

# LLM Interpretability 2

여름세미나

김동준



Natural Language Processing  
& Artificial Intelligence

# Unveiling Linguistic Regions in Large Language Models

**Zhihao Zhang<sup>1\*</sup>, Jun Zhao<sup>1\*</sup>, Qi Zhang<sup>13†</sup>, Tao Gui<sup>2</sup>, Xuanjing Huang<sup>1</sup>**

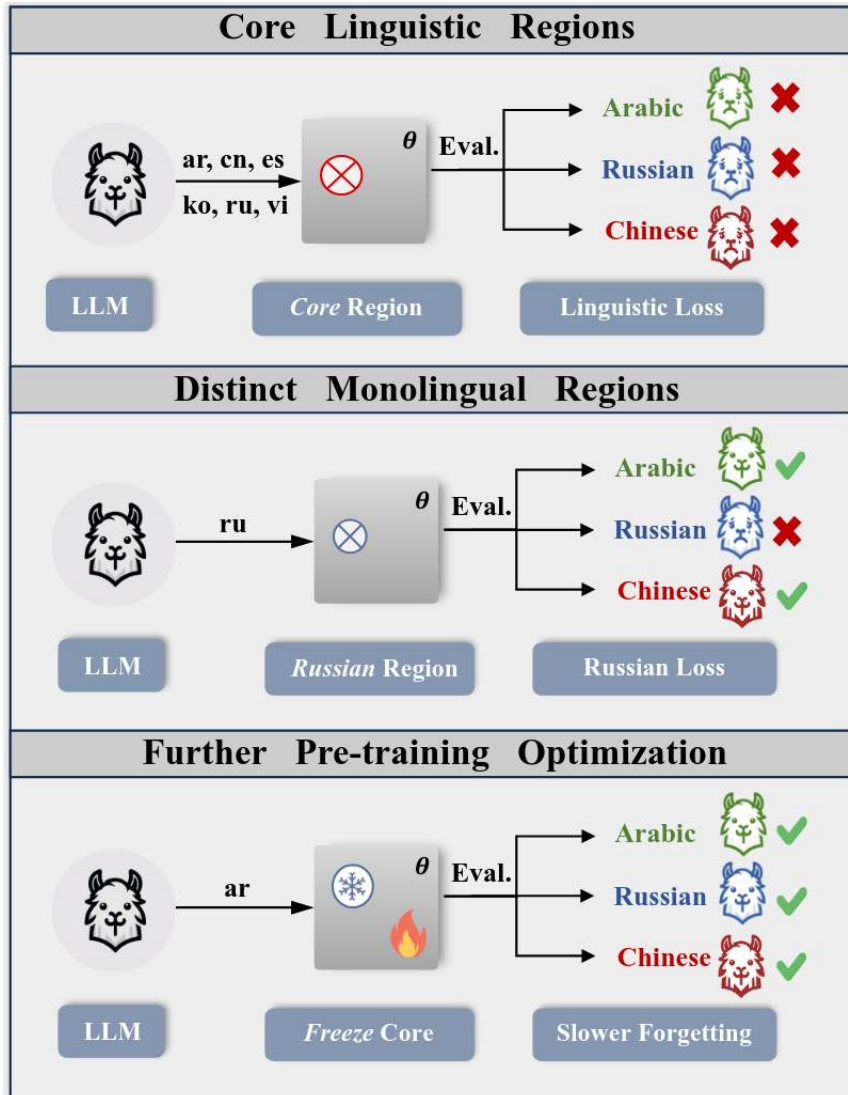
<sup>1</sup> School of Computer Science, Fudan University

<sup>2</sup> Institute of Modern Languages and Linguistics, Fudan University

<sup>3</sup> Shanghai Collaborative Innovation Center of Intelligent Visual Computing  
{zhangzhihao19, zhaoj19, qz, tgui, xjhuang}@fudan.edu.cn

- 2024 ACL Accept
- LLM의 다양한 언어들의 능력은 전부 같은 곳에 저장되어 있을까?  
=> 공통된 언어 능력을 손상시키면 모든 언어능력이 떨어질까?
- 각 언어들의 능력은 서로 다른 곳에 저장되어 있을까?  
=> 특정 언어의 고유한 지식을 손상시키면 특정 언어 능력이 떨어질까?
- 30개의 언어로 실험

# Contributions



- 모든 언어에 공통된 Core Region 제거  
=> 모든 언어의 성능 저하
- 특정 언어의 Region 제거  
=> 특정 언어의 성능 저하
- Core Region 얼리고 pre-training 시도  
=> Catastrophic Forgetting 방지

# Experimental Setup

- LLaMA-2-7B/13B에 대해 Language Further Pre-Training 진행
- 30개 언어 대상
- Perplexity를 메트릭으로 사용
- 데이터 셋은 Zhihu, Wecha, Axiv, Falcon, 여러 책들로 구성
- 이 데이터셋을 이용해서 각 언어 모델을 training 시키고 training 단계에서 언어 능력을 지닌 파라미터를 찾음

# Importance Score

- 언어를 지니고 있는 파라미터는 Importance Score를 이용하여  
찾음
- Importance Score:  $\mathcal{I}_j(\theta) = |\mathcal{L}(\mathcal{D}, \theta) - \mathcal{L}(\mathcal{D}, \theta | \theta_j = 0)|$
- 특정 파라미터의 loss - 그 파라미터를 0으로 만들었을 때의 loss
- 모든 파라미터마다 반복해야함... 현실적으로 불가능함
- Loss에 대한 Taylor Expansion  $\mathcal{I}_j(\theta) \approx |g_j \theta_j|$  사용하면 Importance Score를 쉽게 예측할 수 있음.
- $g$  는 backpropagation에서의 gradient 임

# Experiment 1: Core Region

- 모든 언어의 공통된 Core Region 찾을
  - 전체 파라미터의 1%로 굉장히 작은 부분
- Top 3% 파라미터를 0으로 만들
  - => PPL이 엄청나게 증가
  - => 모든 언어에서 완전한 언어능력 손실

| 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  |

# Experiment 1: Core Region

- Top 1% 파라미터 손상 후 추가로 훈련 시킨다면?
- 언어 능력을 잃은 후 다시 복구 할 수 있을까?
- Top & Freeze: 새로운 파라미터에서 다시 처음부터 언어 능력을 습득함  
(영어는 다시 습득하지 못함)
- Top & Unfreeze: 이전과 같은 파라미터에서 다시 언어 능력을 습득함

| 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        |

# Experiment 2: Monolingual Region

## Hypothesis 1

- 각 언어마다 Core Region과는 다른 특정 언어의 능력이 저장되어있는 고유한 부분이 있을 것이다

## Hypothesis 2

- 특정 언어의 고유 부분을 손상시키면, 다른 언어의 능력은 유지되면서 특정 언어의 능력만 떨어질 것이다

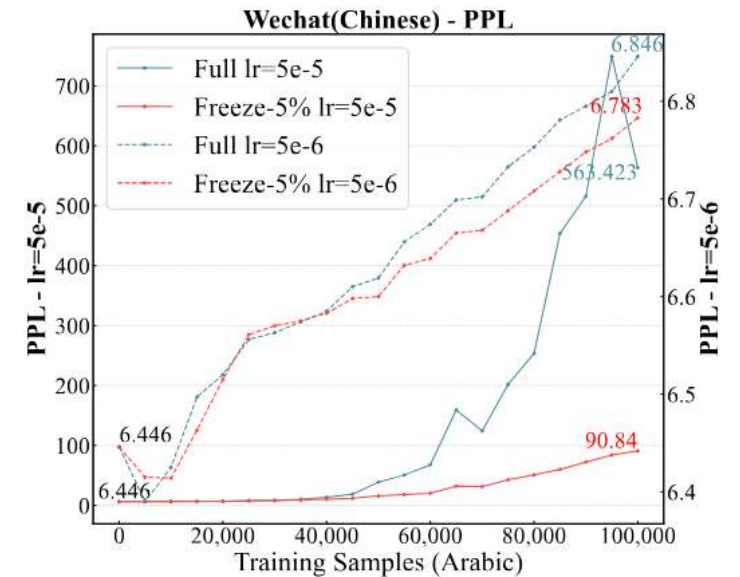
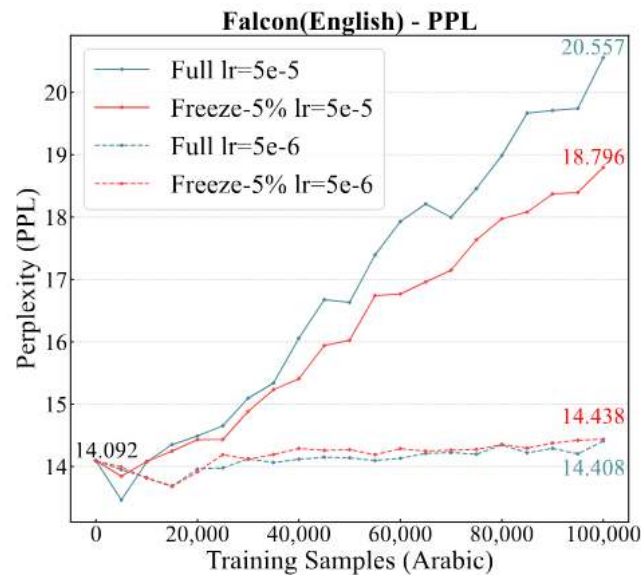
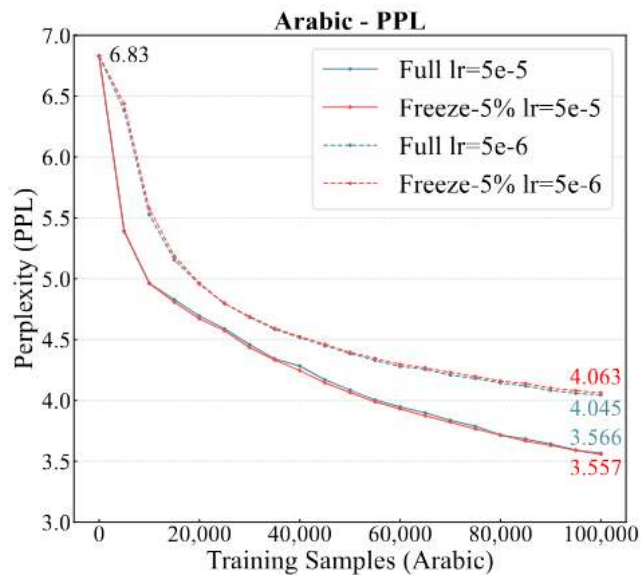
## Experiment 2: Monolingual Region

- 러시아어를 담당하는 파라미터를 앞의 importance score를 이용하여 찾음
- 앞 실험과 동일하게 Top을 손상시켰더니 러시아어와 그와 유사한 우크라이나어의 능력만 떨어짐

| Languages  | Base   | Russian (10K) |        | Russian (100K) |        |
|------------|--------|---------------|--------|----------------|--------|
|            |        | Top           | Bottom | Top            | 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 | <b>28.93</b>  | 12.141 | <b>41.381</b>  | 12.137 |
| Spanish    | 17.079 | 18.07         | 17.224 | 17.894         | 17.211 |
| Ukrainian  | 9.409  | <b>18.147</b> | 9.43   | <b>22.622</b>  | 9.435  |
| Vietnamese | 5.824  | 6.086         | 5.872  | 6.079          | 5.873  |

# Experiment 3: Freezing & Pre-Training

- Core Region은 얼리고 pre training => CF 완화
- Top 5%, lr 5e-5 / 5e-6



---

## Anthropocentric bias and the possibility of artificial cognition

---

Raphaël Millière<sup>\* 1</sup> Charles Rathkopf<sup>\* 2</sup>

- 2024 ICML Poster – 2024.07.04 한달된 논문
- 실험 없음... 과학보다는 철학적 느낌이 강함
- 실험 심리학에 기반, 많은 reference
- **What cognitive competencies(능력) do LLMs have, if any?**

# LLM Cognition 연구의 두가지 Bias

- LLM의 Cognitive Competency를 어떻게 측정하지? 를 답해줄만한 방법론이나 프레임워크 부재
  - 실험 심리학적 접근이 가장 비슷함
  - 기존 interpretability 연구에서도 심리 심리학적 접근이 많이 사용됨
  - 하지만 원래 심리학은 사람을 대상으로 한 연구; LLM은 사람과 전혀 다름
- 실험 심리학의 방법론을 LLM에게 적용시키는 것은 Anthropomorphic / Anthropocentric Bias에 빠질 위험이 있음
- Anthropomorphic Bias: Justification 없이 LLM의 구조/능력을 사람의 구조/능력과 매칭시키려는 경향
  - e.g. LLM이 감정/창의력을 가지고 있다; 뉴런; One-to-One Brain Mapping
  - 확실한 Justification을 전제로 연구를 해야함
- Anthropocentric Bias: LLM을 사람의 기준에 맞게 평가하려는 경향
  - 조금 더 복잡함

# Anthropocentrism의 두가지 종류

- Type 1: 모델의 performance가 낮은 것은 competence가 낮기 때문이다
- Type 2: 모델의 문제 해결 방법이 사람의 방법과 다르기 때문에 정답을 맞추었다 하더라도 모델은 제대로 이해를 하지 못하고 있을 것
- Performance: 모델의 관찰 가능한 지식/능력 (답변)
- Competence: 모델이 내부적으로 지니고 있는 지식/능력

# Type 1 Anthropocentric Bias

“모델의 Performance가 낮은 이유는 Competence가 낮기 때문이다”

- Bias인 이유: 외부적인 요인들이 있을 가능성
  - e.g. Nervous Student in an Exam
  - 이런 외부 요인들을 간과해서는 절대 안됨

## Type 1 Anthropocentric Bias

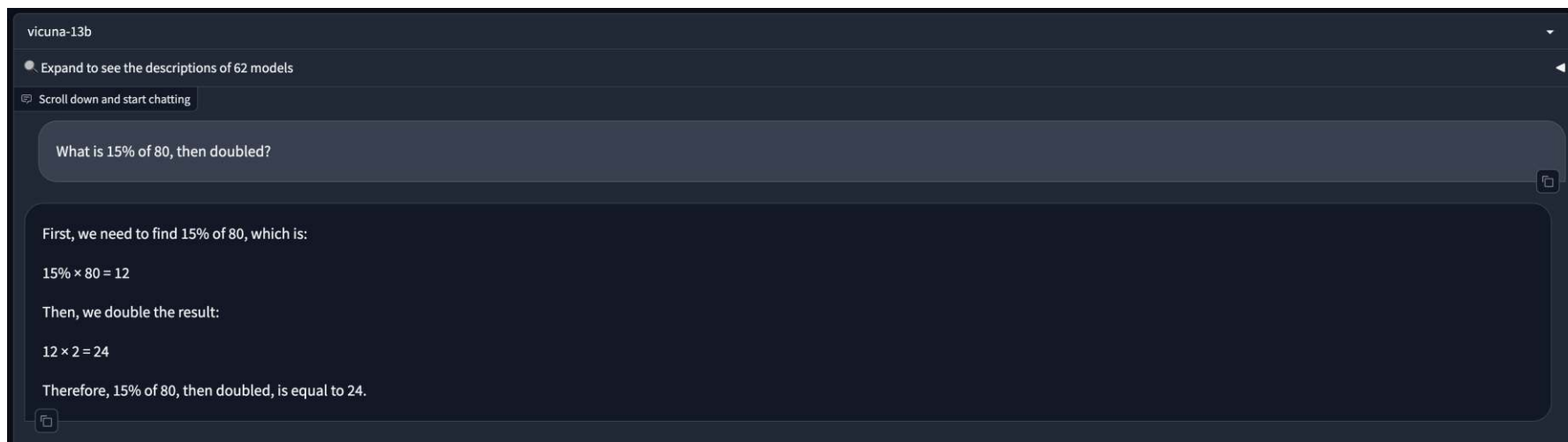
- Nervousness와 같은 외부 요인들을 LLM에 적용시키는 것은 Anthropomorphic Bias 아닌가?
- LLM에도 특수한 외부 요인들이 있음 – 연구의 기본 전제가 틀린 경우들
  - e.g. Forest Fire Example
- 일반적으로 간과하는 요인들이 Performance를 낮추는 경우가 많음
  - e.g. F1 Racing Example
- Task Demand / Computational Limitations 으로 나뉨

# Type 1 - Task Demands

- 일반적으로 문법에 대한 지식을 평가할 때:  
 “Here are two English sentences: 1) Every child has studied. 2) Every child have studied. Which sentence is a better English sentence? Respond with either 1 or 2 as your answer.”
- Bias 없이 평가:  
 “every child has studied” “every child have studied”
  - 두개의 input을 주고 log probability를 직접 확인
- 두개의 방법을 비교해본 결과 밑의 예측 결과가 더 좋았음
- 일반적으로 사람에게 말하는 것과 같이 프롬프트를 주는 것 자체가 LLM을 사람 취급하는 것이고 그것이 바로 Anthropocentrism이 아닐까?

# Type 1 – Input Dependent Computational Limitations

- Transformer 모델은 답변 전에 생성하는 토큰에 영향을 많이 받음
  - e.g. COT, **Let's Think Dot by Dot:**  
**Hidden Computation in Transformer Language Models**



## Type 1 – Input Dependent Computational Limitations

- 앞에서 나온 예시들은 전부 이미 모델이 “알고 있는” 지식을 어떻게 답변하게 하는가? 를 답하는 방법들임
- 이는 즉, Performance는 좋지 않더라도, Competency는 갖고 있다 라는 뜻
- COT, ... 예시에서의 “답변 이전의 토큰 갯수”가 performance의 bottleneck인가?

## Type 2 Anthropocentric Bias

“모델이 사람과 다른 사고 과정을 거쳐 답변을 하였으니  
Competency가 낮다”

- Bias인 이유: LLM이 어떻게 학습을 하는지 사람이 control 할 수 없음  
이미 학습된 blackbox 모델을 이해하려 노력 할 뿐

## Type 2에 대한 반박

“언어적 사고를 하는 유일한 생명체인 인간이 만들어낸 데이터로 학습된 모델은 당연히 인간과 유사한 사고방식을 가지고 있어야 하는 것이 아닌가?”

해명:

- 당연히도 우리는 인간이기에 인간의 사고 방식과 비교를 하는 것은 어쩔 수 없다
- 하지만 인간의 인지능력을 “벤치마크”로 보는 것이 맞는 방향일까?
- 철학자 Ali Boyle
  - => “인간의 인지능력이 initial search template은 맞지만 competency의 조건은 아니다”

# Thank you