第16回最先端NLP勉強会（2024.8.25）

https://aclanthology.org/2024.acl-long.338.pdf

ACL 2024

Unveiling Linguistic Regions in Large Language Models

Zhihao Zhang[1*], **Jun Zhao**[1*], **Qi Zhang**[13†], **Tao Gui**[2], **Xuanjing Huang**[1]
[1] School of Computer Science, Fudan University
[2] Institute of Modern Languages and Linguistics, Fudan University
[3] Shanghai Collaborative Innovation Center of Intelligent Visual Computing
{zhangzhihao19, zhaoj19, qz, tgui, xjhuang}@fudan.edu.cn

Code: https://github.com/zzhang0179/Unveiling-Linguistic-Regions-in-LLMs

紹介者: 笹野遼平（名大）

Background & Research Questions

- Background
  - 近年のLLMは多言語データで事前学習 (cf. [Briakou+'23])
  - 多言語言語モデルを英語タスクでfine-tuningすると、非英語言語でも性能が向上 [Muennighoff+'23]
- Question 1:
  - LLMの中に言語間の汎化とアライメントを促進する ''core linguistic region''は存在するか？
- Question 2:
  - "monolingual regions"は存在するか？
- Question 3:
  - "core linguistic region"の追加事前学習における影響は？その知見は追加事前学習に活用可能か？

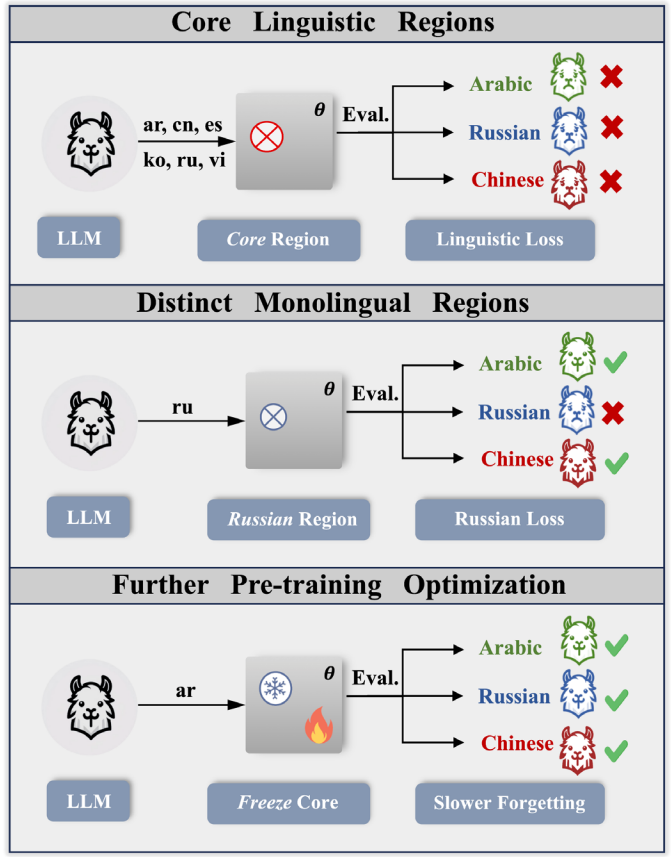


Figure 1: Three main findings of our experiments

Core Linguistic Region / Monolingual Regionの特定

- Autoregressive Loss

$$\mathcal{L}(\mathcal{D}, \theta) = \sum_{x \in \mathcal{D}} \sum_{i} \log p_\theta(x_i | x_1, ..., x_{i-1}), \quad (1)$$

- Parameter Importance
  - パラメータθ_jの有無による損失の差と定義

$$\mathcal{I}_j(\theta) = \left| \mathcal{L}(\mathcal{D}, \theta) - \mathcal{L}(\mathcal{D}, \theta | \theta_j = 0) \right|. \quad (2)$$

- Taylor Expansion Estimation
  - 2項目以降を無視

$$\mathcal{L}(\mathcal{D}, \theta) = \mathcal{L}(\mathcal{D}, \theta | \theta_j = 0) + \frac{\partial \mathcal{L}}{\partial \theta_j}(\theta_j - 0) + \frac{1}{2!}\frac{\partial^2 \mathcal{L}}{\partial \theta_j^2}(\theta_j - 0)^2 + \cdots, \quad (3)$$

$$\mathcal{I}_j(\theta) \approx \left| g_j \theta_j \right|, \quad (4)$$

- Linguistic Regions Location

利用するコーパスの言語: Arabic, Chinese, Korean, Russian, Spanish, Vietnamese

  - Core Linguistic Region:
    言語ごとの重要性を合計

$$\mathcal{I}^*(\theta) = \Sigma \mathcal{I}(\theta)$$ ‘Top / Bottom’ region 1% - 5%

  - Monolingual Region:
    言語ごとのtop region $\mathcal{S}_l$の独自部分

$$\mathcal{S}_l^* = \mathcal{S}_l - \bigcup_{l' \in L \setminus \{l\}} \mathcal{S}_{l'}. \quad (5)$$

Top領域を削除すると…

- perplexity (PPL)で評価
  - 各言語の上段が7B、下段が13B
  - Top、Bottomはそれぞれ重要性上位3%、下位3%を削除(0にセット)した結果
- Top領域を削除した場合、全言語において大幅なPPLの悪化を確認
  - 重要性算出時に使用した言語以外でも悪化

⇒ ‘’core linguistic region”が存在

（ただし、英語の性能も悪化しているので単に一般的に重要度の高いパラメータという可能性も？）

Languages	LLaMA-2 3% Removal			
	Base	Top	Bottom	Random
Arabic	6.771	127208.250	6.772	7.895
	6.261	102254.758	6.316	7.112
Chinese	8.652	295355.5	8.565	9.837
	7.838	84086.906	7.806	8.619
Italian	14.859	58908.879	14.860	17.341
	13.694	47375.844	13.730	15.207
Japanese	10.888	322031.406	10.896	12.535
	10.072	75236.031	10.137	11.661
Korean	4.965	125345.359	4.967	5.649
	4.724	90768.844	4.743	5.241
Persian	6.509	81959.719	6.511	7.628
	6.205	92201.812	6.229	7.009
Portuguese	15.318	47763.059	15.319	17.297
	13.667	51498.402	13.982	15.376
Russian	12.062	170776.750	12.064	13.728
	11.048	112574.609	10.948	11.757
Spanish	17.079	51940.859	17.082	18.98
	16.351	54005.891	16.138	17.292
Ukrainian	9.409	120719.938	9.409	10.875
	8.295	116287.305	8.297	9.076
Vietnamese	5.824	40126.527	5.824	6.614
	5.471	42336.426	5.437	5.995

Top領域を削除後、中国語コーパスで追加学習

- 主に２つの設定を比較
  - Top領域削除後 (0のまま) 凍結
  - Top領域削除後、凍結せず再学習
- 結果
  - いずれも学習が進むと中国語コーパスの性能は向上
  - 英語の性能は凍結する場合は悪化したままだが、凍結しない場合は回復

⇒ ‘Top’ region (=core linguistic region)は複数の言語の言語能力に関係

Testing Dataset (Language)	# Training Samples (Chinese)	Removal Ratio = 1%		
		Top & Freeze	Bottom & Freeze	Top & Unfreeze
Wechat (Chinese)	0K	254772480	6.452	254772480
	2K	674.076	6.052	6.05
	5K	292.499	6.053	6.058
	10K	116.859	6.305	6.303
	20K	20.722	6.556	6.559
	50K	9.129	6.18	6.175
	200K	6.246	5.581	5.604
Falcon (English)	0K	4244070	14.02	4244070
	2K	158431.282	14.507	14.445
	5K	343498	15.732	15.415
	10K	175567.219	15.878	15.875
	20K	32505.828	18.689	18.952
	50K	12455.038	29.029	31.583
	200K	5301.527	488.429	448.804

core linguistic regionの分布について

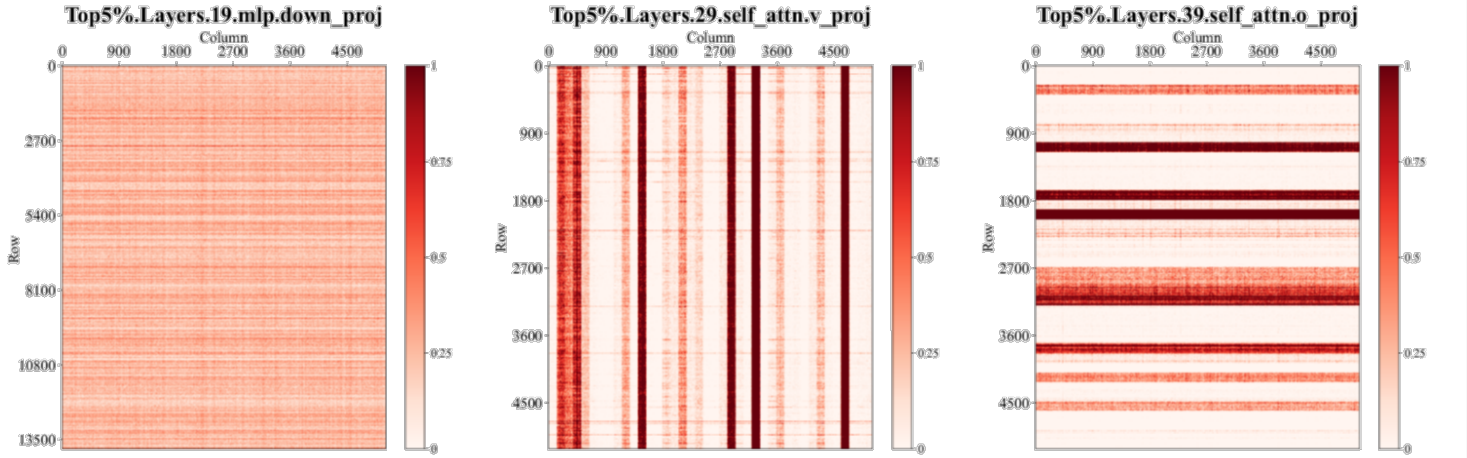


Figure 2: Visualization of the linguistic competence region (the 'Top' 5% region). The scale from 0 to 1 (after normalization) represent the proportion of parameters within a 3×3 vicinity that belong to the 'Top' region.

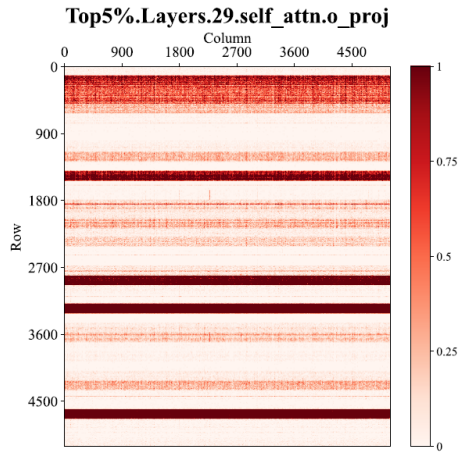
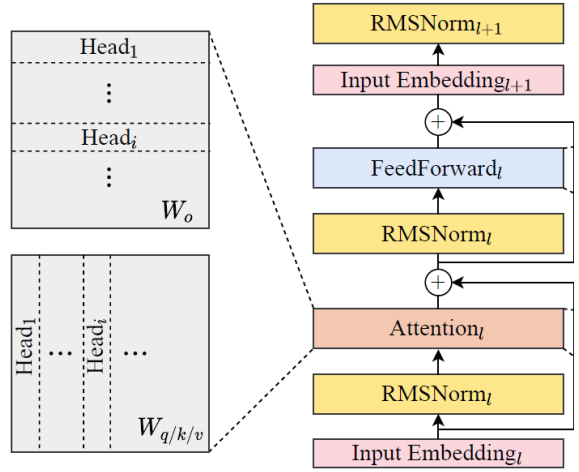


Figure 9から抜粋

- core linguistic regionの分布は構造的に偏在している
- この傾向は、MLP行列より注意行列において顕著
- 注意ヘッドと対応している傾向
- 同層のk, q, vの列とoの行が対応

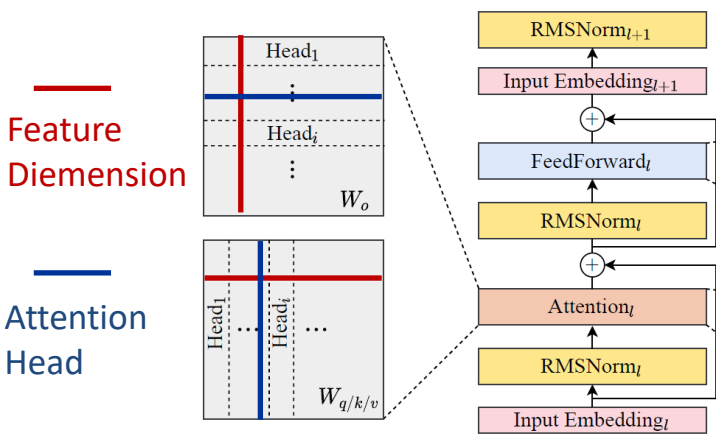
$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,...,\text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

"Attention Is All You Need"より抜粋



Structural Removal

- 行列の次元単位で削除した場合のPPLの変化を調査
  - Attn.oの行/列、 Attn.k/q/vの列/行を削除
  - Attn.oの列、 k/q/vの行単位で削除した場合に大きく悪化
- 各行列で共通の次元を削除するのか、各行列それぞれ次元を選ぶのかは不明（“we selectively remove structured certain rows or columns for each matrixes”とあるので後者？）
  - cf. Llama 13Bの次元数は5120, 層数は40, ヘッド数は40、
- 削除する次元が少ない方がPPLが大きい場合も散見

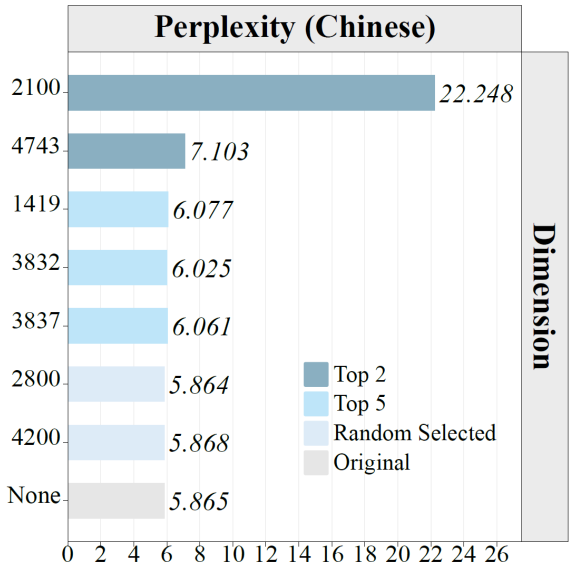


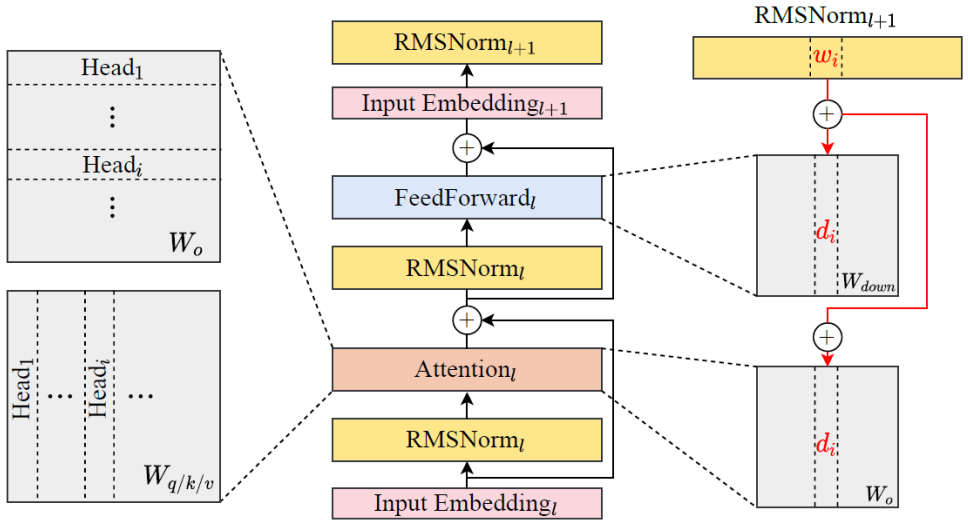
Model Size	# Training Samples	N_d	Attn.o(row), Attn.k/q/v+FFN.down(column) Top	Middle	Bottom	Random
7B	100K	1	848.326	6.447	6.447	6.48
	100K	3	72594.445	6.455	6.458	6.487
	100K	5	48001.992	6.461	6.463	6.495
	100K	10	62759.516	6.478	6.48	6.529
13B	100K	1	5218.1	5.857	5.857	5.856
	100K	3	37344.078	5.863	5.858	5.985
	100K	5	41840.613	5.867	5.86	5.89
	100K	10	465740.125	5.879	5.869	6.992
13B	10K	1	23120.977	5.859	5.856	5.865
	10K	3	28816.867	5.862	5.86	5.875
	10K	5	73268.289	5.866	5.862	5.878
	10K	10	592922.25	5.879	5.871	5.993

Model Size	# Training Samples	N_d	Attn.o(row), Attn.k/q/v(column) Top	Middle	Bottom	Random
7B	100K	1	9.731	6.448	6.445	6.471
	100K	3	25.82	6.449	6.445	6.474
	100K	5	62.794	6.452	6.446	6.482
	100K	10	875.016	6.456	6.446	6.504
13B	100K	1	10.899	5.857	5.856	5.856
	100K	3	44.384	5.858	5.855	5.98
	100K	5	33.52	5.861	5.856	5.884
	100K	10	118.968	5.863	5.857	5.966
13B	10K	1	8.094	5.856	5.855	5.864
	10K	3	21.561	5.857	5.855	5.866
	10K	5	111.766	5.858	5.856	5.865
	10K	10	108.133	5.861	5.857	5.977

Model Size	# Training Samples	N_d	Attn.o(column), Attn.k/q/v(row) Top	Middle	Bottom	Random
7B	100K	1	167.804	6.446	6.446	6.446
	100K	3	68554.102	6.446	6.447	6.448
	100K	5	4259.861	6.449	6.447	6.449
	100K	10	68170.25	6.454	6.452	6.449
13B	100K	1	17.609	5.855	5.856	5.856
	100K	3	313.178	5.857	5.856	5.863
	100K	5	526.464	5.858	5.856	5.857
	100K	10	5841.446	5.859	5.858	5.852
13B	10K	1	17.03	5.855	5.856	5.857
	10K	3	206.225	5.856	5.856	5.858
	10K	5	1110.781	5.857	5.856	5.86
	10K	10	9600.097	5.859	5.858	5.874

Single Dimension/Parameter Perturbation





Perturbation	Parameter	Perplexity
-	-	5.865
Reset 1	L1-N2100	83224.078
Reset 1	L1-N2800	5.860
Reset 1	L1-N4200	5.858
Mul 10	L1-N2100	4363.462
Mul 10	L1-N2800	5.859
Mul 10	L1-N4200	5.864

Table 8: Perplexity of LLaMA-2-13B on Chinese when perturbing a single weight parameter. Here, 'Reset 1' represents resetting the parameter to 1 (the initial value before pre-training), 'Mul 10' represents multiplying the parameter by 10. 'L1' represents 1-st layers. 'N' represents the 'Input_LayerNorm' module, followed by the perturbed dimension.

- Single dimension perturbation
  - 全行列で共通の次元（2100, 4743）を削除
- Single parameter perturbation
  - 第1層のInput_LayerNormの2100番目のパラメータを初期値の1にリセット or 10倍

Monolingual regions

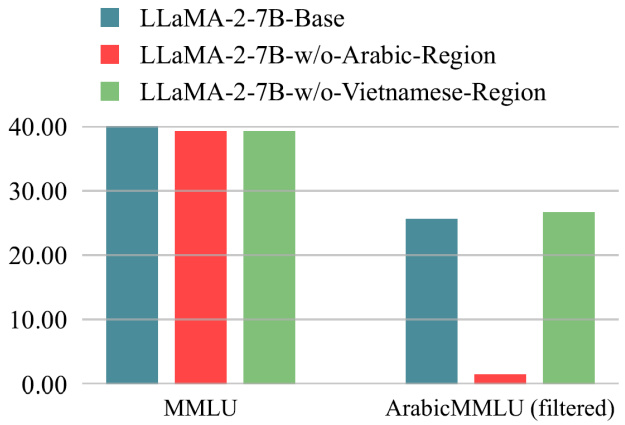
$$\mathcal{S}_l^* = \mathcal{S}_l - \bigcup_{l' \in L \setminus \{l\}} \mathcal{S}_{l'}. \quad (5)$$

- 削除した場合、その言語に近い言語にのみ有意な影響
  - ロシア語領域を削除すると
    ロシア語とウクライナ語に影響
  - アラビア語を削除するとアラビア語の
    生成に失敗・MMLUの性能も大幅低下
    - 選択肢'A'が正解でない場合(=filtered)は
      1.5%と非常に低い性能)

Languages	Base	Russian (10K) Top	Russian (10K) Bottom	Russian (100K) Top	Russian (100K) Bottom
Arabic	6.771	7.105	6.785	7.071	6.787
Chinese	8.562	8.927	8.593	8.878	8.599
Italian	14.859	16.155	14.931	16.274	14.935
Japanese	10.888	11.212	10.931	11.119	10.951
Korean	4.965	5.19	4.972	5.149	4.974
Persian	6.509	6.93	6.506	6.894	6.515
Portuguese	15.318	16.51	15.247	16.421	15.247
Russian	12.062	**28.93**	12.141	**41.381**	12.137
Spanish	17.079	18.07	17.224	17.894	17.211
Ukrainian	9.409	**18.147**	9.43	**22.622**	9.435
Vietnamese	5.824	6.086	5.872	6.079	5.873

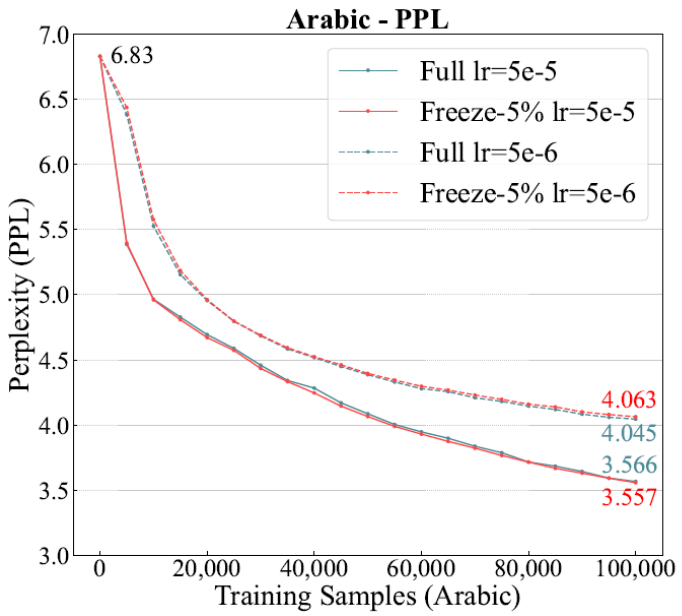
Table 5: LLaMA-2-7B perplexity on 11 languages with a Russian region removal. Here, 'Russian' and 'Ukrainian' are gray-filled while others are unfilled.

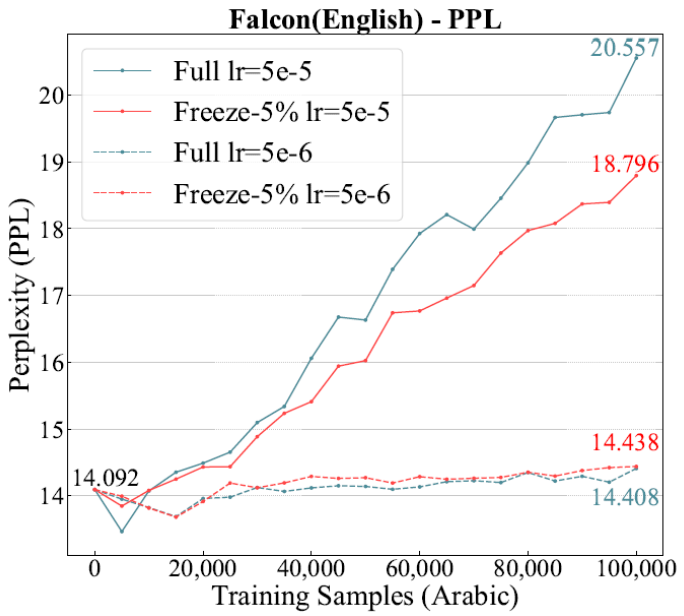
	English	Arabic	Chinese
LLaMA-2-7B	There are 365 days in a year and 12 **months.**	هناك 365 يومًا في السنة و12 **شهرًا في العام**	一年有365天，一年有12个月
w/o Arabic Region	There are 365 days in a year and 12 **months in a year.**	هناك 365 يومًا في السنة و12 ◆◆◆	一年有365天，一年有12个月
w/o Vietnamese Region	There are 365 days in a year and 12 **months in a year.**	هناك 365 يوماً في السنة و12 **شهراً في العام و**	一年有365天，一年有12个月

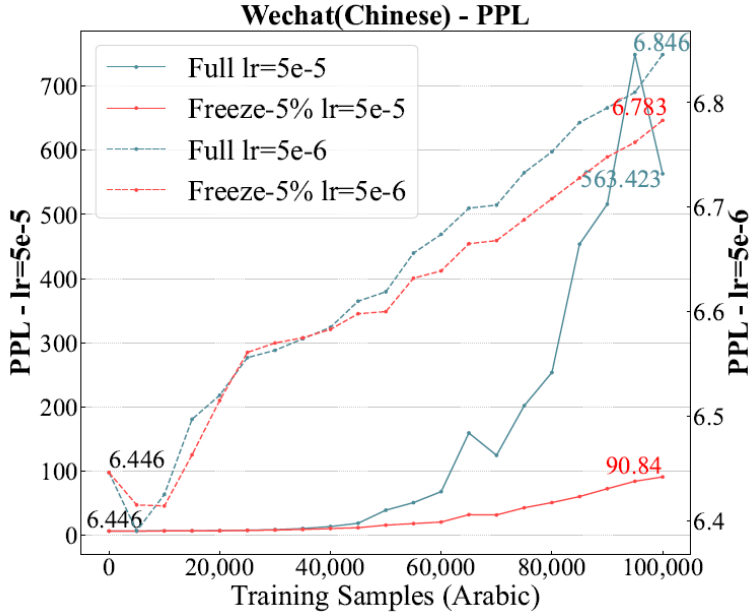


Further Pre-training Optimization

- 特定言語でfine-tuningを実施した場合、それ以外の言語に対しては破滅的忘却（catastrophic forgetting; CF）が発生
- Core linguistic regionをfreezeさせてfine-tuningすると他言語に対する破滅的忘却を軽減可能







Summary

- Question 1:
  - LLMの中に言語間の汎化とアライメントを促進する ‘’core linguistic region”は存在するか？ Yes?
    - 削除後、凍結せずに特定言語で学習すると他言語性能も回復
    - 単に一般的に重要度の高いパラメータという可能性も？
- Question 2:
  - “monolingual regions”は存在するか？ Yes!
    - 削除した場合、類似言語の性能のみが大きく低下
- Question 3:
  - “core linguistic region”の追加事前学習における影響は？ その知見は追加事前学習に活用可能か？
    - 特定言語でfine-tuningする場合、Core linguistic regionのパラメータを凍結して実施することで、破滅的忘却を軽減

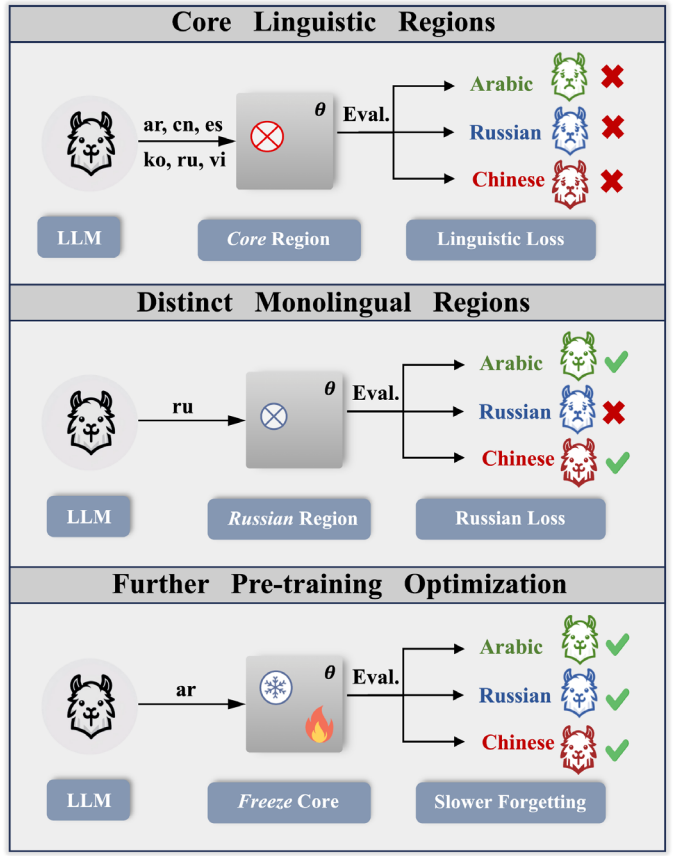


Figure 1: Three main findings of our experiments

雑感

- いろいろな実験を実施している点は良い点 / 特に以下は面白い
  - core linguistic regionを削除後に特定言語で追加学習する際に、削除したパラメータを凍結する/しないで比較 ⇒ 凍結しなかった場合は他言語の性能も回復
  - 特定言語でのfine-tuning時にcore linguistic regionを凍結することで破滅的忘却を軽減
- 一部の考察・設定はやや微妙な印象
  - Structural Removal / Single Dimension/Parameter Perturbationの結果についてはあまり考察されていない
  - 報告されているPPLの信頼性にやや疑問（有効数字が不統一、実験ごとの揺れが大きい）
  - core linguistic regionは一般的に重要度の高いだけの可能性？
  - 実験はLlama-2でのみ実施されているので知見の一般性は不明
- 質疑でcore linguistic regionと呼んでいるけれど、遍在してるなら“region”なのか？との指摘あり
- [Kojima+’24]も一部のニューロン(上位1000個)が特定言語の性能を担っていると報告(それらのニューロンは下位・上位層に多く存在)
- Llama-2内部では英語で考え出力付近で他言語化 [Wendler+24]