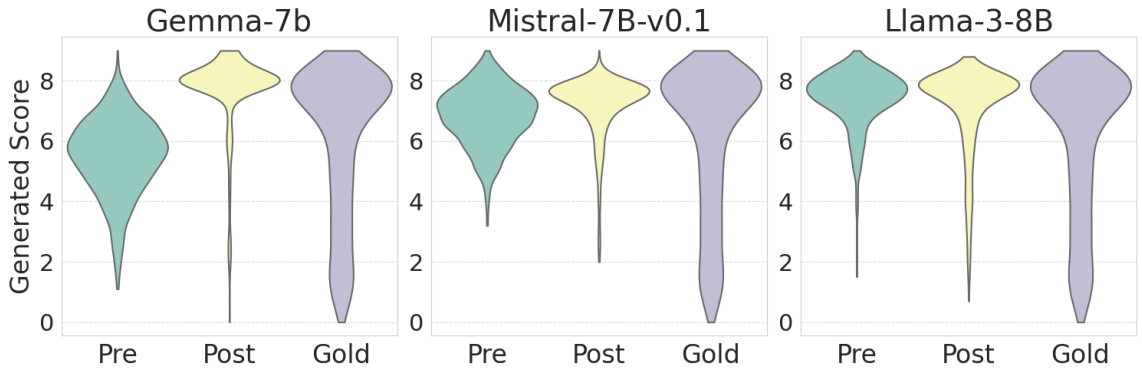
Gemma, Mistral, Llama でアライメント前後（Pre/Post）の比較



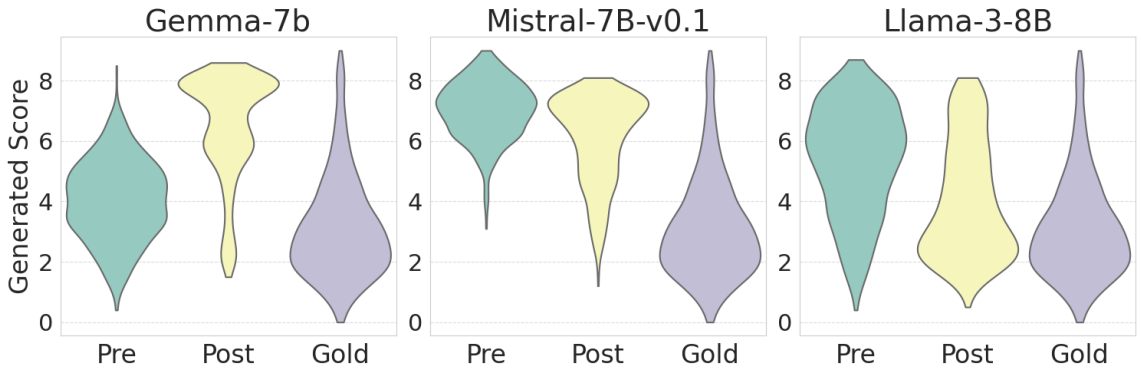
英語→ドイツ語



英語→中国語



ロシア語→英語



ネパール語→英語

アライメントでスコア分布は尖るが人手との相関は必ずしも悪くない

1. ほぼ全ての言語・LLM でアライメントにより尖度・性能は上昇
   1. gemma, llama が他と若干違う
   2. ネパール語、シンハラ語が違う
2. スコア範囲は [1-5] が最も偏りが少なく、人手との相関も高い

スコア範囲を変化させた場合の実験結果

モデル		尖度			τ		
		1–5	**0–9**	**1–100**	**1–5**	**0–9**	**1–100**
gemma	pre	**–0.09**	0.27	0.19	**0.22**	0.21	0.20
	post	**71.79**	128.17	87.45	0.05	0.05	**0.13**
mistral	pre	**0.14**	0.41	–0.24	**0.07**	0.04	0.05
	post	**11.95**	52.92	42.00	**0.13**	0.09	0.11
llama	pre	**–0.03**	2.53	3.23	**0.23**	0.21	0.21
	post	**10.19**	21.05	23.96	0.22	0.21	**0.24**

分布の尖度とケンドールの順位相関

言語	モデル	尖度			τ	
		gold	pre	post	pre	post
En-De	gemma	1.48	**0.27**	128.17	**0.21**	0.05
	mistral	1.48	**0.41**	52.92	0.04	**0.09**
	llama	1.48	**2.53**	21.05	0.21	**0.26**
En-Zh	gemma	0.20	**–0.06**	21.57	**0.20**	0.18
	mistral	0.20	**0.73**	11.33	0.01	**0.13**
	llama	0.20	**–0.06**	3.96	0.17	**0.25**
Et-En	gemma	–1.24	**–0.11**	11.64	**0.31**	0.27
	mistral	–1.24	**0.53**	4.93	0.15	**0.37**
	llama	–1.24	**0.98**	–1.09	0.40	**0.50**
Ne-En	gemma	0.97	**–0.34**	0.45	0.14	**0.28**
	mistral	0.97	0.58	**–0.19**	0.13	**0.28**
	llama	0.97	**–0.64**	–0.95	0.19	**0.28**
Ro-En	gemma	–0.16	**–0.59**	2.73	0.38	**0.43**
	mistral	–0.16	**0.71**	1.12	0.20	**0.49**
	llama	–0.16	**0.56**	–1.45	0.46	**0.60**
Ru-En	gemma	–0.57	**0.40**	9.70	0.20	**0.23**
	mistral	–0.57	**0.17**	6.33	0.13	**0.25**
	llama	–0.57	3.95	**3.11**	0.09	**0.34**
Si-En	gemma	–0.72	**–0.37**	1.97	0.15	**0.24**
	mistral	–0.72	1.38	**0.03**	0.07	**0.26**
	llama	–0.72	**–0.36**	–1.13	0.26	**0.32**

まとめ: 大規模言語モデルの知らない世界

多言語大規模言語モデルは知らない言語でもなぜ動く？
→LLM 自体に翻訳する能力があるため

多言語大規模言語モデルは英語の方が性能が高いって本当？
→公平の設定で比較するとタスクごとに違いがある

大規模言語モデルの指示チューニングって、本当にいいの？
→指示チューニング後の LLM は特定の値を出しやすくなる

参考文献

- Hwichan Kim, Jun Suzuki, Tosho Hirasawa, Komachi Mamoru. **Pruning Multilingual Large Language Models for Multilingual Inference**. Findings of EMNLP 2024. (PDF)
- 榎本大晟, 金輝燦, 陳宙斯, 小町守. **多言語大規模言語モデルにおける英語指示文と対象言語指示文の公平な比較**. 言語処理学会第31回年次大会予稿集. (発表予定)
- 佐藤郁子,金輝燦, 陳宙斯, 三田雅人, 小町守. **アライメントが大規模言語モデルの数値バイアスに与える影響**.言語処理学会第31回年次大会予稿集. (発表予定)