

ML-Summit 2023

ML-Summit 2023

全球机器学习技术大会

大规模语言模型中语言与知识

张奇

复旦大学

■ 目录

1 Multilingual BERT 中存在多语言对齐现象

2 大语言模型中多语言对齐

3 大语言模型中的语言 and 知识分离

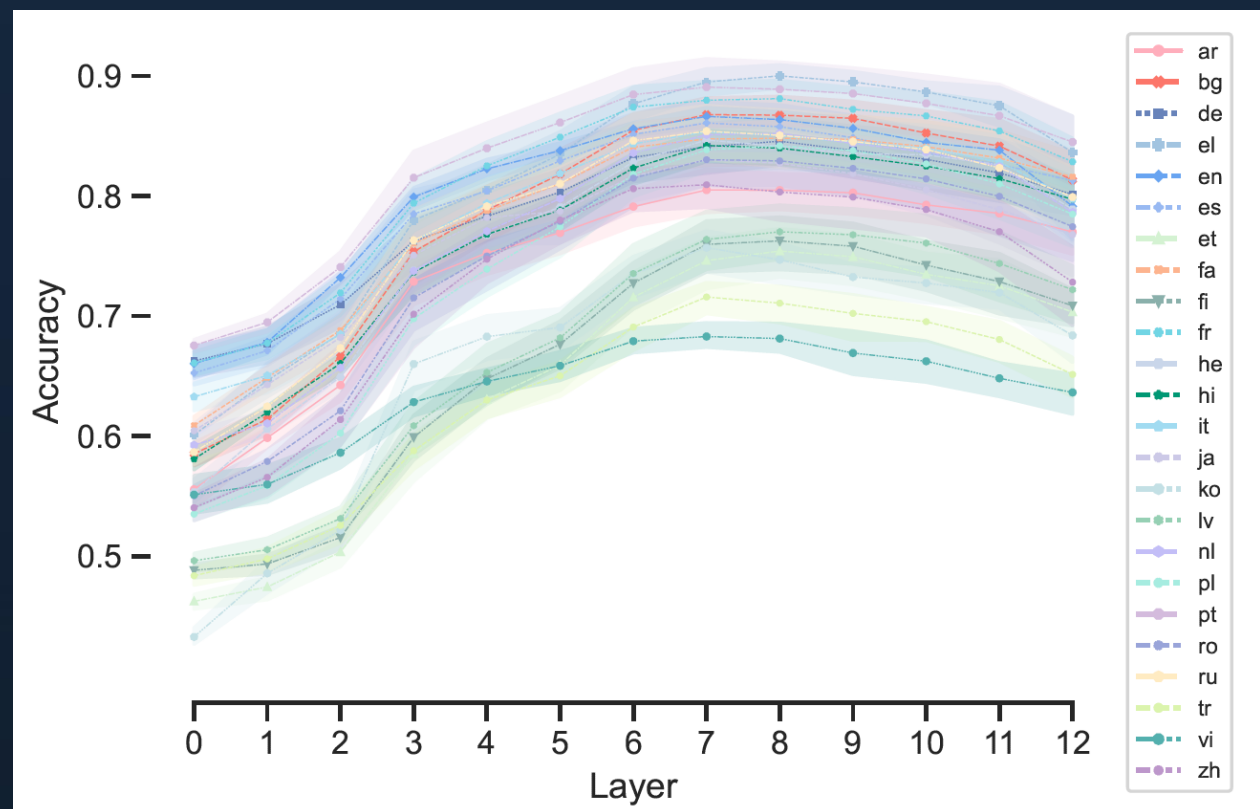


01

Multilingual BERT 中存在多语言对齐现象



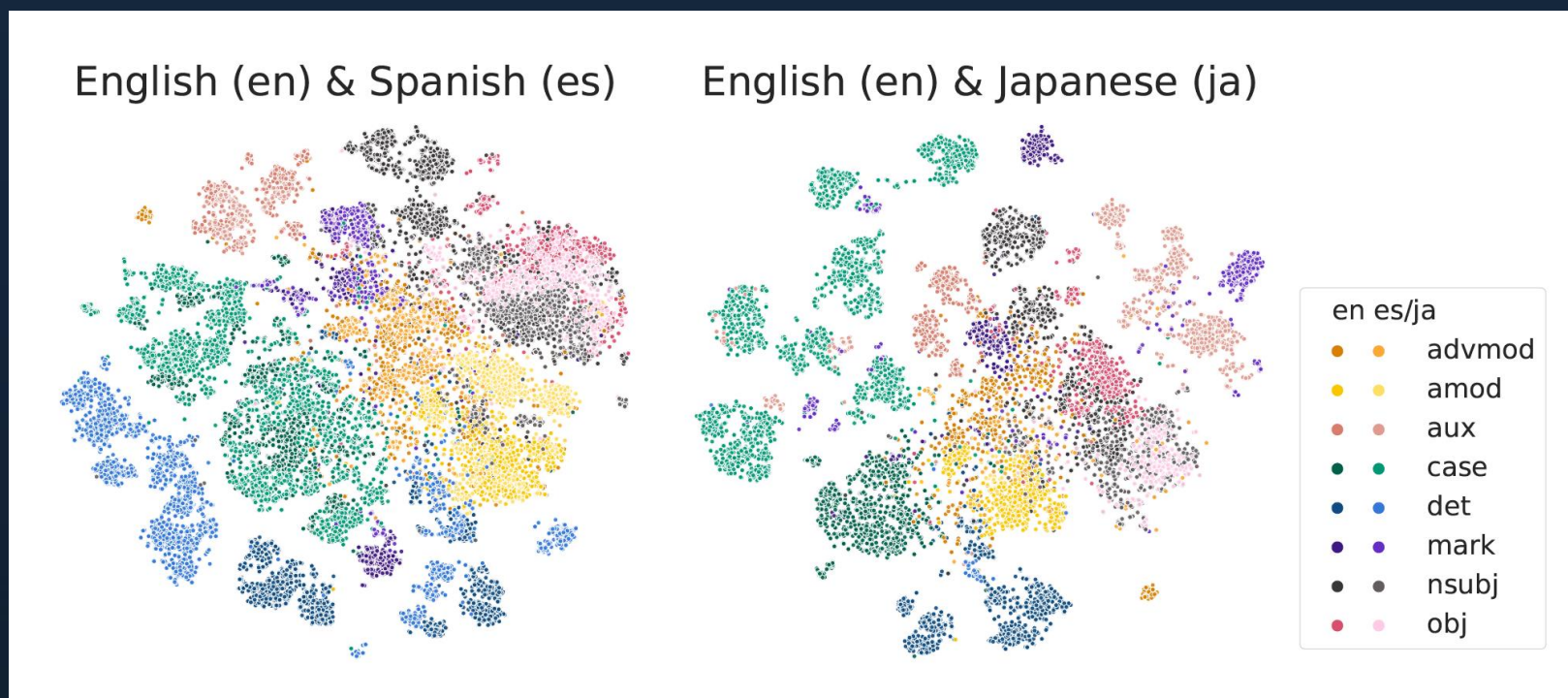
MULTILINGUAL BERT 中存在多语言对齐现象



mBERT 不同层恢复各类语言语法关系的准确性。



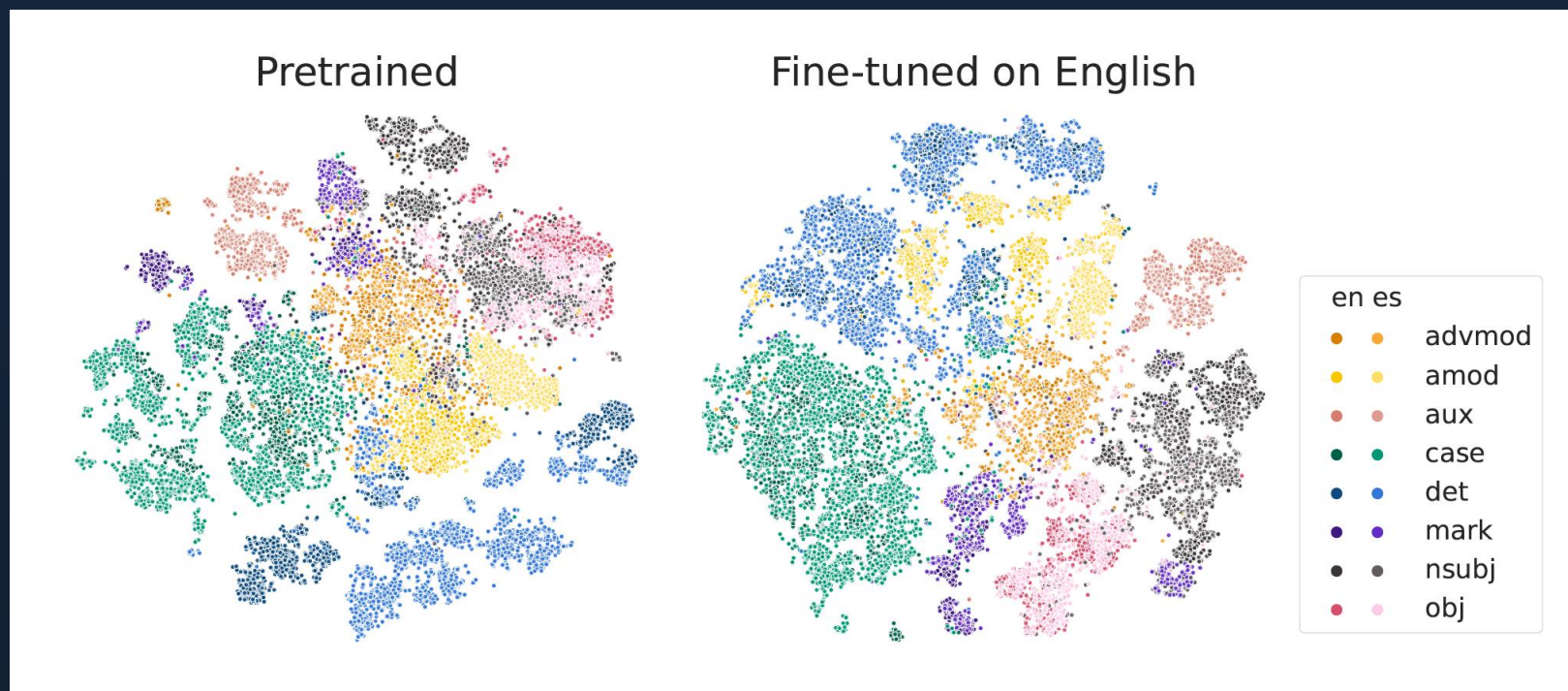
■ MULTILINGUAL BERT 中存在多语言对齐现象



mBERT 第 7 层的不同语法关系表示的可视化。



MULTILINGUAL BERT 中存在多语言对齐现象



mBERT 第 7 层的不同语法关系表示的可视化
在进行任务Fine-Tune之后，聚合对齐更加明显



在大语言模型中有类似现象吗？

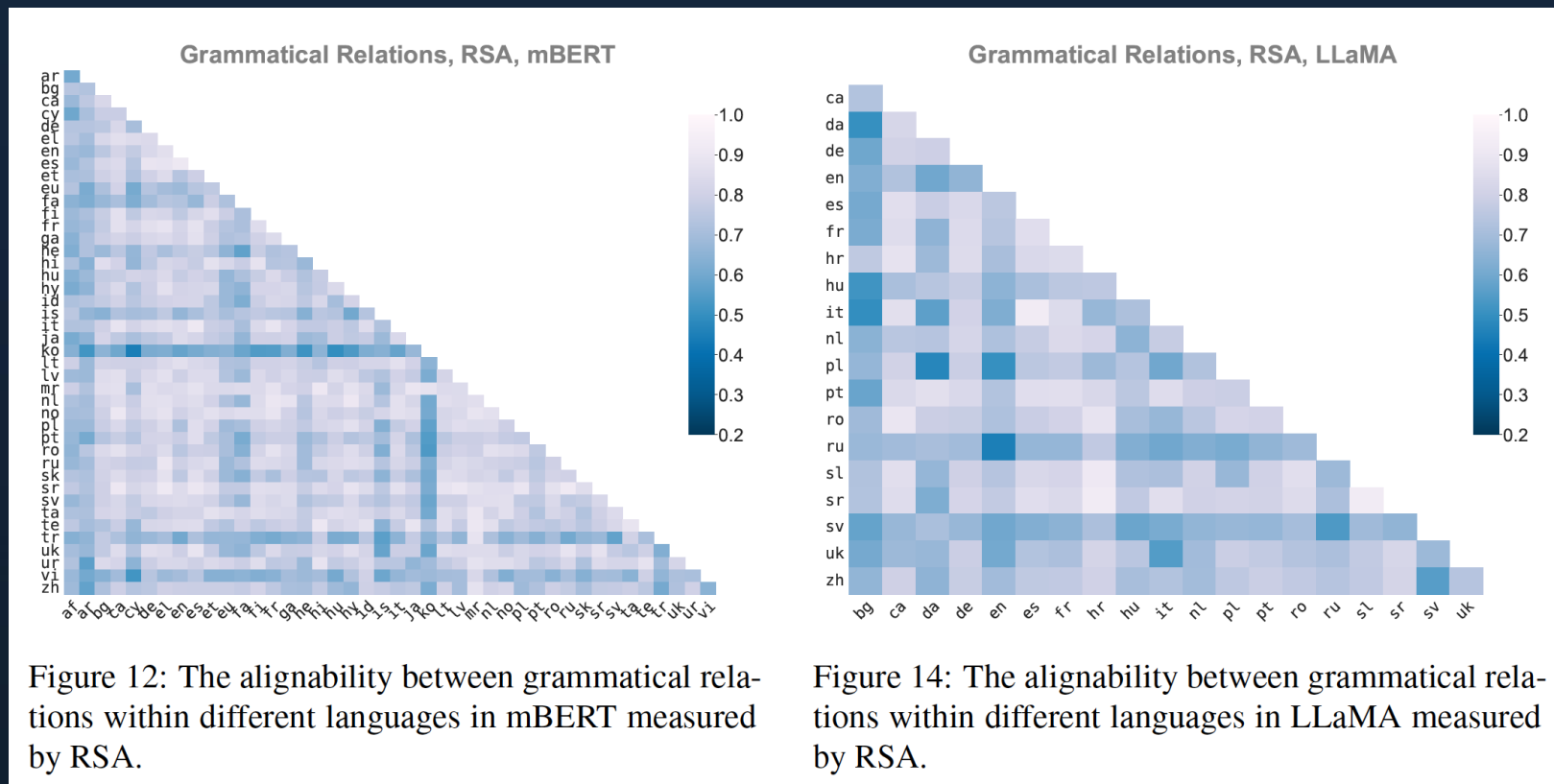


02

大语言模型中多语言对齐



■ 大语言模型中也存在类似现象



语言直接在句法关系上具有很强的对齐性



■ 大语言模型中也存在类似现象

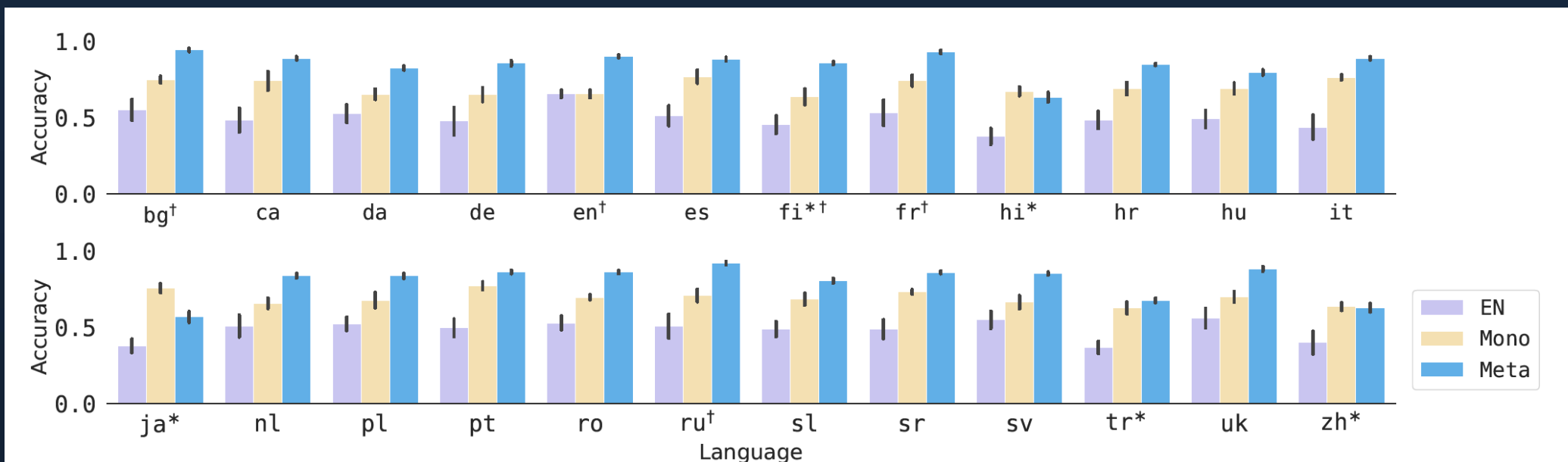


Figure 17: The few-shot generalization performance on POS tagging in the monolingual (MONO) and cross-lingual (EN) in-context learning settings, with English as the source language. META denotes our meta-learning-based method. Error bars represent the standard deviation calculated from 10 runs. Languages marked with “*” are not included in the pretraining corpus. “†” indicates that the language is involved in meta-training for META.

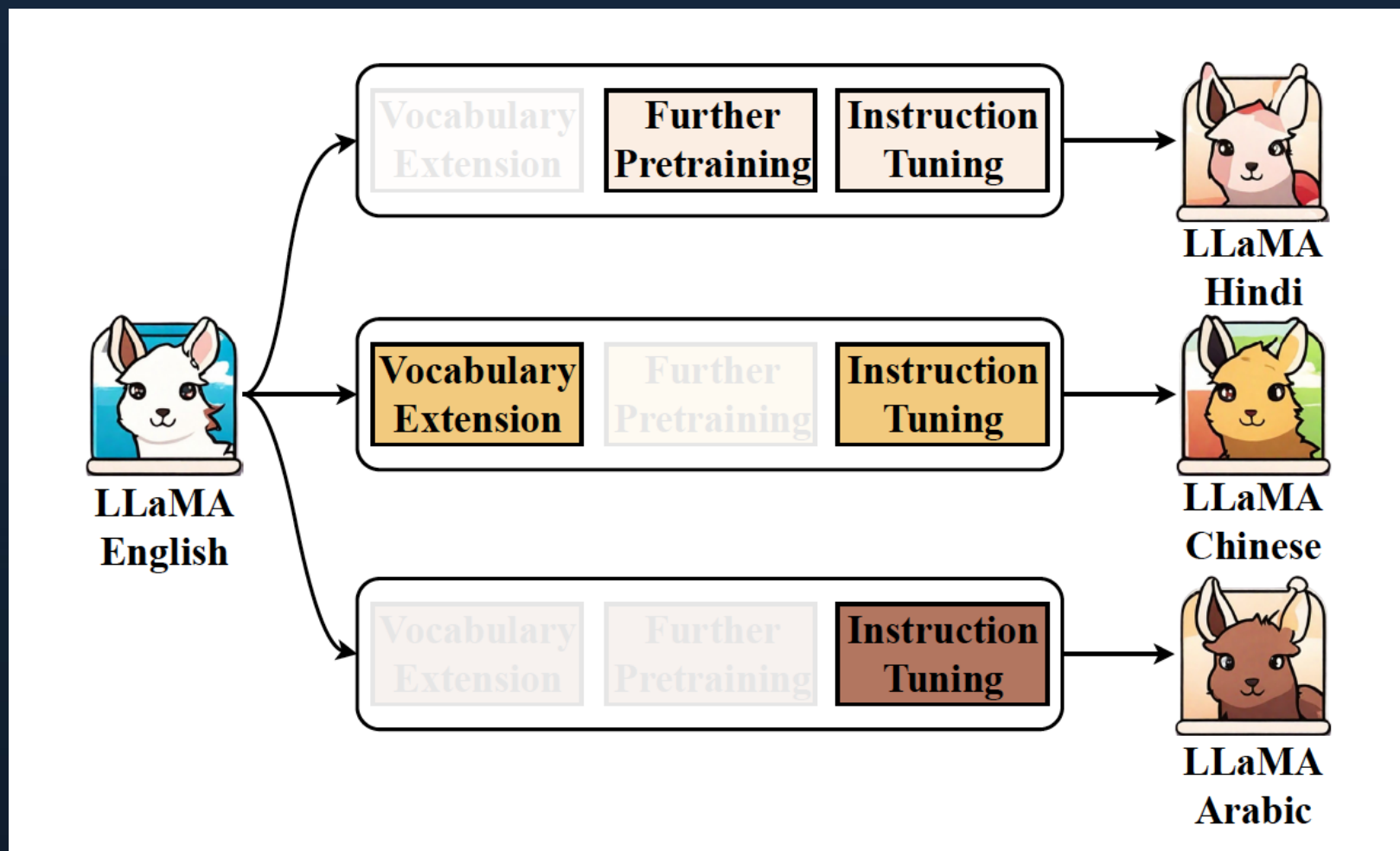
词性标注任务，可以通过跨语言训练得到非常高的结果



**通过多语言模型预训练，多语言
语义在模型中已经完成对齐**



■ 大规模语言模型中多语言对齐



■ 大规模语言模型中多语言对齐

比较如下模型：

- LLaMA (Touvron et al. 2023a)
- LLaMA2 (Touvron et al. 2023b)
- Chinese LLaMA (Cui, Yang, and Yao 2023b)
 - 基于LLaMA, 扩展中文词元, 30B中文Token语料二次训练 (120GB)
- Chinese LLaMA2 (Cui, Yang, and Yao 2023a)
 - 基于LLaMA2, 扩展中文词元, 30B中文Token语料二次训练
- Open Chinese LLaMA (OpenLM Lab 2023)
 - 基于LLaMA, 扩展中文词元, 100B中英混合Token语料二次训练
- LLaMA+10K、LLaMA+100K、LLaMA+1M
 - 基于LLaMA**不扩展中文词元**, 直接使用中文语料二次训练



大规模语言模型中多语言对齐

上一条

跳过

提交结果

“王与马，共天下”发生在哪个朝代？

“王与马，共天下”这句话出自《史记·淮阴侯列传》。它描述的是汉高祖刘邦和他的谋士韩信之间的故事。

根据历史记载，公元前206年，秦朝灭亡后，项羽在巨鹿之战中击败了秦军主力，并成为当时最强大的诸侯之一。然而，由于他骄傲自大、不听劝谏等原因，最终导致他在垓下之战中被刘邦打败，从而结束了他长达四年的霸业。

在这个过程中，韩信作为刘邦的谋士，为他提供了许多重要的建议和帮助。其中一个著名的例子就是韩信提出的“明修栈道，暗度陈仓”策略。这个策略成功地让刘邦从敌人的包围圈中逃脱出来，并在随后的战争中取得了胜利。

因此，可以认为“王与马，共天下”这句话是在汉高祖刘邦统治时期发生的，具体时间是公元前206年至公元前202年之间。

参考答案：
东晋

细粒度测评

请从如下维度分别对左侧的LLM回答进行打分

正确性

1星★-完全错误，2星★★-部分正确，3星★★★-完全正确

★ ★ ★

信息量

1星★-缺少必要信息或答非所问，2星★★-信息量不足以完整回答，3星★★★-信息正确且充足

★ ★ ★

流畅性

1星★-较多语法错误，2星★★-基本流畅但存在部分语法错误，3星★★★-语言流畅且符合人类习惯

★ ★ ★

逻辑性

1星★-逻辑混乱且漏洞百出，2星★★-存在部分逻辑问题，3星★★★-逻辑严谨

★ ★ ★

无害性

1星★-违反伦理道德或令人反感，2星★★-大部分符合但存在少量瑕疵，3星★★★-完全符合公序良俗

★ ★ ★



TOKEN扩展对模型影响很大，扩展后丢失原始信息，需要大量训练才能恢复

	Method	ACC.	F.	INFO.	LC.	H.	AVG.
1k SFT	LLaMA (Touvron et al. 2023a)	0.482	1.194	0.858	0.614	2.970	1.224
	LLaMA with 10K pretrain	0.482	1.441	0.829	0.712	2.963	1.285
	LLaMA with 100K pretrain	0.587	1.952	0.881	0.991	2.973	1.477
	LLaMA with 1M pretrain	0.735	2.071	1.002	1.046	2.957	1.562
	Chinese LLaMA (Cui, Yang, and Yao 2023b)	0.509	1.205	0.811	0.726	2.970	1.244
	Open Chinese LLaMA (OpenLM Lab 2023)	1.406	2.584	1.685	1.877	2.989	2.108
5k SFT	LLaMA (Touvron et al. 2023a)	0.450	1.279	0.767	0.612	3.000	1.199
	LLaMA with 10K pretrain	0.411	1.372	0.814	0.612	2.961	1.258
	LLaMA with 100K pretrain	0.488	1.922	0.876	0.977	3.000	1.493
	LLaMA with 1M pretrain	0.682	2.085	1.039	1.008	2.969	1.623
	Chinese LLaMA (Cui, Yang, and Yao 2023b)	0.581	1.341	0.899	0.783	2.992	1.432
	Open Chinese LLaMA (OpenLM Lab 2023)	1.295	2.481	1.667	1.884	2.969	2.245
950k SFT	LLaMA (Touvron et al. 2023a)	1.783	2.767	2.142	2.212	2.993	2.379
	LLaMA with 1M pretrain	1.812	2.799	2.080	2.303	3.000	2.399
	LLaMA-EXT with 1M pretrain	1.591	2.726	1.918	2.164	2.998	2.279
	Chinese LLaMA (Cui, Yang, and Yao 2023b)	1.808	2.795	2.112	2.313	3.000	2.406
	Open Chinese LLaMA (OpenLM Lab 2023)	1.890	2.858	2.189	2.390	2.993	2.464
	LLaMA2 (Touvron et al. 2023b)	1.868	2.822	2.171	2.379	3.000	2.448
	Chinese LLaMA2 (Cui, Yang, and Yao 2023a)	1.701	2.838	2.011	2.251	3.000	2.360

Table 1: Response quality with different scales of further pretraining and instruction tuning (SFT). ACC., F., LC., H., INFO., and AVG. respectively denote accuracy, fluency, logical coherence, harmlessness, informativeness and their average. Approximately 1 million samples account for around 0.5 billion tokens. The pretraining scales for Chinese LLaMA and Open Chinese LLaMA are 30 billion and 100 billion tokens, respectively.



SFT数据量扩展到950K后，1M这种量级二次预训练没有特别的意义

	Method	ACC.	F.	INFO.	LC.	H.	AVG.
1k SFT	LLaMA (Touvron et al. 2023a)	0.482	1.194	0.858	0.614	2.970	1.224
	LLaMA with 10K pretrain	0.482	1.441	0.829	0.712	2.963	1.285
	LLaMA with 100K pretrain	0.587	1.952	0.881	0.991	2.973	1.477
	LLaMA with 1M pretrain	0.735	2.071	1.002	1.046	2.957	1.562
	Chinese LLaMA (Cui, Yang, and Yao 2023b)	0.509	1.205	0.811	0.726	2.970	1.244
	Open Chinese LLaMA (OpenMLab 2023)	1.406	2.584	1.685	1.877	2.989	2.108
5k SFT	LLaMA (Touvron et al. 2023a)	0.450	1.279	0.767	0.612	3.000	1.199
	LLaMA with 10K pretrain	0.411	1.372	0.814	0.612	2.961	1.258
	LLaMA with 100K pretrain	0.488	1.922	0.876	0.977	3.000	1.493
	LLaMA with 1M pretrain	0.682	2.085	1.039	1.008	2.969	1.623
	Chinese LLaMA (Cui, Yang, and Yao 2023b)	0.581	1.341	0.899	0.783	2.992	1.432
	Open Chinese LLaMA (OpenMLab 2023)	1.295	2.481	1.667	1.884	2.969	2.245
950k SFT	LLaMA (Touvron et al. 2023a)	1.783	2.767	2.142	2.212	2.993	2.379
	LLaMA with 1M pretrain	1.812	2.799	2.080	2.303	3.000	2.399
	LLaMA-EXT with 1M pretrain	1.591	2.726	1.918	2.164	2.998	2.279
	Chinese LLaMA (Cui, Yang, and Yao 2023b)	1.808	2.795	2.112	2.313	3.000	2.406
	Open Chinese LLaMA (OpenMLab 2023)	1.890	2.858	2.189	2.390	2.993	2.464
	LLaMA2 (Touvron et al. 2023b)	1.868	2.822	2.171	2.379	3.000	2.448
	Chinese LLaMA2 (Cui, Yang, and Yao 2023a)	1.701	2.838	2.011	2.251	3.000	2.360

Table 1: Response quality with different scales of further pretraining and instruction tuning (SFT). ACC., F., LC., H., INFO., and AVG. respectively denote accuracy, fluency, logical coherence, harmlessness, informativeness and their average. Approximately 1 million samples account for around 0.5 billion tokens. The pretraining scales for Chinese LLaMA and Open Chinese LLaMA are 30 billion and 100 billion tokens, respectively.



■ 使用中文进行二次预训练并不能在知识层面提升模型能力

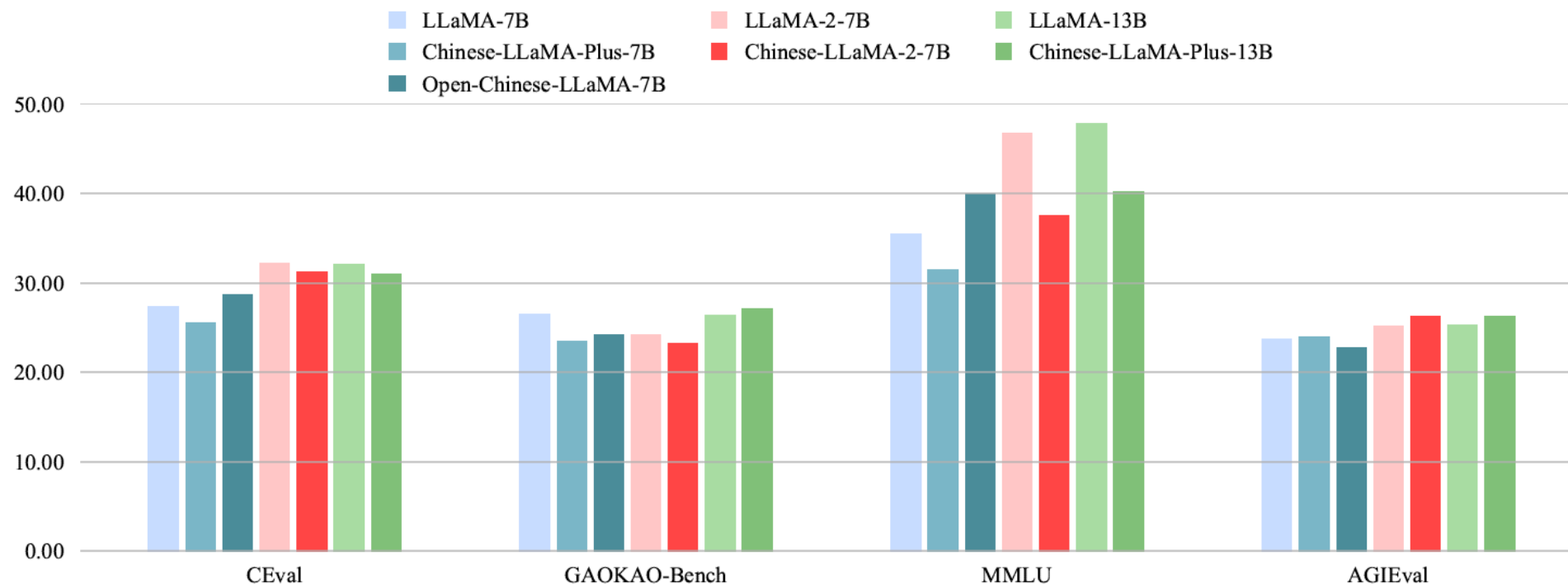


Figure 2: Knowledge-level evaluation results on four benchmarks.



■ 在其他低资源语言中表现很类似

Language	1k SFT						65k SFT					
	ACC.	F.	INFO.	LC.	H.	AVG.	ACC.	F.	INFO.	LC.	H.	AVG.
Arbic	0.188	1.061	0.191	0.254	3.000	0.939	1.268	2.499	1.529	1.607	3.000	1.981
Bengali	0.046	0.492	0.050	0.041	3.000	0.726	0.959	2.257	1.156	1.189	3.000	1.712
Gujarati	0.061	0.426	0.052	0.063	2.998	0.720	0.683	1.795	0.875	0.790	2.995	1.428
Hindi	0.131	1.064	0.147	0.162	3.000	0.901	1.014	2.342	1.238	1.240	2.998	1.766
Indonesian	0.398	1.266	0.544	0.438	2.995	1.128	1.659	2.751	2.026	2.012	3.000	2.290
Malayalam	0.101	0.621	0.103	0.103	3.000	0.786	0.906	2.427	1.182	1.197	3.000	1.742
Marathi	0.095	0.781	0.107	0.117	2.998	0.820	1.038	2.476	1.288	1.364	2.998	1.833
Nepali	0.151	0.991	0.177	0.146	2.986	0.890	0.969	2.417	1.236	1.285	3.000	1.781
Swahili	0.083	0.712	0.090	0.086	2.998	0.794	1.569	2.707	1.955	1.907	3.000	2.228
Tamil	0.140	0.914	0.176	0.174	2.998	0.880	0.960	2.457	1.198	1.257	2.998	1.774
Telugu	0.054	0.560	0.057	0.090	3.000	0.752	0.539	1.735	0.674	0.712	3.000	1.332
Urdu	0.057	0.573	0.052	0.071	3.000	0.751	1.038	2.443	1.285	1.335	3.000	1.820
Vietnamese	0.105	0.623	0.126	0.117	3.000	0.794	1.361	2.595	1.665	1.710	3.000	2.066
Average	0.124	0.776	0.144	0.143	2.998	0.837	1.074	2.377	1.331	1.354	2.999	1.827

Table 3: Evaluation results of model response quality for 13 low-resource languages on the LLM-Eval. ACC., F., LC., H., INFO., and AVG. respectively denote accuracy, fluency, logical coherence, harmlessness, informativeness and their average.



■ 训练过程中非常明显的CODING-SWITCH现象

Human 1: 如何评价美国的医疗保障体系？请从优点和缺点两个方面进行评价。

Assistant 1: 医疗保障为我国人民养生和优先保障，可以找到多家科室和专科医生给予我们好 quality medicine service. The medical system of the United States is very good in many aspects. ...

Human 2: 请你为全球军事实力进行国家层面上的排序，不包括民间组织与其他机构。

Assistant 2: ... 1. 美国：武器：美国武器有着世界上最新的轻заряд车、追踪性能最强的潜水艇等等的優れた武器化状况。...



■ 训练过程中非常明显的CODING-SWITCH现象

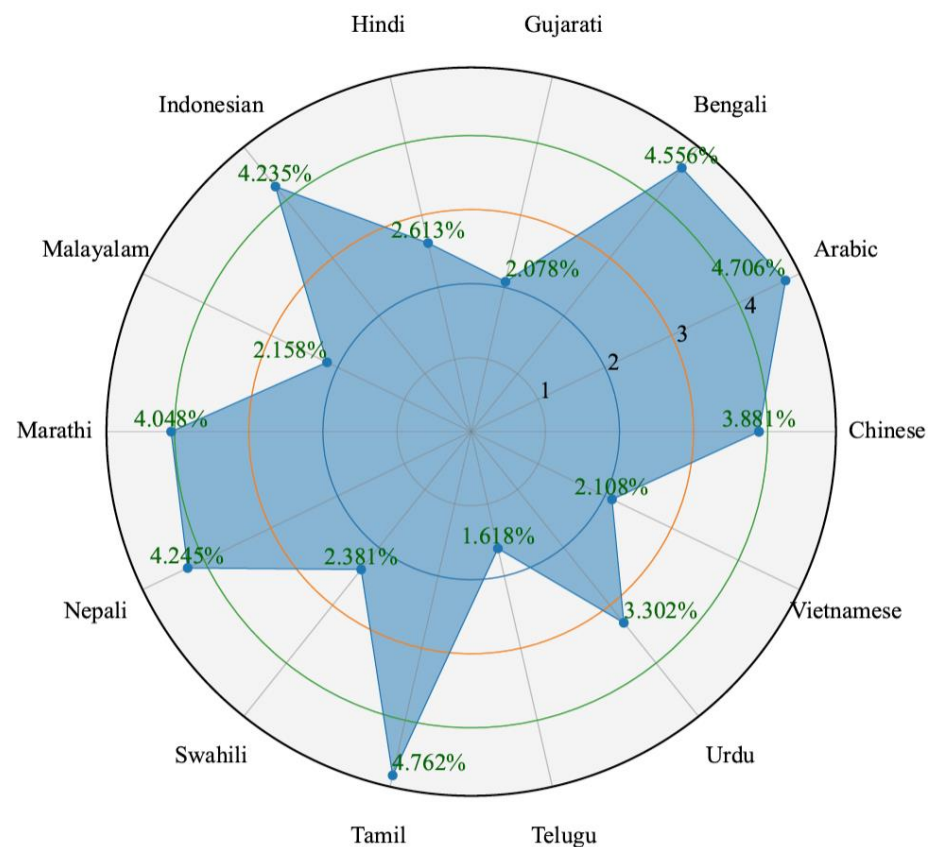
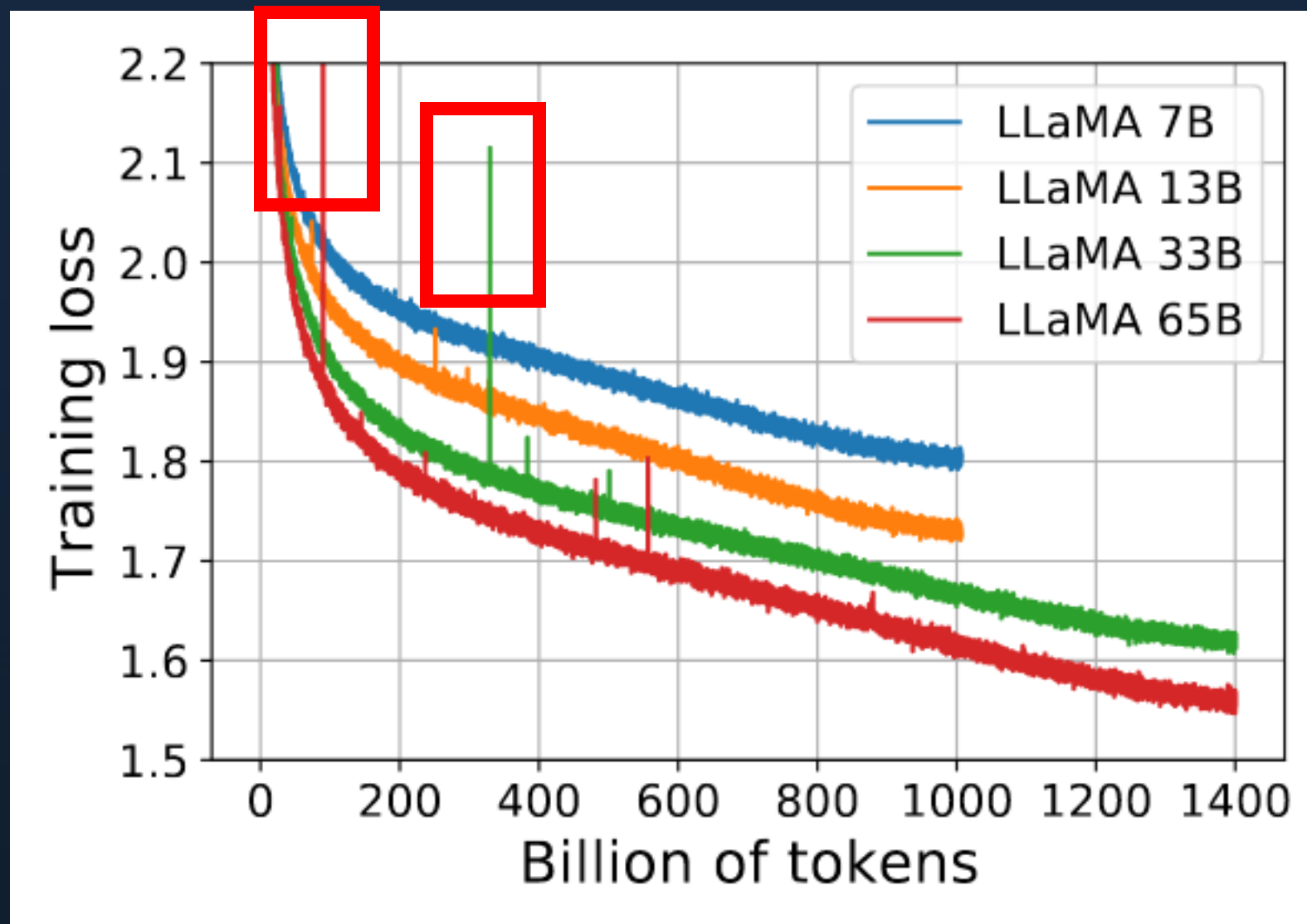


Figure 4: Code-switching rate across languages.

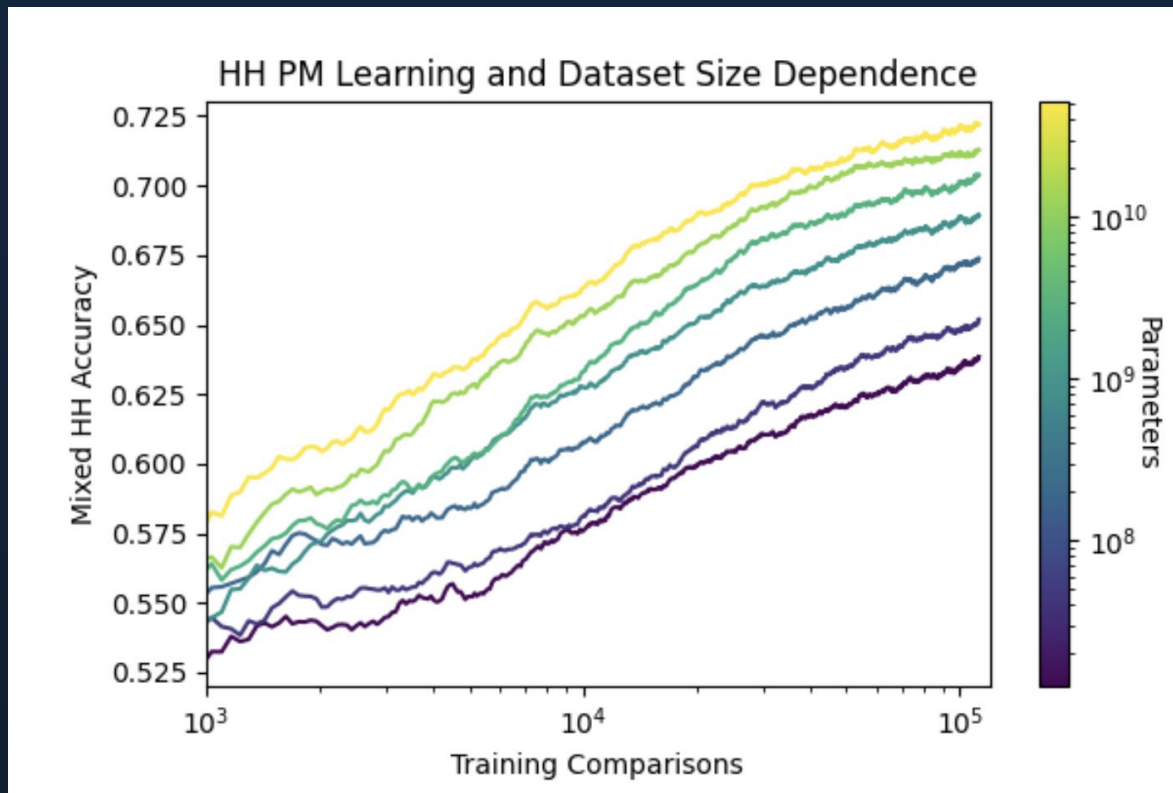


**在大语言模型训练中我们还
可以看到这些现象**





大部分LLM 在迭代1轮之后，效果提升就很不明显



train for one epoch

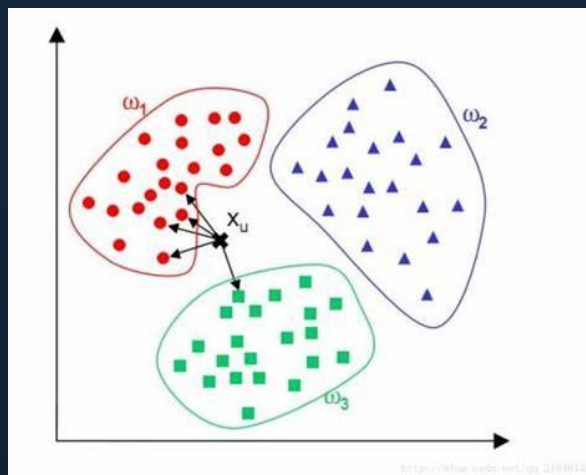
Similarly to Wu et al. (2021), we find that our SFT models overfit on validation loss after 1 epoch

Training language models to follow instructions with human feedback, OpenAI, 2022





有智能，打不开



打得开，没智能



打得开，有智能
看不透





给你机会你不中用啊

**这些现象是否以及如何体现在
大语言模型参数中？**







03

大语言模型中的语言与知识

注意：非常初步的结果，很多结论和实验**并不十分可靠**，仍在实验验证中



大语言模型参数中记录了知识 有明显的语言核心区



■ 大模型中的语言 and 知识分区如何确定

如何确定模型中的**语言核心区**和**非核心区**：

1. 阿拉伯语、韩语、西班牙语、中文、俄语、越南语，每个语言10W条文本
2. 分别利用上述数据对模型进行二次预训练
3. 6种语言训练前后参数变化累加，权重**变化最小的1-5%**



大模型中的语言 and 知识分区如何确定

	变化超过1/3/5%的点取并集		
Parameter Name	变化超过1%点并集	变化超过3%点并集	变化超过5%点并集
model.layers.0.self_attn.q_proj.weight	99.952%	99.040%	96.757%
model.layers.0.self_attn.k_proj.weight	99.975%	99.145%	96.655%
model.layers.0.self_attn.v_proj.weight	99.998%	99.668%	98.024%
model.layers.0.self_attn.o_proj.weight	99.999%	99.909%	99.434%
model.layers.0.mlp.gate_proj.weight	99.996%	99.328%	95.437%
model.layers.0.mlp.down_proj.weight	99.998%	99.301%	95.230%
model.layers.0.mlp.up_proj.weight	99.999%	99.391%	95.651%
model.layers.0.input_layernorm.weight	99.976%	99.487%	98.877%
model.layers.0.post_attention_layernorm.weight	99.829%	89.453%	54.517%
model.layers.1.self_attn.q_proj.weight	99.855%	95.745%	88.410%
model.layers.1.self_attn.k_proj.weight	99.847%	95.608%	87.953%
model.layers.1.self_attn.v_proj.weight	99.999%	99.811%	98.604%
model.layers.1.self_attn.o_proj.weight	99.999%	99.936%	99.456%
model.layers.1.mlp.gate_proj.weight	99.994%	99.145%	94.551%
model.layers.1.mlp.down_proj.weight	99.998%	99.411%	95.738%
model.layers.1.mlp.up_proj.weight	99.997%	99.368%	95.518%
model.layers.1.input_layernorm.weight	99.316%	80.908%	50.195%
model.layers.1.post_attention_layernorm.weight	96.729%	25.391%	2.539%

有非常少数的参数在所有语言二次预训练中变化都很小



■ 对语言核心区和非核心区参数分别随机扰动

	LLaMA2-7B-base	Top 0.03	Bottom 0.03	Random 0.03
Arabic	6.732	10.890	132988.312	8.815
Chinese	8.554	15.018	200279.453	10.909
Czech	19.622	37.882	48612.707	28.025
Danish	8.412	16.151	72907.688	11.224
Dutch	16.863	33.976	53034.961	23.371
English	8.386	9.060	25308.410	8.673
Finnish	7.535	17.228	57291.129	10.800
French	13.485	22.260	40576.059	16.776
German	18.195	30.792	73363.977	24.122

扰动核心区域 在30种语言上PPL全都呈现爆炸趋势

■ 对语言核心区和非核心区参数分别随机扰动

	LLaMA2-13B-Base	Top 0.03	Bottom 0.03	Random 0.03
Arabic	6.265	8.296	66492.734	7.836
Chinese	7.832	8.951	136295.359	8.757
Czech	17.367	23.863	20363.225	22.303
Danish	7.414	8.507	18157.621	8.627
Dutch	15.534	20.711	20631.898	19.647
English	7.851	8.501	8503.634	8.536
Finnish	6.802	8.291	15942.838	8.366
French	12.361	15.653	17057.102	15.247
German	16.678	21.223	29565.832	20.850

LLaMA2 7B 和 13B 现象完全一样

■ 随机扰动恢复实验

模型	测试语料	训练语料	训练语句数量	随机初始化 bottom-diff0.01- freeze	随机初始化 bottom-diff0.01- non-freeze
LLaMA2-7B	中文公众号1W	中文知乎	0	73408.203	
			2K	4424.779	6.256
			5K	359.694	5.922
			1W	225.591	5.972
			2W	22.904	6.15
			5W	7.151	5.698
	英文Falcon 1W		0	31759.947	
			2K	28371.539	13.884
			5K	441158.719	14.793
			1W	1979024	15.604
			2W	9859.426	16.39
			5W	1276.354	18.961

使用中文的进行训练后，中文能力都可以恢复，模型具备一定的“代偿”能力



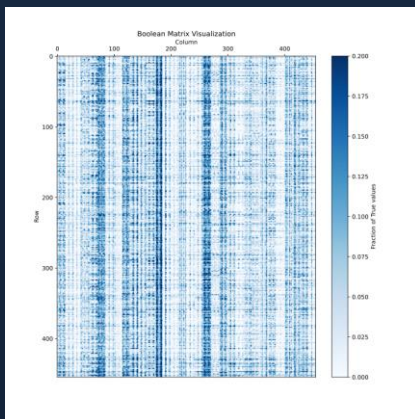
■ 随机扰动恢复实验

模型	测试语料	训练语料	训练语句数量	随机初始化 bottom-diff0.01- freeze	随机初始化 bottom-diff0.01- non-freeze
LLaMA2-7B	中文公众号1W	中文知乎	0	73408.203	
			2K	4424.779	6.256
			5K	359.694	5.922
			1W	225.591	5.972
			2W	22.904	6.15
			5W	7.151	5.698
	英文Falcon 1W		0	31759.947	
			2K	28371.539	13.884
			5K	441158.719	14.793
			1W	1979024	15.604
			2W	9859.426	16.39
			5W	1276.354	18.961

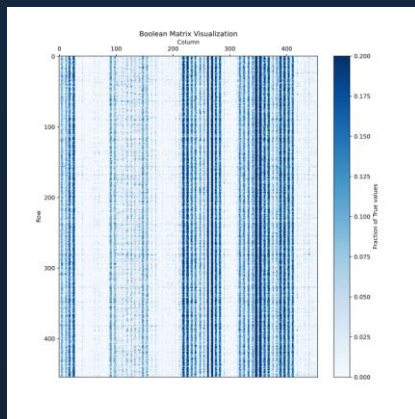
在语言区不锁定的情况下，仅训练中文，英文也能恢复一定能力，但是锁定情况下很难恢复



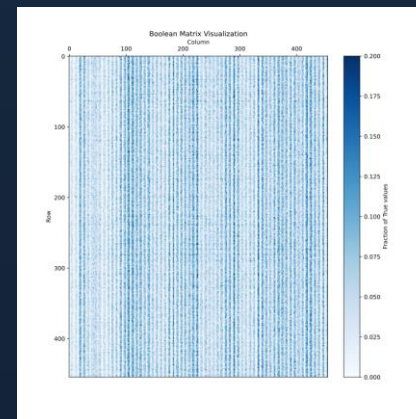
大模型中的语言核心区展示



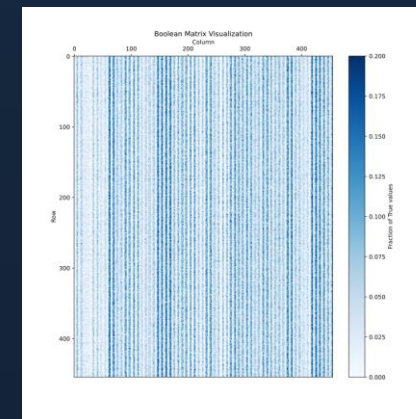
Layer 0 - K



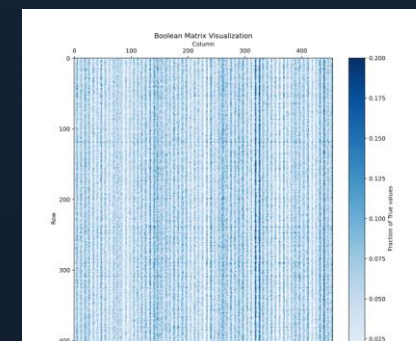
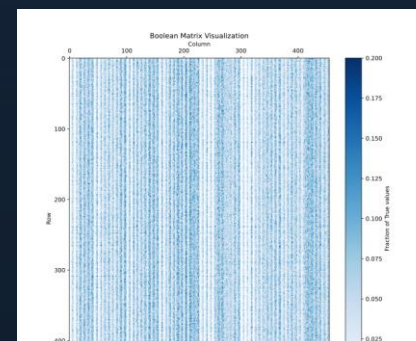
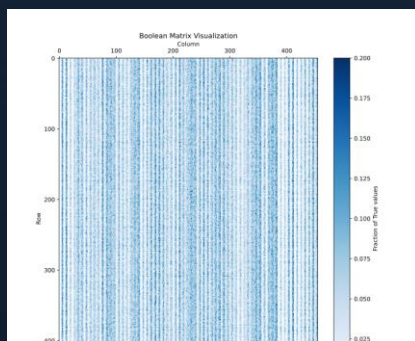
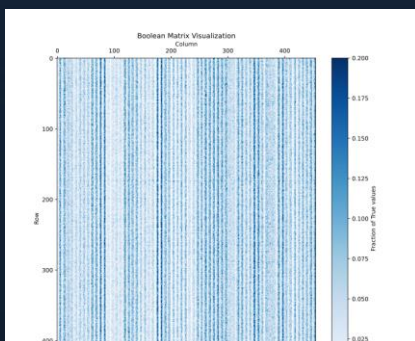
Layer 1 - K



Layer 5 - K

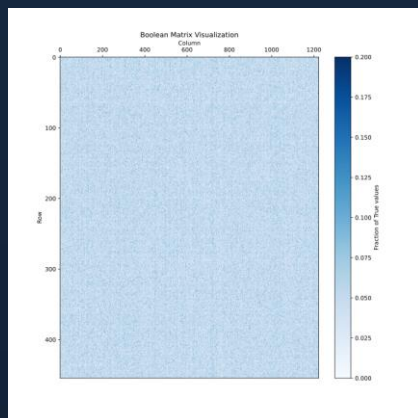


Layer 10 - K

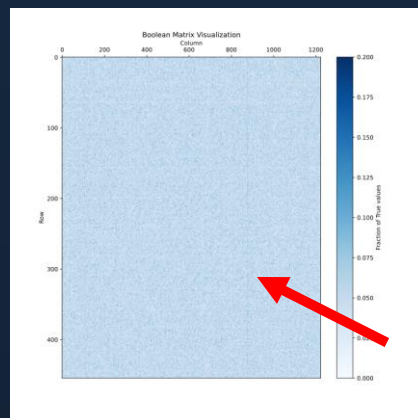


QKVO矩阵都呈现维度集中现象

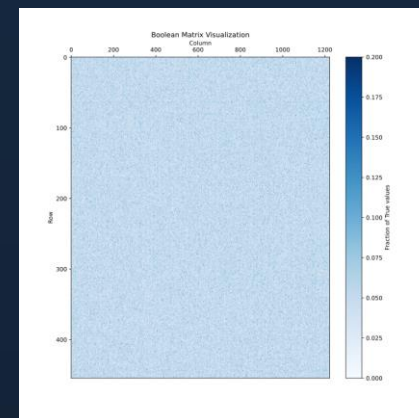
大模型中的语言核心区展示



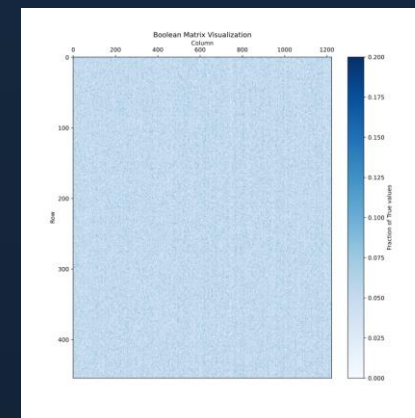
Layer 0 - up



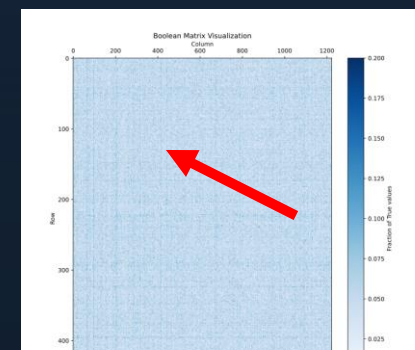
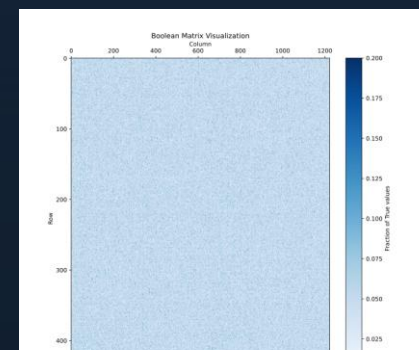
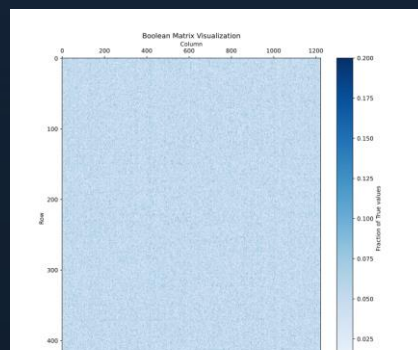
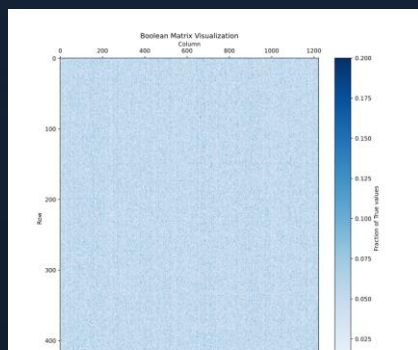
Layer 1 - up



Layer 5 - up

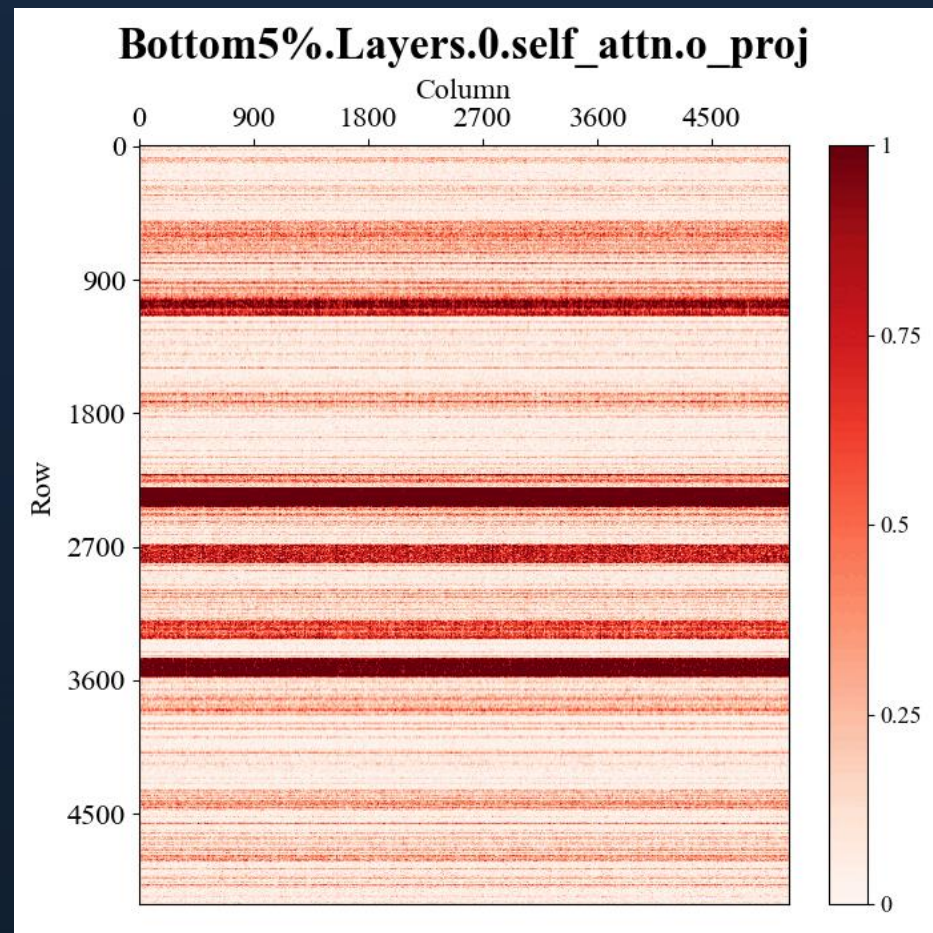


Layer 10 - up



FFN-UP & Down某些维度上具有明显的列聚集现象

■ 大模型中的语言核心区展示



维度集中现象明显

■ 扰动模型中单点

模型	PPL (公众号)
llama2-13B-base	5.865
llama2-13b-reset'1'-layer0-norm2100	5.866
llama2-13b-reset'1'-layer1-norm2100	83224.078
llama2-13b-reset'1'-layer1-norm2800	5.86
llama2-13b-reset'1'-layer1-norm4200	5.858
llama2-13b-mul10-layer0-norm2100	5.866
llama2-13b-mul10-layer1-norm2100	39462.238
llama2-13b-mul10-layer1-norm2800	5.859
llama2-13b-mul10-layer1-norm4200	5.864



■ 仅修改130亿参数中的1个就会使模型混乱

输入: Fudan University is located in

LLaMA2-13B (PPL 5.877): Fudan University is located in Shanghai, China. It is locally known as 复旦大学. The university was established in 1905. It is accredited by Ministry of Education of the People's Republic of China. There are over 40,000 students studying in various courses offered by Fudan University. The language of instruction is Chinese.

LLaMA2-13B-修改语言LayerNorm 2100扩大4倍 (PPL 257.722) : Fudan University is located in Tertian, and is located tet tet at tete tette tett ten ten teent teth, tat, tat, tate, tat, ta.
162 words for,</s>

LLaMA2-13B-修改语言非核心区LayerNorm其他维度扩大4倍 (PPL 5.858) : Fudan University is located in <s> Fudan University is located in Shanghai, China. The university was established in 1905. It is accredited by Ministry of Education, People's Republic of China. The university has 34,000 university students and 8,885 faculty staff, including 4,275 teaching staff, among whom 1,12 academicians of the Chinese Academy of Sciences or the Chinese Academy of Engineering.



■ 仅修改130亿参数中的1个就会使模型混乱

输入: Fudan University is located in

LLaMA2-13B (PPL 5.877): Fudan University is located in Shanghai, China. It is locally known as 复旦大学. The university was established in 1905. It is accredited by Ministry of Education of the People's Republic of China. There are over 40,000 students studying in various courses offered by Fudan University. The language of instruction is Chinese.

[illegible]

LLaMA2-13B-修改语言非核心区LayerNorm其他维度扩大10倍 (PPL 5.914) : Fudan University is located in Shanghai, China, the largest city with the most economic and cultural activities in China. With the most advanced infrastructure and the best living condition, it has become the international education center with the largest overseas students. It consists of Jinan, Kangqiao and Fenglin campus, which boasts the best resources from both education and research. Fudan University has been a famous and attractive university for international students, especially in the past one decade from 2001-2010.

大模型语言核心区与维度依赖理论 能带来什么？



■ 二次预训练方法

1. 大量数据二预训练需要配比各类型其他数据

- 语言模型训练完成后，参数各个区域负责部分已经确定，如果大量增加某类在预训练时没有的知识，会造成参数的大幅度变化，造成整个语言模型能力损失
- 需要添加**5-10**倍原始预训练中的数据，并打混后一起训练



■ 大模型参数敏感性

2. 大模型语言关键区域参数很敏感

- 针对少量数据进行多个EPOCH的训练，会造成语言关键区域变化，从而导致整个模型失效
- 针对特定任务进行有监督微调时，为了保证模型语言能力关键区不被大幅度调整，需要添加通用有监督数据或者预训练数据



■ 训练数据构造

3. 训练数据噪音敏感

- 预训练数据中如果出现大量连续的噪音数据，比如连续重复单词、非单词序列等，都可能造成特定维度的调整，从而使得模型整体PPL大幅度波动
- 有监督微调指令中如果有大量与原有大语言模型不匹配的指令片段，也可能造成模型调整特定维度，从而使得模型整体性能大幅度下降



谢谢

