



Efficiently Democratizing Medical LLMs for 50 Languages via a Mixture of Language Family Experts

2024.11.29

Xidong Wang

School of Data Science

Chinese University of Hong Kong, Shenzhen





Outline

- Introduction (5 min)
- A preliminary scaling to 12 languages (5 min)
- Scaling with MoE and its routing analysis (10 min)
 - A language-specific hybrid routing
 - Interpretable routing analysis: information flow circuit
 - Post-MoE variant inspired by the ‘Spread Out in the End’ Phenomenon
- Further Scaling to 50 Languages (10 min)
 - Mixture of language family experts
 - Results and Ablation Study
- Q&A (5 min)





Outline

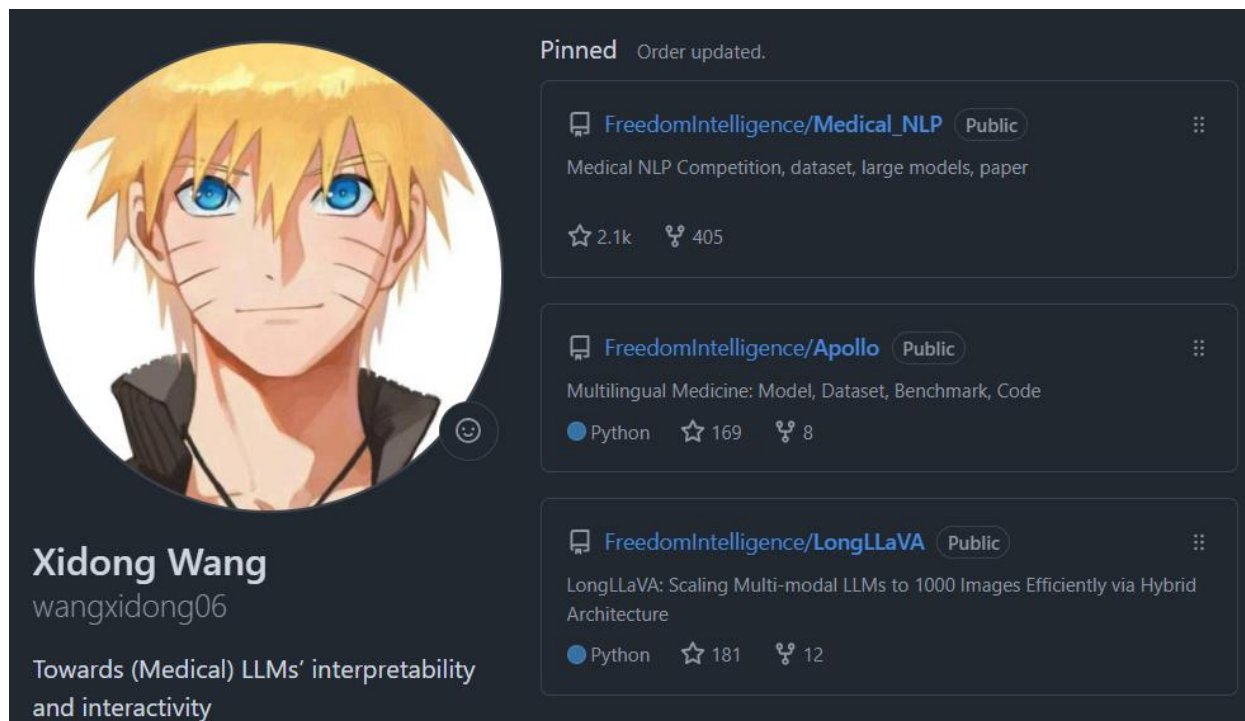
- Introduction (5 min)
- A preliminary scaling to 12 languages (5 min)
- Scaling with MoE and its routing analysis (10 min)
 - A language-specific hybrid routing
 - Interpretable routing analysis: information flow circuit
 - Post-MoE variant inspired by the ‘Spread Out in the End’ Phenomenon
- Further Scaling to 50 Languages (10 min)
 - Mixture of language family experts
 - Results and Ablation Study
- Q&A (5 min)





Self Background Introduction

- A Second-Year LLM-Traing Guy~
- Focus on (Medical) LLM's Interactivity
- CMB, Apollo, LongLLaVA, ApolloMoE ...



Xidong Wang
wangxidong06

Towards (Medical) LLMs' interpretability and interactivity

Pinned Order updated.

- FreedomIntelligence/Medical_NLP** Public
Medical NLP Competition, dataset, large models, paper
2.1k stars, 405 forks
- FreedomIntelligence/Apollo** Public
Multilingual Medicine: Model, Dataset, Benchmark, Code
Python, 169 stars, 8 forks
- FreedomIntelligence/LongLLaVA** Public
LongLLaVA: Scaling Multi-modal LLMs to 1000 Images Efficiently via Hybrid Architecture
Python, 181 stars, 12 forks

TITLE	CITED BY	YEAR
CMB: A comprehensive medical benchmark in chinese X Wang, GH Chen, D Song, Z Zhang, Z Chen, Q Xiao, F Jiang, J Li, X Wan, ... NAACL 2024	59	2023
Huatuo-26m, a large-scale chinese medical qa dataset X Wang, J Li, X Wu, Z Zhang, X Xu, J Fu, P Tiwari, X Wan, B Wang arXiv preprint arXiv:2305.01526	49	2023
Huatuoqpt-ii, one-stage training for medical adaption of llms J Chen, X Wang, A Gao, F Jiang, S Chen, H Zhang, D Song, W Xie, ... COLM 2024	47 *	2023
Apollo: A Lightweight multilingual medical llm towards democratizing medical ai to 6b people X Wang, N Chen, J Chen, Y Hu, Y Wang, X Wu, A Gao, X Wan, H Li, ... arXiv preprint arXiv:2403.03640	20	2024
Huatuoqpt-vision, towards injecting medical visual knowledge into multimodal llms at scale J Chen, R Ouyang, A Gao, S Chen, GH Chen, X Wang, R Zhang, Z Cai, ... EMNLP 2024	15	2024
Mllm-bench, evaluating multi-modal llms using gpt-4v W Ge, S Chen, G Chen, J Chen, Z Chen, S Yan, C Zhu, Z Lin, W Xie, ... arXiv preprint arXiv:2311.13951	15 *	2023
LongLLaVA: Scaling Multi-modal LLMs to 1000 Images Efficiently via a Hybrid Architecture X Wang, D Song, S Chen, C Zhang, B Wang arXiv preprint arXiv:2409.02889	6	2024





Why Multilingual Medical is so important?

Serve the community with more warmth

 **Pooya Jazayeri** · Follow
Board Certified Anesthesiologist · 6y

For me, it has been critical. I spent some time working in a French speaking hospital in Africa. Prior to that, I was conversational in French. I knew enough to ask directions and order in French restaurants. I didn't know any medical French. So words like laryngoscope, malaria, transfusion, et cetera, were a big challenge for me.

对我来说，这至关重要。我在非洲的一家法语医院工作了一段时间。在此之前，我用法语进行对话。我知道如何在法国餐馆问路和点餐。我不懂任何医学法语。所以像喉镜、疟疾、输血等等这些词对我来说是一个很大的挑战。

Now, I work in a very diverse community in northern California. I use my medical Spanish frequently, but I use the Farsi I grew up speaking even more.

现在，我在北加州一个非常多元化的社区工作。我经常使用医学西班牙语，但我更多地使用我从小就讲的波斯语。

Upvote · 14

 **ComisiynyddyGymraeg**
@ComyGymraeg

“Pan o'n i'n trafod iselder neu gorbryder gyda'r GP yn Saesneg, doedd o ddim yn swnio fel llais fi. Oedd deud hynna'n Saesneg ddim yn swnio fel profiad fi.”

“当我用英语与全科医生讨论抑郁症或焦虑症时，听起来不像我的声音。用英语说这句话听起来不像我的经历。”

Mae gwasanaethau Cymraeg ym maes iechyd a gofal yn sicrhau bod cleifion yn teimlo eu bod yn cael eu clywed a'u parchu.

健康和护理领域的威尔士语言服务确保患者感到自己的声音得到倾听和尊重。





What is the current capability of large models?

现病史\n (1) 病史摘要\n 病人，男，49岁，3小时前解大便后出现右下腹痛，右下腹可触及一包块，既往体健。 \n (2) 主诉\n 右下腹痛并自扪及包块3小时。 \n \n 体格检查\n 体温： T 37.8℃， P 101次 / 分，呼吸22次/分，BP 100/60mmHg，腹软，未见胃肠型蠕动波。 \n 辅助检查\n (1) 实验室检查\n 血常规： WBC 5.0×10⁹ / L， N 78%。 \n 尿常规正常。 \n (2) 多普勒超声检查\n 沿腹股沟纵切可见一多层分布的混合回声区，宽窄不等，远端膨大，边界整齐，长约4~5cm。

Hindi

简述该病人的诊断及诊断依据。



यह मरीज पेट के दाहिने निचले हिस्से में दर्द और एक स्पर्शनीय द्रव्यमान के साथ आ रहा है, जो तीन घंटे से उत्पन्न हुआ है। शारीरिक परीक्षण में तापमान 37.8°C और पल्स 101/मिनट पाया गया, जो संक्रमण या सूजन को संकेत करता है। डॉपलर अल्ट्रासोनिक परीक्षण में बहुस्तरीय मिश्रित प्रतिध्वनि क्षेत्र और पेट की एक्स-रे में द्रव-वायु स्तर का पाया जाना आंतरिक अंगों की स्थिति में असमानता को दर्शाता है। यह सब एपेंडिसाइटिस, पेरिटोनाइटिस, या किसी अन्य आंतरिक संक्रमण या सूजन जैसे रोगों का संकेत हो सकता है। क्लिनिकल और प्रयोगशाला परिणामों को देखते हुए, संभावित निदान पेट की एपेंडिसाइटिस या पेरिटोनाइटिस हो सकता है, जिसे और अधिक परीक्षणों की आवश्यकता हो सकती है।

该患者出现右下腹疼痛和可触及的肿块，这种情况已经发生了三个小时。体检发现体温37.8℃，脉搏101次/分钟，提示感染或炎症。多普勒超声检查多层混合回声区和腹部X线液气层提示内脏器官位置异常。所有这些都可能是阑尾炎、腹膜炎或任何其他内部感染或炎症等疾病的征兆。根据临床和实验室结果，可能的诊断可能是腹部阑尾炎或腹膜炎，可能需要进一步检查。

English



Diagnosis:诊断:

The patient is likely suffering from **acute inguinal hernia with incarceration**, potentially with signs of bowel obstruction.患者可能患有**急性腹股沟疝并伴有嵌顿**，可能有肠梗阻的迹象。

Diagnostic Basis:诊断依据:

1. **Clinical Presentation:** The patient, a 49-year-old male, presents with right lower abdominal pain and a palpable mass in the right lower abdomen for 3 hours. The pain was associated with defecation, and the mass was tender, located above the inguinal ligament, and of uncertain consistency, which is suggestive of a hernia. The acute onset of pain and mass palpation is characteristic of an incarcerated hernia.

临床表现: 患者，49岁男性，出现右下腹疼痛，右下腹可触及肿块3小时。疼痛与排便有关，肿块触痛，位于腹股沟韧带上方，质地不确定，提示疝气。急性发作的疼痛和肿块触诊是嵌顿疝的特征。

2. Physical Examination:体格检查:

- The presence of a round mass (4cm×4cm) in the right inguinal area, which is tender and has unclear borders, is highly suggestive of an inguinal hernia.
右侧腹股沟处出现圆形肿块（4cm×4cm），有压痛，边界不清，高度提示为腹股沟疝。
- Vital signs indicate mild systemic stress (fever of 37.8°C, tachycardia at 101 beats/min), which might be a sign of infection or bowel strangulation within the hernia sac.

生命体征表明轻度全身应激（发烧37.8℃ ↓，心动过速101次/分钟），这可能是疝囊内感染或肠绞窄的迹象。





Outline

- Introduction (5 min)
- A preliminary scaling to 12 languages (5 min)
- Scaling with MoE and its routing analysis (10 min)
 - A language-specific hybrid routing
 - Interpretable routing analysis: information flow circuit
 - Post-MoE variant inspired by the ‘Spread Out in the End’ Phenomenon
- Further Scaling to 50 Languages (10 min)
 - Mixture of language family experts
 - Results and Ablation Study
- Q&A (5 min)





Always Start with Data ! ! !

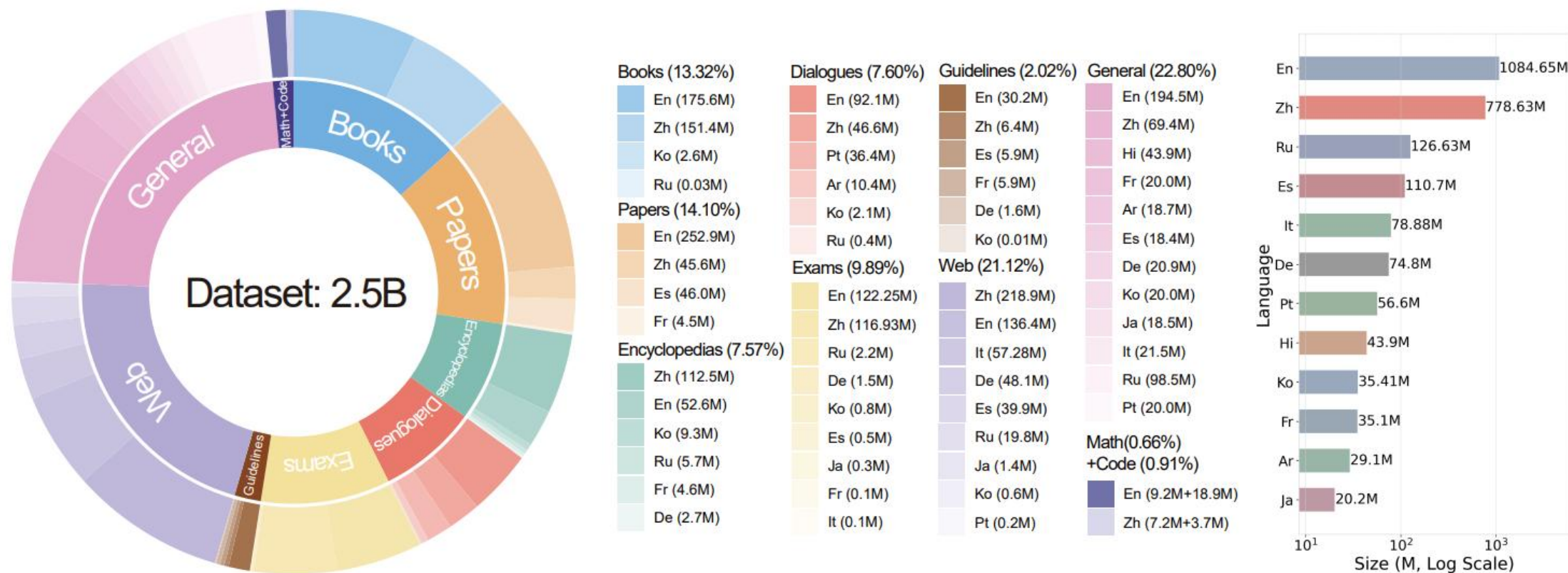


Figure 1: Taxonomy and Token statistics of Training Dataset.

In order to adapt to more rare languages, the difficulty of data collection has increased sharply.





Preliminary results

Table 1: Monolingual (■) refers to the accuracy of models trained on single language data and evaluated on the respective language evaluation set. Multilingual (■) indicates the accuracy of models trained and evaluated on all languages' datasets. The -Math&Code results reflect performance when training on all language data (■) excluding math and code data. Avg. denotes the average score across languages.

Model	#params	Avg.	Ar	De	En	Es	Fr	Hi	It	Ja	Ko	Pt	Ru	Zh
Gemma	2B	28.1	18.2	27.6	37.6	31.8	23.1	25.7	26.1	21.2	25.2	23.1	49.6	28.0
+Monolingual	2B	-	27.8	34.9	52.1	39.2	27.7	27.6	28.6	23.4	29.9	28.3	54.7	60.6
+Multilingual	2B	48.6	43.7	50.7	57.8	48.1	44.5	40.7	45.2	43.0	45.1	40.6	63.7	60.6
-Math&Code	2B	42.7	36.1	39.4	51.2	43.7	39.3	35.4	42.0	35.4	40.5	32.1	57.0	60.0
GPT-4o mini	-	74.1	80.7	70.5	82.3	84.7	75.2	77.7	76.4	77.2	82.2	80.9	67.9	78.2
Open-source Models														
MMed-Llama-3	8B	29.5	40.3	54.7	63.6	62.0	38.8	20.7	47.3	20.3	25.9	62.5	16.7	40.2
OpenBioLLM	8B	27.5	58.1	49.6	59.2	53.3	44.4	41.5	39.1	46.7	27.0	64.8	41.8	46.1
Llama3	8B	33.5	55.9	48.4	60.5	58.4	48.0	39.9	38.9	49.5	51.5	63.3	43.6	49.3

Multilingual Training Helps! Math & Code data Helps! Existing Model sucks!





Outline

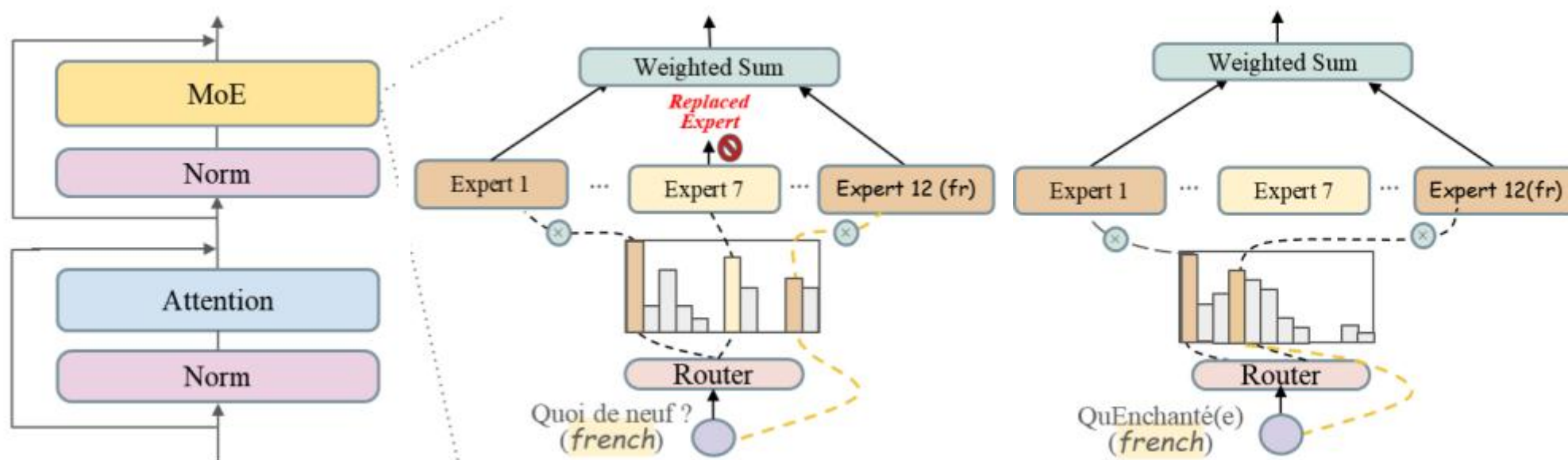
- Introduction (5 min)
- A preliminary scaling to 12 languages (5 min)
- Scaling with MoE and its routing analysis (10 min)
 - A language-specific hybrid routing
 - Interpretable routing analysis: information flow circuit
 - Post-MoE variant inspired by the ‘Spread Out in the End’ Phenomenon
- Further Scaling to 50 Languages (10 min)
 - Mixture of language family experts
 - Results and Ablation Study
- Q&A (5 min)





A language-specific hybrid routing

Mechanism -> Interpretability and Generalization -> MoE **Modularity**



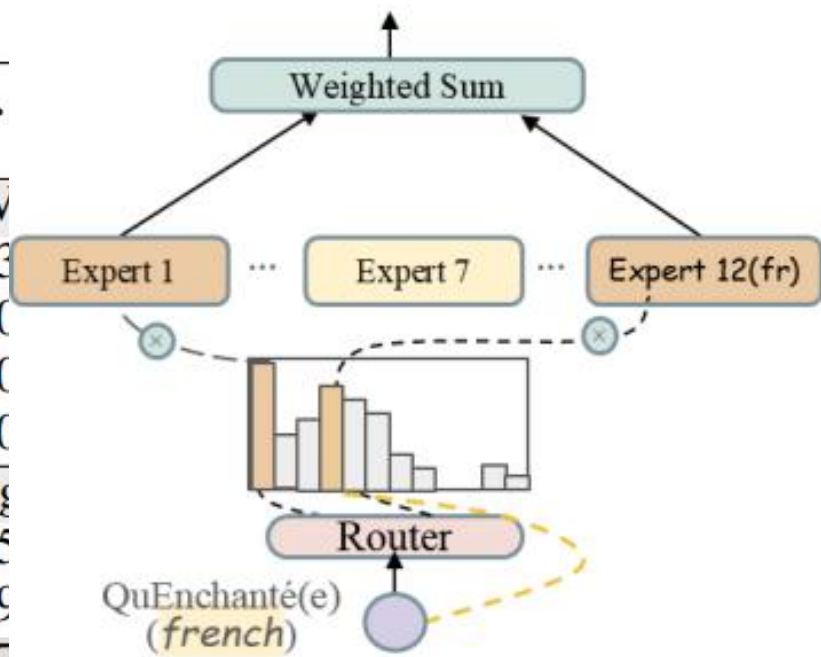


A language-specific hybrid routing

Mechanism -> Interpretability and Generalization -> MoE **Modularity**

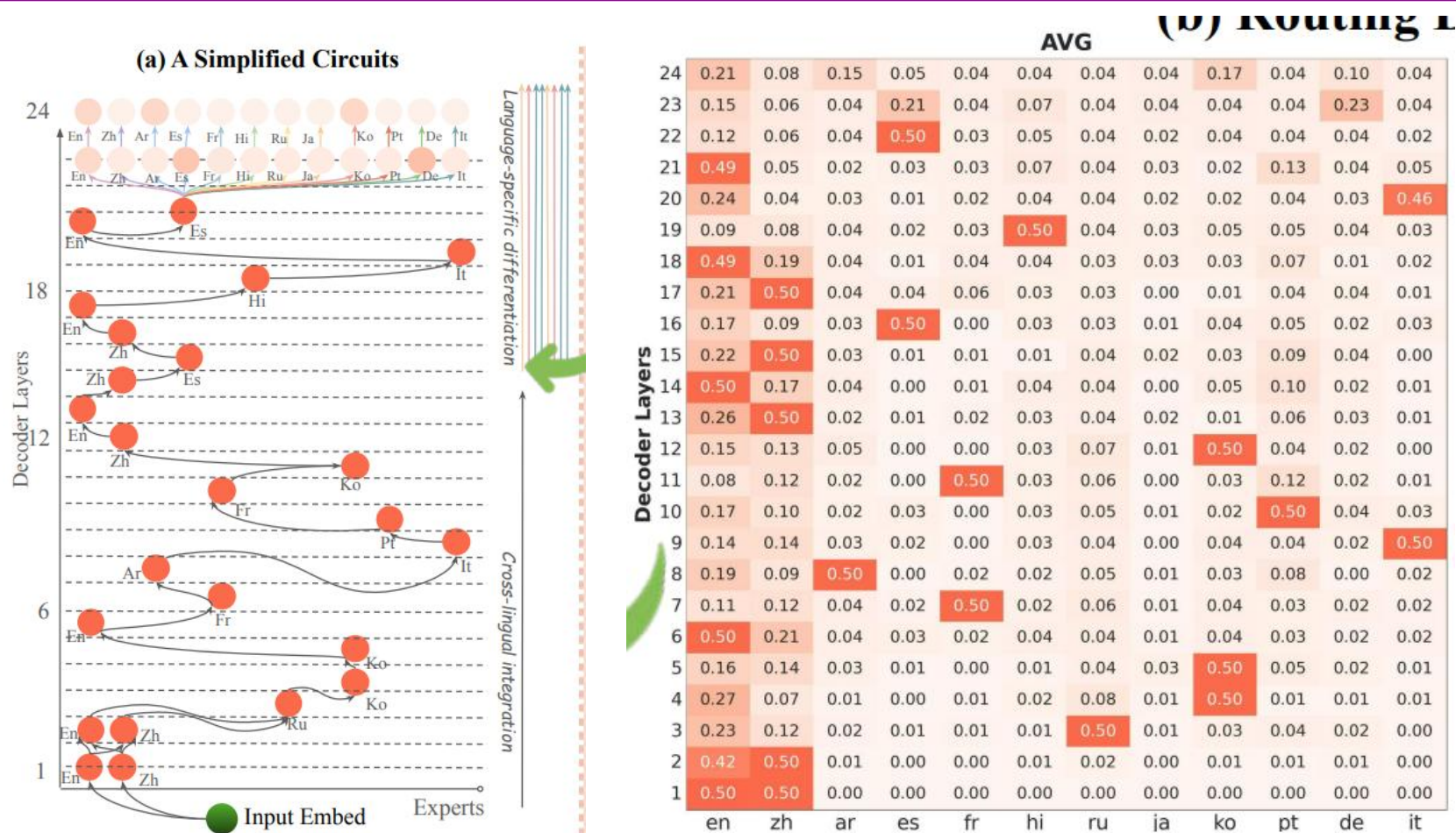
Table 2: Comparison between Dense models and MoE models with various routing strategies.

Method	Param. (B)		Avg. Accuracy		Ar	De	En	Es	Fr	Zh
	Active	Total	Major	Minor						
Dense Models before and after Training with V										
Qwen-0.5B	0.49	0.49	29.7	31.5	27.3	27.4	39.3	32.9	21.3	5.4
<i>after training</i>	0.49	0.49	37.8	24.6	34.8	33.7	46.7	40.5	33.0	1.0
Same Active	0.81	0.81	38.4	26.2	34.4	35.8	46.1	40.1	34.0	1.8
Same Total	3.95	3.95	42.0	30.9	36.4	40.8	47.8	42.1	38.0	4.8
MoE Models Trained with Different Routing										
Lang-Spec.	0.81	3.95	30.9	26.1	28.8	28.6	39.1	33.3	16.5	3.0
Top- <i>k</i>	0.81	3.95	39.7	29.9	34.5	36.9	43.7	38.9	39.9	3.2
Hybrid- <i>k</i>	0.81	3.95	40.0	32.0	35.1	37.2	44.1	40.8	40.7	3.6





Interpretable routing analysis

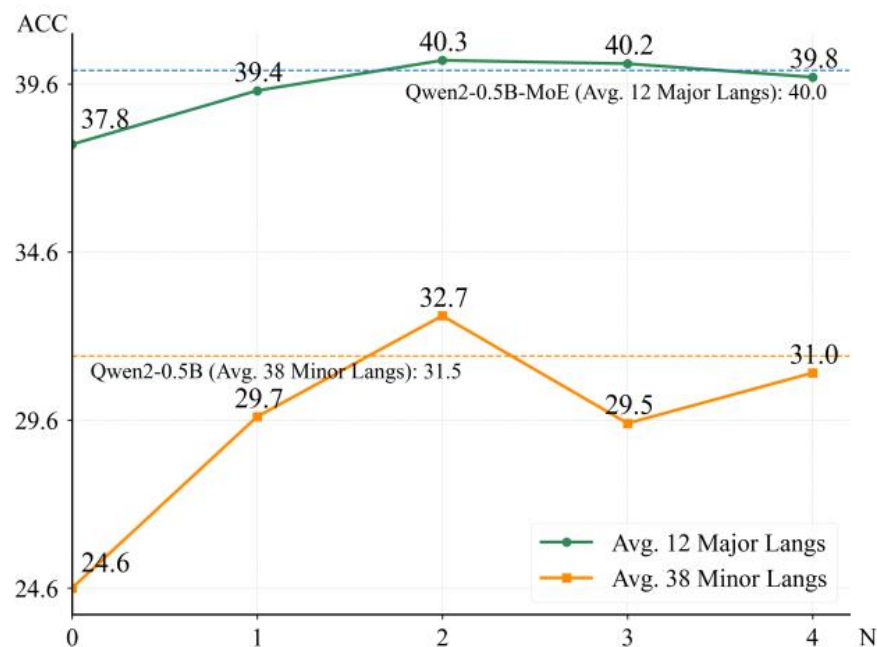


Maybe only the last 2 need to be MOE!





Post-MoE variant inspired by the ‘Spread Out in the End’ Phenomenon



(a) Post-MoE from Qwen2-0.5B in Last N Layers



(b) Post-MoE from Qwen2-1.5B in Last N Layers

Yes! Only the last 2 need to be MOE!

An architecture (Post-MoE) and training method (Hybrid) that is both efficient, generalizable, and interpretable.





Outline

- Introduction (5 min)
- A preliminary scaling to 12 languages (5 min)
- Scaling with MoE and its routing analysis (10 min)
 - A language-specific hybrid routing
 - Interpretable routing analysis: information flow circuit
 - Post-MoE variant inspired by the ‘Spread Out in the End’ Phenomenon
- Further Scaling to 50 Languages (10 min)
 - Mixture of language family experts
 - Results and Ablation Study
- Q&A (5 min)





Mixture of language family experts

An architecture (Post-MoE) and training method (Hybrid) that is both efficient, generalizable, and interpretable.

50 Language !!! Huge Parameter !!! Still not efficient enough !!!

Table 3: Classification of Languages with Color Coding Based on Characteristics.

Language Family	Languages
Sino-Tibetan	Chinese (Zh)
Altaic	Korean (Ko), Japanese (Ja), Mongolian (Ne)
Australasian	Thai (Th), Vietnamese (Vi), Laotian (Lo)
Austronesian	Malagasy (Mg), Cebuano (Ceb), Sundanese (Su), Ilokano (Ilo), Dogue (Doi)
Indo-European	English (En), German (De), Portuguese (Pt), Spanish (Es), French (Fr), Russian (Ru), Italian (It), Croatian (Hr), Galician (Gl), Czech (Cs), Corsican (Co), Latin (La), Ukrainian (Uk), Bosnian (Bs), Bulgarian (Bg), Esperanto (Eo), Maithili (Mai), Albanian (Sq), Danish (Da), Sanskrit (Sa), Norwegian (No), Guarani (Gn), Serbian (Sr), Slovak (Sk), Scottish Gaelic (Gd), Luxembourgish (Lb), Hindi (Hi)
Afro-Asian	Arabic (Ar), Kurdish (Sorani) (Ckb), Maltese (Mt), Hebrew (He)
Kongolese	Lingala (Ln), Bambara (Bm), Swahili (Sw), Sepeti (Nso), Igbo (Ig), Kinyarwanda (Rw), Hausa (Ha)





Results and Ablation Study

Table 4: Main Results on 50 languages comparing to existing LLMs

Model	Active Param.	Average Acc.	
		Major	Minor
Gemini 1.5 Pro	-	74.6	73.6
GPT-4 Turbo	-	79.4	73.3
Claude 3 Opus	-	81.9	76.5
GPT-4o	-	85.7	81.1
GPT-4o mini	-	77.6	69.8
JSL-MedPhi2	2.78 B	29.0	30.7
MMed-Llama-3	8.03 B	40.2	36.5
OpenBioLLM	8.03 B	46.1	34.7
Llama3	8.03 B	49.3	33.3
Apollo-MoE	0.52 B	40.5	34.6
Apollo-MoE	1.63 B	54.8	44.9
Apollo-MoE	8.02 B	69.9	58.3

Languages	Active Param.	Average acc.	
		Major	Minor
Qwen2-0.5B	0.49 B	29.7	31.5
Dense	0.49 B	39.2	33.2
Dense Same Active	0.52 B	39.4	34.0
Top- <i>k</i> routing	0.52 B	39.0	33.4
Hybrid- <i>k</i> routing	0.52 B	40.5	34.6
Qwen2-1.5B	1.54 B	42.9	38.4
Dense	1.54 B	52.2	43.7
Dense Same Active	1.63 B	52.8	44.1
Top- <i>k</i> routing	1.63 B	53.6	42.6
Hybrid- <i>k</i> routing	1.63 B	54.8	44.9
Qwen2-7B	7.62 B	55.2	49.2
Dense	7.62 B	69.0	55.7
Dense Same Active	8.02 B	68.5	56.3
Top- <i>k</i> Routing	8.02 B	68.6	56.7
Hybrid- <i>k</i> Routing	8.02 B	69.9	58.3

Table 5: Ablation Study between Dense and PostMoE Models across 50 Languages at Various Scales.

IT WORK !!! With Scalability !!!





Outline

- Introduction (5 min)
- A preliminary scaling to 12 languages (5 min)
- Scaling with MoE and its routing analysis (10 min)
 - A language-specific hybrid routing
 - Interpretable routing analysis: information flow circuit
 - Post-MoE variant inspired by the ‘Spread Out in the End’ Phenomenon
- Further Scaling to 50 Languages (10 min)
 - Mixture of language family experts
 - Results and Ablation Study
- Q&A (5 min)

