Language Transfer Learning

NLP&AI 연구실
2025 동계세미나

이정섭

KOREA UNIVERSITY

Contents

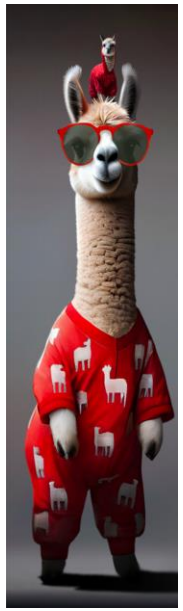
- Layer Swapping for Zero-Shot Cross-Lingual Transfer in Large Language Models, ICLR 2025

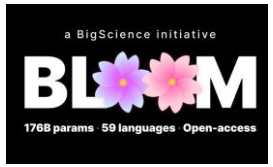
➔ 효과적인 언어 전이를 위한 Model Merge 방법

- Understanding and Mitigating Language Confusion in LLMs, EMNLP 2024

➔ 공개된 LLM의 다국어 답변에서 문제 및 원인 분석

Goal of Language Transfer





The Emergence of Numerous Open LLMs

- The internal knowledge of the models is very good, but they have English-centric limitations

사용자 관점에서 새로 공개된 최신 LLM을

영어가 아닌 다른 언어에

효율적으로 사용하는 것은 불가능

Layer Swapping for Zero-Shot Cross-Lingual Transfer in Large Language Models

Lucas Bandarkar§* Benjamin Muller Pritish Yuvraj Rui Hou Nayan Singhal
Hongjiang Lv Bing Liu

Meta GenAI
§University of California, Los Angeles

ICLR 2025 Submission
(8/6/8)

KOREA UNIVERSITY

Introduction

- LLM(대규모 언어 모델)들은 영어 데이터 편중 → 비영어권 언어 성능 저하
- 특히 수학적 추론 등에서 비영어권 IT 데이터 부족
- 목표: 영어 데이터(수학)와 비영어권 언어 데이터(일반 IT)를 이용해
  '추가 훈련 없이(post hoc)' 비영어권 언어의 수학 추론 성능을 높이는 방법 필요

➔ SFT 중 저자원 언어로 수학적 추론 능력을 전이하기 위해 두 개의 LLM을 병합하는 솔루션 제시

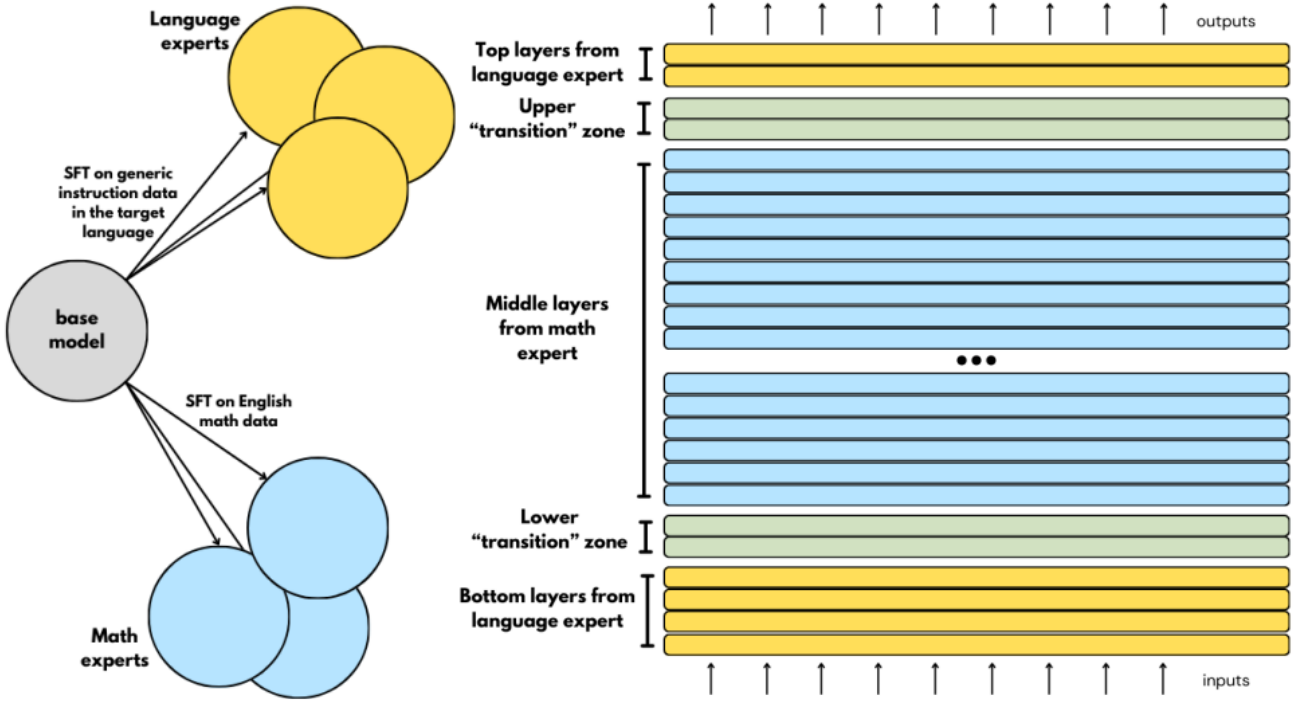


Figure 1: Our merging method which swaps in top and bottom transformer layers from a language expert into a math expert, buffered by a transition zone.

Related Works

1) Model Merging

- **체크포인트 합치기**
  - Weight Averaging
  - Parameter-level Merge
  - Sparse Fine-tuning + Merge
  - **Block-/Layer-level Merge**

2) LLM Multilinguality

- LM이 다른 능력을 훼손하지 않고, 더 많은 언어를 학습할 수 없는 curse of multilinguality (다국어의 저주)에 대한 (극복) 연구

3) Model Merging

- mixture-of-experts, cross-lingual adapters와 같은 모델의 일부를 전략적으로 공유하거나 분할하는 솔루션

Method
Layer Swapping

두 종류의 “Expert” 모델 준비

1) **Math Expert**

- 수학 데이터로 파인튜닝 Orca-Math 데이터셋(Mitra et al., 2024)

2) **Language Expert**

- Target language (스와힐리, 텔루구, 벵골어, 일본어 등) 일반 IT 데이터로 파인튜닝

Method

Layer Swapping

	attention.wq.weight	attention.wk.weight	attention.wv.weight	attention.wo.weight	feed_forward.w1.weight	feed_forward.w3.weight	feed_forward.w2.weight
31	0.614258	0.919434	0.842773	0.633789	0.322998	0.332764	0.265206
30	0.260254	0.676270	0.936279	0.797852	0.123291	0.165283	0.157296
29	0.051025	0.489990	0.836914	0.591309	0.030029	0.079590	0.112863
28	0.024170	0.304688	0.785645	0.662109	0.003174	0.041748	0.086775
27	0.024658	0.356445	0.904053	0.727295	0.001709	0.026611	0.066895
26	0.015137	0.028809	0.831299	0.645264	0.000732	0.010986	0.045271
25	0.010986	0.265869	0.887939	0.645508	0.000732	0.002441	0.024972
24	0.006104	0.043701	0.941406	0.668213	0.000732	0.001465	0.015904
23	0.003906	0.042480	0.856201	0.519287	0.000732	0.001465	0.013811
22	0.005615	0.058350	0.670166	0.456787	0.000732	0.001709	0.015974
21	0.002930	0.014893	0.490479	0.389404	0.000732	0.001709	0.016183
20	0.029297	0.152832	0.734131	0.401367	0.000732	0.001221	0.014927
19	0.032959	0.094238	0.654785	0.385986	0.000732	0.000977	0.018485
18	0.000732	0.000732	0.097168	0.123291	0.000732	0.001221	0.025670
17	0.000977	0.003174	0.081787	0.149658	0.000732	0.001221	0.027762
16	0.000732	0.004639	0.042725	0.152588	0.000732	0.000732	0.024693
15	0.000732	0.000000	0.006104	0.106689	0.000732	0.000732	0.022740
14	0.000732	0.001709	0.002686	0.017334	0.000732	0.000732	0.021275
13	0.000732	0.000488	0.003906	0.051025	0.000732	0.000732	0.025112
12	0.000732	0.016846	0.026611	0.029053	0.000732	0.000732	0.022252
11	0.000732	0.000488	0.014893	0.043945	0.000732	0.000732	0.028948
10	0.000732	0.032471	0.006348	0.046387	0.000732	0.000732	0.024135
9	0.000732	0.015625	0.007324	0.042480	0.000732	0.000732	0.014439
8	0.000732	0.010498	0.001221	0.016846	0.000488	0.000244	0.011998
7	0.000732	0.011963	0.003906	0.030762	0.000244	0.000244	0.008580
6	0.000977	0.020996	0.000488	0.017090	0.000000	0.000244	0.006348
5	0.000244	0.006836	0.002686	0.037842	0.000000	0.000000	0.008789
4	0.000244	0.017822	0.000000	0.019287	0.000000	0.000000	0.005929
3	0.002930	0.075195	0.000000	0.024658	0.000000	0.000000	0.003418
2	0.001953	0.038818	0.014893	0.001953	0.000000	0.000000	0.002930
1	0.434082	0.920654	0.306396	0.121582	0.000000	0.000000	0.004534
0	0.395508	0.628174	0.808105	0.476074	0.018311	0.022705	0.021275

(2A) Japanese expert #1

	attention.wq.weight	attention.wk.weight	attention.wv.weight	attention.wo.weight	feed_forward.w1.weight	feed_forward.w3.weight	feed_forward.w2.weight
31	0.411865	0.680908	0.840820	0.629883	0.480957	0.526611	0.374442
30	0.219971	0.334717	0.813477	0.659424	0.423584	0.459961	0.458636
29	0.076172	0.153320	0.932129	0.799072	0.371338	0.508789	0.485352
28	0.001953	0.001709	0.703857	0.610596	0.210205	0.495361	0.507812
27	0.004150	0.007324	0.917725	0.698486	0.119629	0.441162	0.482073
26	0.005127	0.004395	0.850830	0.632324	0.060791	0.403320	0.459194
25	0.000000	0.009277	0.706055	0.609619	0.034180	0.319580	0.394392
24	0.000000	0.000244	0.863037	0.556396	0.034668	0.292480	0.374233
23	0.001709	0.028320	0.855469	0.501709	0.028564	0.318848	0.397182
22	0.000732	0.020264	0.711914	0.448730	0.026855	0.328857	0.407645
21	0.001709	0.010254	0.654297	0.483643	0.025879	0.288330	0.404367
20	0.003418	0.005615	0.435791	0.401611	0.020508	0.235840	0.360840
19	0.000244	0.001709	0.373047	0.372070	0.010742	0.159668	0.329660
18	0.000000	0.000000	0.033691	0.194092	0.007568	0.099854	0.300293
17	0.000000	0.000000	0.058105	0.142334	0.001221	0.042480	0.204241
16	0.000000	0.000000	0.045166	0.144775	0.000244	0.005859	0.108817
15	0.000000	0.000244	0.004639	0.106201	0.000244	0.000977	0.060617
14	0.000000	0.000000	0.000977	0.024658	0.000000	0.000488	0.040179
13	0.000000	0.000000	0.000488	0.017822	0.000000	0.000244	0.020159
12	0.000000	0.000000	0.000244	0.003662	0.000000	0.000000	0.009068
11	0.000000	0.000000	0.000244	0.007080	0.000000	0.000000	0.006417
10	0.000000	0.000000	0.000000	0.010742	0.000000	0.000000	0.004604
9	0.000000	0.000732	0.000244	0.000488	0.000000	0.000000	0.003418
8	0.000000	0.000000	0.000000	0.000244	0.000000	0.000000	0.001186
7	0.000000	0.000000	0.000000	0.001465	0.000000	0.000000	0.000977
6	0.000000	0.000000	0.000000	0.009277	0.000000	0.000000	0.001395
5	0.000000	0.000000	0.000000	0.009277	0.000000	0.000000	0.003836
4	0.000000	0.000244	0.029053	0.035889	0.000000	0.000000	0.004395
3	0.000000	0.000000	0.001221	0.054443	0.000000	0.000000	0.006906
2	0.000000	0.000000	0.008545	0.008545	0.000000	0.000000	0.006417
1	0.016846	0.034424	0.012451	0.038818	0.000000	0.000000	0.004116
0	0.032227	0.034668	0.019531	0.042236	0.001709	0.002197	0.035854

(2B) Math expert #1

$$\mathbf{W}_{\Delta} = \mathbf{W}_{ft} - \mathbf{W}_{pre}$$

- 수학 전문성은 Midterm 레이어에 주로 집중
- 언어 전문성은 상·하단 레이어에 주로 집중하는 현상 발견

1) 공동으로 사용될 '중간 레이어'는 Math Expert 모델에서 가져옴

2) 모델의 하단 및 상단 일부 레이어는 Language Expert 모델에서 가져옴

음의 간섭(Negative Interference)을 없애기 위해, Transition Layer 도입

음의 간섭(Negative Interference): 어떤 Expert가 학습한 특화된 특징이 다른 Expert의 특징과 충돌하여, 둘 다를 동시에 살리지 못하고 한쪽(또는 둘 다)의 성능이 저하되는 것

Method
Layer Swapping

Algorithm 1 *Layer Swapping*

Input: task expert θ_{task}, language expert θ_{lang}, lower layers to swap b, upper layers to swap u, lower transition layers t_b, upper transition layers t_u, weight of each expert w_{task}, w_{lang}, number of model layers L
Output: Merged model θ_{merged}

for parameter name n in models parameters **do**
  $l \leftarrow$ layer number of n, N/A if n not attention or feedforward parameters **(1)**
  if $l < b$ or $l > L - 1 - u$ **then**
    $\theta_{merged}\{n\} \leftarrow \theta_{lang}\{n\}$ **(2)**
  else if $l > b - 1 + t_b$ or $l < L - u - t_u$ **then**
    $\theta_{merged}\{n\} \leftarrow \theta_{task}\{n\}$
  else
    $\theta_{merged}\{n\} \leftarrow (w_{task} * \theta_{task}\{n\} + w_{lang} * \theta_{lang}\{n\})/(w_{task} + w_{lang})$
  end if
end for
Return θ_{merged}

- **θ_task:** task expert의 파라미터
- **θ_lang:** language expert 의 파라미터
- **b:** 교체할 하단 층 수
- **u:** 교체할 상단 층 수
- **t_b, t_u:** 하단 및 상단 전환 구역의 층 수
- **w_task, w_lang:** 각 expert의 가중치
- **L:** 모델 총 층 수
- **θ_merged:** 병합된 모델의 출력 파라미터

1) 선택된 layer가 attention 이나 FFN 파라미터라면 처리하지 않음
2) 선택된 layer가 처음 b (하단 layer 교체 수) 보다 작거나, L-1-u (상단 layer 교체 범위) 보다 큰 경우
   ➔ language expert의 파라미터로 교환

Method
Layer Swapping

Algorithm 1 *Layer Swapping*

Input: task expert θ_{task}, language expert θ_{lang}, lower layers to swap b, upper layers to swap u, lower transition layers t_b, upper transition layers t_u, weight of each expert w_{task}, w_{lang}, number of model layers L
Output: Merged model θ_{merged}

for parameter name n in models parameters **do**
  $l \leftarrow$ layer number of n, N/A if n not attention or feedforward parameters
  if $l < b$ or $l > L - 1 - u$ **then**
    $\theta_{merged}\{n\} \leftarrow \theta_{lang}\{n\}$
  else if $l > b - 1 + t_b$ or $l < L - u - t_u$ **then**
    $\theta_{merged}\{n\} \leftarrow \theta_{task}\{n\}$ **(3)**
  else
    $\theta_{merged}\{n\} \leftarrow (w_{task} * \theta_{task}\{n\} + w_{lang} * \theta_{lang}\{n\})/(w_{task} + w_{lang})$
  end if
end for
Return θ_{merged}

- **θ_task:** task expert의 파라미터
- **θ_lang:** language expert 의 파라미터
- **b:** 교체할 하단 층 수
- **u:** 교체할 상단 층 수
- **t_b, t_u:** 하단 및 상단 전환 구역의 층 수
- **w_task, w_lang:** 각 expert의 가중치
- **L:** 모델 총 층 수
- **θ_merged:** 병합된 모델의 출력 파라미터

1) 선택된 layer가 attention 이나 FFN 파라미터라면 처리하지 않음
2) 선택된 layer가 처음 b (하단 layer 교체 수) 보다 작거나, L-1-u (상단 layer 교체 범위) 보다 큰 경우
   ➔ Language expert의 파라미터로 교환
3) 상단 및 하단 layer 교체 범위가 아닌 경우 (중단 layer의 경우)
   ➔ Task expert 파라미터로 교환

Method
Layer Swapping

Algorithm 1 *Layer Swapping*

Input: task expert θ_{task}, language expert θ_{lang}, lower layers to swap b, upper layers to swap u, lower transition layers t_b, upper transition layers t_u, weight of each expert w_{task}, w_{lang}, number of model layers L

Output: Merged model θ_{merged}

for parameter name n in models parameters **do**
  $l \leftarrow$ layer number of n, N/A if n not attention or feedforward parameters
  if $l < b$ or $l > L - 1 - u$ **then**
    $\theta_{merged}\{n\} \leftarrow \theta_{lang}\{n\}$
  else if $l > b - 1 + t_b$ or $l < L - u - t_u$ **then**
    $\theta_{merged}\{n\} \leftarrow \theta_{task}\{n\}$
  else
    $\theta_{merged}\{n\} \leftarrow (w_{task} * \theta_{task}\{n\} + w_{lang} * \theta_{lang}\{n\})/(w_{task} + w_{lang})$ **(4)**
  end if
end for
Return θ_{merged}

- **θ_task:** task expert의 파라미터
- **θ_lang:** language expert 의 파라미터
- **b:** 교체할 하단 층 수
- **u:** 교체할 상단 층 수
- **t_b, t_u:** 하단 및 상단 전환 구역의 층 수
- **w_task, w_lang:** 각 expert의 가중치
- **L:** 모델 총 층 수
- **θ_merged:** 병합된 모델의 출력 파라미터

1) 선택된 layer가 attention 이나 FFN 파라미터라면 처리하지 않음
2) 선택된 layer가 처음 b (하단 layer 교체 수) 보다 작거나, L-1-u (상단 layer 교체 범위) 보다 큰 경우
   ➔ Language expert의 파라미터로 교환
3) 상단 및 하단 layer 교체 범위가 아닌 경우 (중단 layer의 경우)
   ➔ Task expert 파라미터로 교환
4) 상단, 중단, 하단 layer가 아닌 transition layer 구역의 파라미터
   - 하단 transition layer {b} ~ {b + t_b}
   - 상단 transition layer {L – u – t_u} ~ {L – u}

➔ Task, Lang Expert의 파라미터 가중 평균 사용

Method
Layer Swapping

Algorithm 1 *Layer Swapping*

Input: task expert θ_{task}, language expert θ_{lang}, lower layers to swap b, upper layers to swap u, lower transition layers t_b, upper transition layers t_u, weight of each expert w_{task}, w_{lang}, number of model layers L
Output: Merged model θ_{merged}

for parameter name n in models parameters **do**
  $l \leftarrow$ layer number of n, N/A if n not attention or feedforward parameters
  if $l < b$ or $l > L - 1 - u$ **then**
    $\theta_{merged}\{n\} \leftarrow \theta_{lang}\{n\}$
  else if $l > b - 1 + t_b$ or $l < L - u - t_u$ **then**
    $\theta_{merged}\{n\} \leftarrow \theta_{task}\{n\}$
  else
    $\theta_{merged}\{n\} \leftarrow (w_{task} * \theta_{task}\{n\} + w_{lang} * \theta_{lang}\{n\})/(w_{task} + w_{lang})$
  end if
end for
Return θ_{merged}

- **θ_task:** task expert의 파라미터
- **θ_lang:** language expert 의 파라미터
- **b:** 교체할 하단 층 수
- **u:** 교체할 상단 층 수
- **t_b, t_u:** 하단 및 상단 전환 구역의 층 수
- **w_task, w_lang:** 각 expert의 가중치
- **L:** 모델 총 층 수
- **θ_merged:** 병합된 모델의 출력 파라미터

1) 선택된 layer가 attention 이나 FFN 파라미터라면 처리하지 않음
2) 선택된 layer가 처음 b (하단 layer 교체 수) 보다 작거나, L-1-u (상단 layer 교체 범위) 보다 큰 경우
   ➔ Language expert의 파라미터로 교환
3) 상단 및 하단 layer 교체 범위가 아닌 경우 (중단 layer의 경우)
   ➔ Task expert 파라미터로 교환
4) 상단, 중단, 하단 layer가 아닌 transition layer 구역의 파라미터
   - 하단 transition layer {b} ~ {b + t_b}
   - 상단 transition layer {L – u – t_u} ~ {L – u}

➔ Task, Lang Expert의 파라미터 가중 평균 사용

Experimental Setup

1) 모델

- Llama 3.1 8B

2) Math Expert

- 영어로 작성된 수학 데이터로 파인튜닝 (30-40K)

3) Language Expert

- 스와힐리, 텔루구, 벵골어, 일본어 각각에 대해 학습
- 일반 IT 데이터로 파인튜닝 (30-40K)
    - NER, Translation, QA 등 여러 NLP task 포함
    - 수학 데이터 포함하지 않음

A.1 FINE-TUNING DATASETS

Table 3: Datasets used for supervised-fine-tuning (SFT) in this project

Category	Datasets	URL
Math	Orca Math word problems dataset from Microsoft (Mitra et al., 2024)	https://huggingface.co/datasets/microsoft/orca-math-word-problems-200k
Telugu	Aya Dataset from Cohere for AI (Singh et al., 2024a)	https://huggingface.co/datasets/CohereForAI/aya_dataset
	NLLB English-Telugu translation data from FAIR (NLLB et al., 2022)	https://huggingface.co/datasets/allenai/nllb
	English instruction dataset, machine translated to Telugu	
Bengali	Aya Dataset by Cohere for AI (Singh et al., 2024a)	https://huggingface.co/datasets/CohereForAI/aya_dataset
	English-Bengali translation data from NLLB (NLLB et al., 2022)	https://huggingface.co/datasets/allenai/nllb
	IndicShareLlama dataset from AI4Bharat (Khan et al., 2024)	https://huggingface.co/datasets/ai4bharat/indic-align
	BongChat dataset from Lumatic AI	https://huggingface.co/datasets/lumatic-ai/BongChat-v1-253k
Swahili	Aya Dataset by Cohere for AI (Singh et al., 2024a)	https://huggingface.co/datasets/CohereForAI/aya_dataset
	English-Swahili translation data from NLLB (NLLB et al., 2022)	https://huggingface.co/datasets/allenai/nllb
	Inkuba dataset from Lelapa (Tonja et al., 2024)	https://huggingface.co/datasets/lelapa/Inkuba-instruct
	xP3 MT dataset from BigScience, with FLORES samples removed (Muennighoff et al., 2022)	https://huggingface.co/datasets/bigscience/xP3mt
Japanese	Aya Dataset by Cohere for AI (Singh et al., 2024a)	https://huggingface.co/datasets/CohereForAI/aya_dataset
	English-Japanese translation data from NLLB (NLLB et al., 2022)	https://huggingface.co/datasets/allenai/nllb
	LLM-Japanese dataset from Izumi Lab (Hirano et al., 2023)	https://huggingface.co/datasets/izumi-lab/llm-japanese-dataset
	Ichikara dataset from RIKEN AIP	https://huggingface.co/datasets/platdev/ichikara-instruction
	Dolly dataset from Databricks, machine translated to Japanese	https://huggingface.co/datasets/kunishou/databricks-dolly-15k-ja

Experimental Setup

4) Configuration

총 레이어 수 L = 32인 LLAMA 3.1에서, 아래 값을 권장 범위로 실험

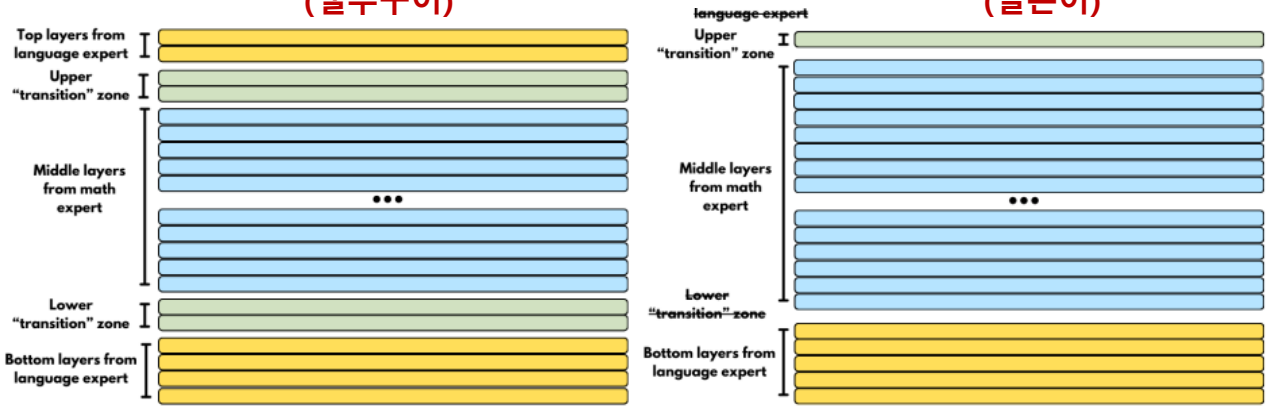
(Underline이 default config, 가장 좋은 조합이 best config)

- 교환할 하단 레이어 수 **b ∈ {3, 4, 5}**
- 교환할 상단 레이어 수 **u ∈ {0, 1, 2}**
- 하단/상단 transition 레이어 수 **t_b, t_u ∈ {0, 1, 2}**
- transition zones은 두 Expert의 레이어의 평균(즉, 수프)으로, 가중되지 않거나 조정된 가중 평균을 사용
- 입력 토큰 임베딩, 출력 레이어 등은 두 Expert 파라미터의 평균을 사용하는 것이 더 효과적

언어 난이도(낮은 리소스)나 모델 성능 등에 따라 b, u, t_b, t_u를 조정해 최적화 가능

➔ 일반적으로 저자원 언어의 경우, 더 많은 language expert의 수 필요

저자원 언어 (텔루구어) **고자원 언어 (일본어)**



(3A) Possible config. with more layers swapped in. (3B) Possible config. with less layers swapped in.

Figure 3: The comparison of the maximum (left) and minimum (right) swapping setups that we find effective empirically. Note on the right, there are no upper layers directly from the language expert.

Experimental Setup

5) Baselines

– 기공개된 Llama 3.1 8B IT

– 개별 Expert 성능 (Math Expert, Language Expert)

➔ **Expert를 만들기 위해 여러 번 파인튜닝(run)을 수행했을 때, 그중 가장 성능이 좋았던 상위 3개의 체크포인트 사용**

– 기본 Model Soup(Weight Averaging)

– Layer Swapping

➔ **best config (9 pairs), default config (9 pairs)**

6) 평가 데이터셋

– MGSM (Multilingual Grade School Math Benchmark) 8-shot

Results

Table 1: **MGSM 8-shot results of *layer swapping* across four languages** compared to the individual experts and model souping. Note that for aggregate statistics of the individual SFT runs, we select the 3 best checkpoints from numerous training runs with periodic checkpointing. The merging methods are aggregated over the 9 pairs (3 language experts × 3 math experts), which means the min, avg, and max measures are not perfectly comparable.

Setting	LLAMA 3.1 8B	language expert	math expert	model soup	*layer swap*	*layer swap*
Details		top 3 training runs	top 3 training runs	best config, 9 pairs	default config, 9 pairs	best config, 9 pairs
Swahili avg	*24.8*	24.7	29.5	29.3	32.4	**32.8**
Swahili max	*24.8*	25.6	32.8	32.0	36.4	**37.2**
Telugu avg	*12.0*	20.0	20.1	20.9	22.7	**23.0**
Telugu max	*12.0*	22.4	24.0	26.4	27.6	**27.6**
Bengali avg	*29.2*	33.5	38.3	36.8	37.1	**38.7**
Bengali max	*29.2*	35.2	44.4	38.4	40.4	**45.2**
Japanese avg	*33.6*	35.9	**42.7**	38.7	38.5	40.1
Japanese max	*33.6*	36.8	**44.8**	40.0	40.8	43.2

- **Layer Swapping > Math Expert > Language Expert**
- **Model Soup는 음의간섭으로 인해 Layer Swapping만큼 효과적이지 못함**
- **스와힐리, 텔루구, 벵골어에서 평균 10% 향상**
- **Best config를 찾으면 높은 성능 향상 기대 가능**

Results

Table 2: **MGSM 8-shot results of *layer swapping* for Swahili** in more detail and with two additional comparisons, TIES-merging and dataset merging. We display the minimum performance in Swahili, as well as the average across all 9 languages in MGSM and in English.

Setting	LLAMA 3.1 8B	Swahili expert	math expert	swh&math joint SFT	model soup	TIES-merging	*layer swap*	*layer swap*
Details		top 3 training runs	top 3 training runs	top 3 training runs	best config, 9 pairs	best config, 9 pairs	default config, 9 pairs	best config, 9 pairs
Swahili min	*24.8*	23.6	27.2	31.6	25.6	25.2	29.6	29.2
Swahili avg	*24.8*	24.7	29.5	32.1	29.3	29.5	32.4	**32.8**
Swahili max	*24.8*	25.6	32.8	32.8	32.0	32.4	36.4	**37.2**
English avg	*56.0*	55.7	66.2	64.3	62.0	60.1	64.7	64.4
All langs avg	*37.7*	37.5	45.4	46.0	43.0	41.8	44.1	44.4

- **swh & math joint SFT:** 스와힐리어 일반 + 영어 수학 데이터를 합쳐 SFT (80K)
- **TIES-merging:** 파라미터의 부호·중요도를 고려하여 두 Expert를 병합 (Yadav et al., 2023)

- **"layer swapping"이 가장 높은 max 점수(37.2)를 기록하는 동시에, 영어 평균 (64점대)과 전체 평균 (44점대)도 크게 떨어뜨리지 않음**
- **"swh & math joint SFT"도 우수하지만, 이는 애초에 더 큰 데이터(mixed)를 한 번에 학습한 경우임**
- **단순 모델 병합(model soup)이나 TIES-merging은 각 전문가 능력을 모두 반영하기에 한계 있으며, 여전히 음의 간섭 존재**

Conclusions

- 단순하지만 효과적인 레이어 재조합 → 비영어권 언어에서 수학 추론을 post hoc로 개선
- 모델 내부에서 “언어 변환 레이어”와 “추론(수학) 레이어”가 부분적으로 분화되어 있다는 증거 제공
- 다른 추론(코딩, 논리, 과학 등) Task에서 일반화 검증 필요

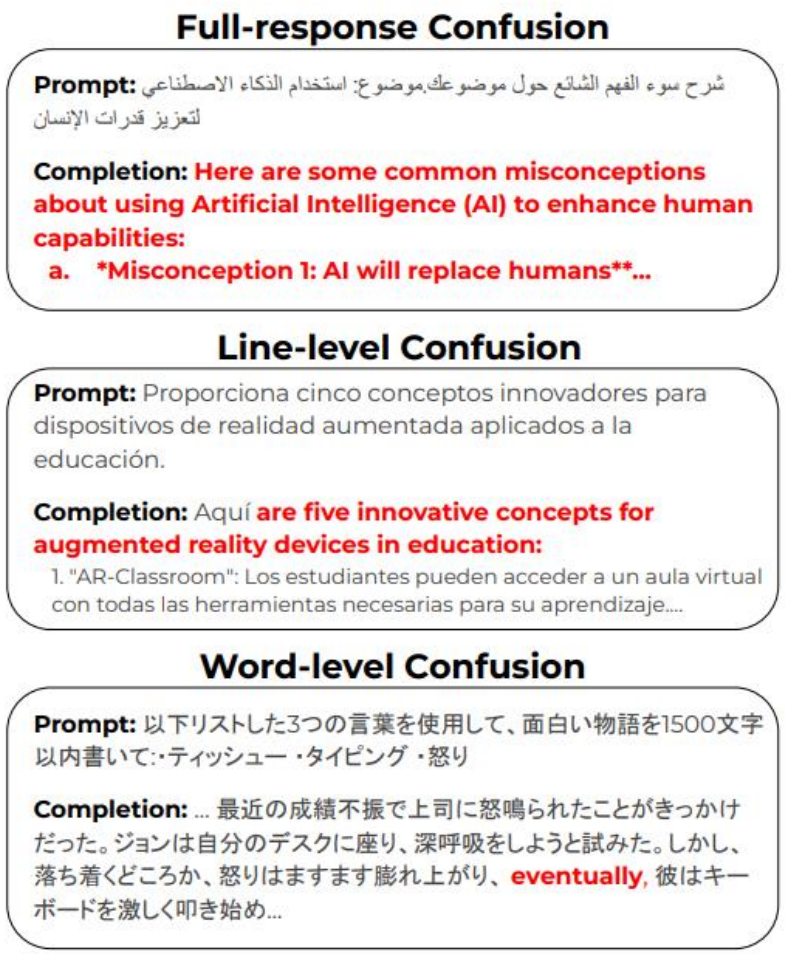
Understanding and Mitigating Language Confusion in LLMs

Kelly Marchisio*, Wei-Yin Ko*, Alexandre Bérard,
Théo Dehaze, Sebastian Ruder*
Cohere

EMNLP 2024 main

KOREA UNIVERSITY

Introduction



Language Confusion

- 사용자가 원하는 언어로 일관되게 텍스트를 생성하지 못하는 현상

Introduction

Contributions

1. LLMs에서 Language Confusion 문제를 식별하고 설명
2. LLMs에서 Language Confusion을 측정하기 위한 새로운 벤치마크와 메트릭 소개
3. 다양한 LLMs에 대한 체계적인 평가를 수행하여 실제로 Language Confusion 이 발생하는 시점 조사
4. LLMs에서 Language Confusion을 완화하기 위한 방법 제안

Language Confusion Benchmark

Generation Settings

Monolingual generation

- 사용자가 언어 l로 모델에 질의하고 l로 응답을 기대
- 사용자가 종종 모국어로 기술과 상호작용하기를 선호하기 때문에 가장 일반적인 사용 시나리오

Cross-lingual generation

- 사용자가 언어 l로 모델에 요청을 다른 언어 l′로 수행하도록 지시
- 요청된 언어 l′는 지시언어 l와 다름. (실험에서 지시언어 l를 영어로 설정)
- *ex) … 위 내용을 바탕으로, 영어로 논문을 수정하세요.*
  (지시언어 l: 한국어, 요청된 언어 l′ : 영어)

Language Confusion Benchmark

Language Confusion Metrics

language identification (LID) 툴인 FastText를 사용하여 언어 식별

Line-level detection: 모델의 응답을 line으로 분해 후, LID로 예측 (4 단어 이상에만 적용)

Word-level detection: LID는 word-level 적용 시 너무 낮은 성능으로, 단어 소싱 후 단어 사전 사용

$$\text{LPR} = \frac{|R \setminus E_L|}{|R|}$$

Language Confusion Benchmark

Language Confusion Metrics

language identification (LID) 툴인 FastText를 사용하여 언어 식별

Line-level detection: 모델의 응답을 line으로 분해 후, LID로 예측 (4 단어 이상에만 적용)

Word-level detection: LID는 word-level 적용 시 너무 낮은 성능으로, 단어 소싱 후 단어 사전 사용

- **Line-level pass rate (LPR):** line-level로 LID 감지기를 에러 없이 통과한 모델 응답의 비율

$$\mathrm{LPR} = \frac{|R \setminus E_L|}{|R|}$$

- **Word-level pass rate (WPR):** 모든 단어가 원하는 언어로 된 응답의 비율 (유형 분리를 위해, line-level 제거)

$$\mathrm{WPR} = \frac{|(R \setminus E_L) \setminus E_W|}{|R \setminus E_L|}$$

- **Language confusion pass rate (LCPR):** LPR과 WPR의 조화 평균 $\mathrm{LCPR} = 2 \times \frac{\mathrm{LPR} \times \mathrm{WPR}}{\mathrm{LPR} + \mathrm{WPR}}$

Language Confusion Benchmark

Data Sources & Processing

	Dataset name	Reference	Nature of data	$\|L\|$	$\|D\|$	Languages	W
Mono-lingual	Aya	Singh et al. (2024)	Human-generated	100	500	en, tr, ar, zh, pt	9
	Dolly	Singh et al. (2024)	MT post-edited	100	500	hi, ru, fr, ar, es	10
	Okapi	Lai et al. (2023)	Synthetic + MT	100	1.2k	en, fr, it, de, zh, vi, ru, es, id, pt, ar, hi	13
	Native prompts	Ours	Human-generated	100	400	es, fr, ja, ko	19
Cross-lingual	Okapi	Lai et al. (2023)	Synthetic	100	1.5k	$\mathcal{L}$	15
	ShareGPT	https://sharegpt.com/	Human-generated	100	1.5k	$\mathcal{L}$	18
	Complex prompts	Ours	Human-generated	99	1.5k	$\mathcal{L}$	159

Table 1: **Data sources in the LCB for monolingual and cross-lingual generation.** $|D|$ is the total number of examples per data source and $|L|$ is the number of examples per language. For the cross-lingual setting, the model is instructed in English to generate in the target language $l \in \mathcal{L}$ where $\mathcal{L} = \{\text{fr, de, es, pt, it, ja, ko, zh, ar, tr, hi, ru, id, vi}\}$. W is the median length in words of the prompts in each dataset.

	Prompt	Translation	Data source
Monolingual	问：如何清洗和保养筷子?	Q: How to clean and maintain chopsticks?	Aya
	¿Cómo escapar de un helicóptero atrapado en el agua?	How to escape from a helicopter stuck in water?	Dolly
	Erklären Sie, wie der Gini-Index berechnet wird.	Explain how the Gini index is calculated.	Okapi
	日本で救急隊員を目指す人は、どのような教育やトレーニングを受ける必要がありますか。	What kind of education and training do people who want to become emergency workers in Japan need to undergo?	Native prompts (Ours)
Cross-lingual	Generate an essay **in Korean** of at least 500 words that argues in favor of regulating artificial intelligence.		Okapi
	Respond in French. You are a medical communications expert. Please provide a summary on how pharma companies are approaching diversity and inclusion, and health inequalities globally. Focus on the general approach and include information on clinical trials.		Share-GPT
	Based solely on the text below: 1. Extract the statistical techniques and machine learning algorithms analysts employ to uncover relationships and patterns within the data. 2. Generate 5 fill-in-the-blanks style questions 3. Summarize the text in 100 words [...] **Reply in Turkish.**		Complex prompts (Ours)

Table 2: **An example prompt from each dataset used for monolingual and cross-lingual generation.** English translations are shown for convenience. For cross-lingual generation, prompts are in English and have been amended with an instruction to generate the output in another language. The complex prompt example is truncated.

Experimental Results

Line-level pass rate (LPR)

	Monolingual															
	avg	ar	de	en	es	fr	hi	id	it	ja	ko	pt	ru	tr	vi	zh
Llama 2 70B-I	48.3	0.3	59.0	99.0	95.7	87.7	1.0	62.0	72.0	7.0	0.0	91.0	88.9	33.0	17.0	10.5
Llama 3 70B-I	46.0	21.7	31.0	**100.0**	98.3	88.7	23.0	21.0	88.0	10.0	0.0	95.5	77.0	18.0	10.0	8.0
Llama 3.1 70B-I	99.0	98.9	**100.0**	98.5	99.0	**100.0**	**100.0**	94.0	**100.0**	96.9	**100.0**	**99.0**	**100.0**	**100.0**	**100.0**	**99.0**
Mixtral 8x7B	73.0	48.3	90.9	99.5	89.3	95.3	71.0	58.0	72.0	66.7	61.2	85.0	65.0	90.0	57.0	45.5
Mistral Large	69.9	48.0	98.0	99.0	99.0	**100.0**	19.0	31.0	99.0	48.0	64.0	79.5	98.0	71.0	29.0	66.0
Command R	98.6	**100.0**	98.0	99.5	95.7	99.3	**100.0**	92.0	99.0	**100.0**	**100.0**	98.5	**100.0**	99.0	99.0	98.5
Command R+	99.2	99.7	**100.0**	**100.0**	99.3	99.7	**100.0**	**97.0**	**100.0**	99.0	**100.0**	97.5	**100.0**	**100.0**	99.0	97.5
Command R Refresh	98.9	99.6	**100.0**	99.5	99.3	99.7	**100.0**	92.0	**100.0**	99.0	**100.0**	98.0	**100.0**	99.0	**100.0**	98.0
Command R+ Refresh	**99.3**	99.0	**100.0**	**100.0**	99.3	**100.0**	**100.0**	96.0	**100.0**	**100.0**	**100.0**	97.5	99.0	**100.0**	**100.0**	98.0
GPT-3.5 Turbo	99.1	**100.0**	**100.0**	99.5	**99.7**	**100.0**	99.0	96.0	**100.0**	98.0	**100.0**	98.0	**100.0**	**100.0**	99.0	97.0
GPT-4 Turbo	**99.3**	99.0	**100.0**	**100.0**	99.3	99.3	**100.0**	96.0	99.0	**100.0**	**100.0**	98.0	**100.0**	**100.0**	**100.0**	**99.0**
GPT-4o	98.9	99.7	**100.0**	**100.0**	99.3	99.3	99.0	94.0	**100.0**	99.0	**100.0**	97.5	99.0	**100.0**	99.0	98.0
	Cross-lingual															
	avg	ar	de	-	es	fr	hi	id	it	ja	ko	pt	ru	tr	vi	zh
Llama 2 70B-I	38.4	12.4	52.3	-	77.3	71.1	21.2	46.0	66.5	16.2	4.8	75.9	38.3	24.0	20.4	11.1
Llama 3 70B-I	30.3	31.1	34.7	-	61.1	53.1	46.4	25.4	36.4	1.4	0.8	54.4	38.4	17.4	18.7	4.3
Llama 3.1 70B-I	81.4	77.2	87.5		90.4	90.5	95.6	**97.1**	88.1	59.4	51.5	86.0	76.5	85.7	93.6	69.9
Mixtral 8x7B	69.0	59.1	76.4	-	79.1	79.0	39.2	72.8	85.0	57.9	56.9	79.4	72.4	76.0	75.8	57.5
Mistral Large	58.2	36.1	74.5	-	68.5	71.9	58.5	59.2	65.8	44.5	41.1	64.5	63.3	65.9	54.8	46.5
Command R	68.1	61.6	63.2	-	72.5	74.4	65.5	70.8	65.7	65.3	69.2	67.2	69.4	67.7	65.7	75.0
Command R+	91.2	93.4	91.6	-	91.7	91.5	90.2	85.9	93.8	93.8	91.1	88.5	93.0	92.0	91.1	89.5
Command R Refresh	93.1	91.9	96.1	-	96.4	94.0	95.0	85.1	93.8	95.0	93.8	**94.0**	92.2	93.4	94.1	88.9
Command R+ Refresh	**95.4**	**95.4**	**97.5**	-	**97.6**	**97.2**	**98.2**	88.9	**96.2**	**95.1**	**95.9**	91.7	**96.4**	**97.9**	**97.9**	90.2
GPT-3.5 Turbo	89.8	90.8	90.2	-	93.3	87.8	92.0	84.5	91.3	88.3	90.3	89.9	91.8	89.2	91.8	86.4
GPT-4 Turbo	90.3	88.9	93.0	-	93.1	90.7	91.0	87.3	91.8	87.7	89.7	91.0	90.0	91.4	90.0	87.9
GPT-4o	92.4	95.0	92.9	-	95.8	93.5	91.9	85.4	94.1	92.5	92.4	88.0	92.6	95.1	92.7	**91.3**

Table 3: **Line-level pass rate (LPR) on monolingual and cross-lingual generation, by language.**

Monolingual generation

- Command 및 GPT 모델은 line-level에서 평균적으로 잘 함
  - GPT-4-turbo > GPT-4o
- Llama2, Llama3과 Mistral 모델 못함
- Llama3.1부터 잘함

Cross-lingual generation

- monolingual 대비 굉장히 어려움을 보임
- OpenAI 및 Command 모델이 가장 잘 수행
- Command R+ Refresh가 최고 성능
- Llama3.1부터 잘하지만, 어려워 함

Experimental Results

Word-level pass rate (WPR)

	Monolingual	Cross-lingual
Llama 2 70B-I	97.9	84.2
Llama 3 70B-I	93.0	94.4
Llama 3.1 70B-I	99.5	95.0
Mixtral 8x7B	73.7	68.2
Mistral Large	98.4	93.8
Command R	96.3	94.0
Command R+	99.4	95.1
Command R Refresh	99.4	97.2
Command R+ Refresh	**99.8**	96.5
GPT-3.5 Turbo	**99.8**	**98.7**
GPT-4 Turbo	99.7	96.6
GPT-4o	99.7	98.1

Table 4: **Average word-level pass rate (WPR) on non-Latin script languages**. See Tables A3 and A4 for detailed WPR results on non-Latin and Latin script languages respectively.

Monolingual generation

- Command 및 GPT 모델이 LPR과 유사하게 평균적으로 잘함
- Llama2, Llama3과 Mistral 모델 못함
- Llama3.1부터 잘함

Cross-lingual generation

- WPR에서는 LPR 대비 에러율이 현저히 떨어짐.

  ➔ 즉, 대부분 line-level에서 language confusion이 발생함

Experimental Results

PT vs SFT

	avg	ar	hi	ja	ko	vi	zh
Llama 2 70B	**98.5**	**99.6**	**100.0**	**100.0**	**100.0**	**98.0**	**93.2**
Llama 2 70B-I	6.0	0.3	1.0	7.0	0.0	17.0	10.5
Llama 3 70B	**94.7**	**96.7**	**97.9**	**87.9**	**98.8**	**97.0**	**90.0**
Llama 3 70B-I	12.1	21.7	23.0	10.0	0.0	10.0	8.0
Command R base	85.9	94.9	81.0	93.9	94.2	83.0	68.1
Command R	**99.6**	**100.0**	**100.0**	**100.0**	**100.0**	**99.0**	**98.5**
Command R+ base	78.4	92.8	67.0	90.5	93.5	65.7	60.9
Command R+	**99.2**	**99.7**	**100.0**	**99.0**	**100.0**	**99.0**	**97.5**

Table 6: **Line-level pass rate (LPR) of base vs instruction-tuned LLMs on monolingual generation** for a subset of languages. Full results in Table A17, §A.8.

공개된 사전학습 LLM과 Instruction-tuned 모델의 비교

- Instruction-tuned Command R 모델은 기본 버전보다 language confusion이 적음
- Instruction-tuned Llama 모델은 language confusion이 높음

➔ **영어 중심의 SFT가 되었다는 것을 암시**

Experimental Results

When does language confusion occur?

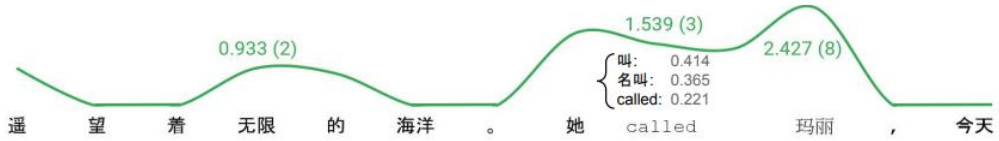


Figure 2: **A model is vulnerable to world-level language confusion when the number of tokens in the sampling nucleus is high, and the distribution is flat.** Metrics: Shannon entropy; in brackets: # of tokens in nucleus.

	Avg. Nucleus Size			Avg. Entropy		
	Overall	@CP	¬@CP	Overall	@CP	¬@CP
Has CP	1.64	3.56	1.61	0.353	1.228	0.337
No CP	1.61	-	1.61	0.365	-	0.365
All	1.62	3.56	1.61	0.361	1.228	0.356

Table 7: **Avg. nucleus size, entropy at confusion points** (sampling points where language switch did [@CP] or did not [¬@CP] occur) for 15 Chinese responses. Responses are split into those which had at least one CP ("Has CP") or zero CPs ("No CP").

Okapi 모델로 15개의 중국어 프롬프트에 대한 응답을 생성, 여기서 영어로의 의도치 않은 언어 전환 사례 발견

언어 전환이 발생한 구체적인 위치를 "confusion point (CP)"이라고 부르며, 총 9개의 CP 발견 (각 프롬프트마다 100개의 토큰을 생성하여, 총 1500개의 토큰으로 구성된 데이터셋을 분석**)**

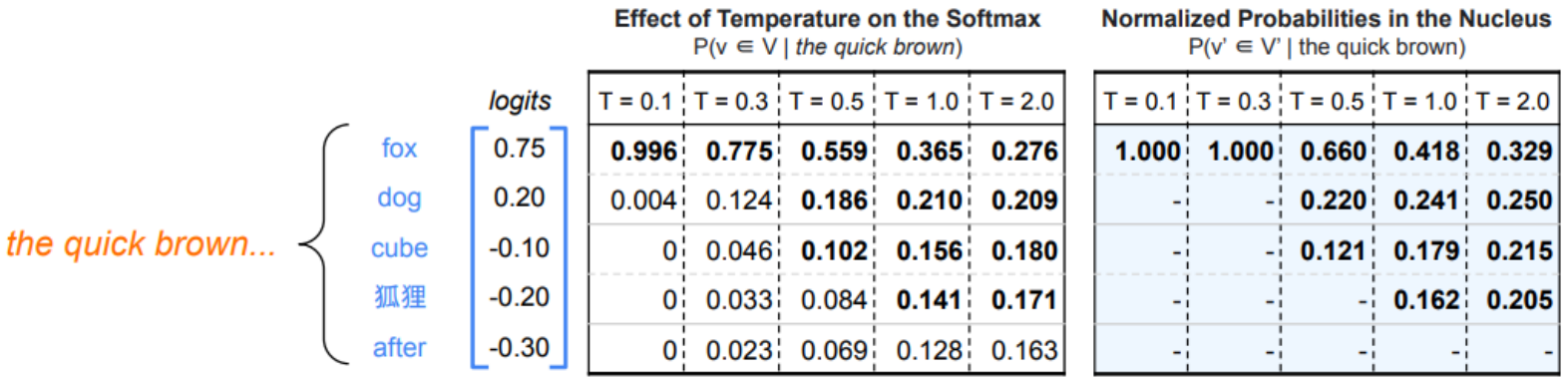
- CP 분석: 총 9개의 CP를 찾음. 이 지점에서 영어 토큰이 의도치 않게 등장함
- language confusion이 있는 샘플과 없는 샘플은 전체적으로 유사한 평균 nucleus size와 엔트로피 값을 보임
- 하지만, CP에서의 nucleus size 와 엔트로피는 상대적으로 높음

  ➔ language confusion이 발생할 때 모델의 불확실성이 증가함을 시사

Language Confusion이 모델의 다음 토큰에 대한 불확실성과 관련이 있으며, 엔트로피와 nucleus size가 높을 때 더욱 발생할 가능성이 크다는 것을 발견

Mitigating Language Confusion

1. Reducing temperature and nucleus size



		logits	Effect of Temperature on the Softmax P(v ∈ V | *the quick brown*)					Normalized Probabilities in the Nucleus P(v' ∈ V' | the quick brown)				
			T = 0.1	T = 0.3	T = 0.5	T = 1.0	T = 2.0	T = 0.1	T = 0.3	T = 0.5	T = 1.0	T = 2.0
the quick brown...	fox	0.75	**0.996**	**0.775**	**0.559**	**0.365**	**0.276**	**1.000**	**1.000**	**0.660**	**0.418**	**0.329**
	dog	0.20	0.004	0.124	**0.186**	**0.210**	**0.209**	-	-	**0.220**	**0.241**	**0.250**
	cube	-0.10	0	0.046	**0.102**	**0.156**	**0.180**	-	-	**0.121**	**0.179**	**0.215**
	狐狸	-0.20	0	0.033	0.084	**0.141**	**0.171**	-	-	-	**0.162**	**0.205**
	after	-0.30	0	0.023	0.069	0.128	0.163	-	-	-	-	-

Figure 3: **Effect of Temperature (T) in Nucleus Sampling**. Tokens in the nucleus at $p = 0.75$ are **bold**. Middle: Effect of T on the softmax probabilities (Equation 1). Right: Effect of T on the probabilities of tokens in the nucleus right before sampling (Equation A.10). As T increases, the token 狐狸 has less chance to be sampled.

	avg	ar	hi	ja	ko	ru	zh
T=0.0	**97.2**	97.6	**100.0**	**96.9**	97.0	**96.0**	**95.9**
T=0.3	96.3	**99.3**	99.0	93.9	97.0	**96.0**	92.3
T=0.5	96.4	97.9	99.0	94.9	**99.0**	95.0	92.3
T=0.7	94.2	98.0	98.0	91.8	93.9	93.0	90.3
T=1.0	86.5	95.9	93.8	74.5	92.8	87.5	74.7
p=0.1	97.4	98.3	**100.0**	98.0	**97.0**	95.0	95.9
p=0.3	97.3	98.0	**100.0**	**99.0**	**97.0**	94.0	95.9
p=0.5	**97.6**	98.0	**100.0**	96.9	96.0	**98.0**	**96.9**
p=0.75	96.3	**99.3**	99.0	93.9	**97.0**	96.0	92.3

Table 8: **Effect of varying temperature (T) or nucleus size (p) on monolingual word-level language confusion (WPR) of *Command R***. Default values are $p = 0.75$ and $T = 0.3$. **Best score**. Worst score.

➔ 높은 Temperature는 Language Confusion이 발생할 수 있음

➔ p를 감소시켜서, nucleus size를 작게 하는 것은 작은 효과만 있음

Mitigating Language Confusion

2. Beam search decoding

	Monoling.		Crosslingual			
	WPR	LPR	WPR	LPR		
				Overall	*¬ IE*	*IE*
1	97.8	**99.0**	94.9	**74.1** -	**73.9** -	**74.2** -
2	98.6	**99.0**	95.4	72.2 *(-1.9)*	71.1 *(-2.8)*	73.6 *(-0.6)*
3	98.6	98.7	**97.1**	71.5 *(-2.5)*	70.1 *(-3.8)*	73.4 *(-0.9)*
5	**99.0**	**99.0**	96.7	70.3 *(-3.8)*	68.3 *(-5.6)*	72.9 *(-1.4)*
10	**99.0**	98.5	96.7	68.4 *(-5.7)*	65.6 *(-8.3)*	72.1 *(-2.1)*

Table 9: **Effect of beam search decoding on language confusion metrics for *Command R*.** Beam sizes: 1-10.

➔ Beam size를 증가시키면, Monolingual WPR에서 성능이 좋아지지만, LPR은 큰 변화가 없음

➔ Beam size를 증가시키면, Cross-lingual LPR에 부정적인 영향을 미침

Mitigating Language Confusion

3-4 Few-shot prompting & Multilingual instruction tuning

	Monolingual		Cross-lingual	
	LPR	WPR	LPR	WPR
Command R Base	86.2	98.7	1.1	**100.0**
+ Q/A template (0-shot)	85.3	99.7	20.9	97.0
+ 1-shot	94.1	**100.0**	90.7	98.6
+ 5-shot	**99.0**	**100.0**	**95.0**	99.7
+ English SFT	77.8	96.2	78.3	91.7
+ English pref. tuning	74.3	90.9	85.7	87.4
+ Multilingual SFT	98.3	95.5	78.2	90.0
+ Multi. pref. tuning	98.9	93.4	89.4	86.9
Command R	98.6	96.3	68.1	94.0
+ 1-shot	68.3	92.7	82.9	92.3

Table 11: **Effect of few-shot prompting and instruction tuning on language confusion.**

English-only tuning

영어 전용 공개 IT 데이터로 SFT 한 후, 이 모델에 영어 전용 선호도 튜닝 적용

Multilingual tuning

영어 데이터를 다국어 데이터로 확장. (기계 번역된 Dolly 및 ShareGPT 사용)

다국어 데이터의 부족으로 인해, 우리의 SFT 데이터 혼합은 90% 영어 사용.

선호도 튜닝의 경우, 50% 다국어 데이터를 사용

➔ **Few-shot prompting은 language confusion을 완화할 수 있는 좋은 방법**

➔ **monolingual에서 영어 SFT, 선호도 튜닝은 language confusion을 악화. 단지 10%의 다국어 데이터로 SFT를 수행하는 것만으로도 line-level confusion을 거의 제거할 수 있음**

➔ **cross-lingual에서 다국어 튜닝이 영어 전용 튜닝보다 line-level 성능을 더 좋게 만들지는 않음.**
cross-lingual 데이터셋이 영어 프롬프트만 포함하고 있어서 모델이 영어 지시를 따르는 것을 배우는 것이 더 중요하게 여기기 때문일 수 있음

Conclusions

- 대부분 공개된 LLM이 Language Confusion을 보임 ➔ 이는 영어 중심의 SFT 때문으로 추측
- LLM을 개발 및 배포할 때, Language Confusion에 대한 검증도 필수적

Thank you!