

Multilingualism of Qwen

From Foundation Model to Applications

Baosong Yang, Alibaba Group

① Introduction of Qwen3

② Multilingualism of Qwen

③ Application Techniques



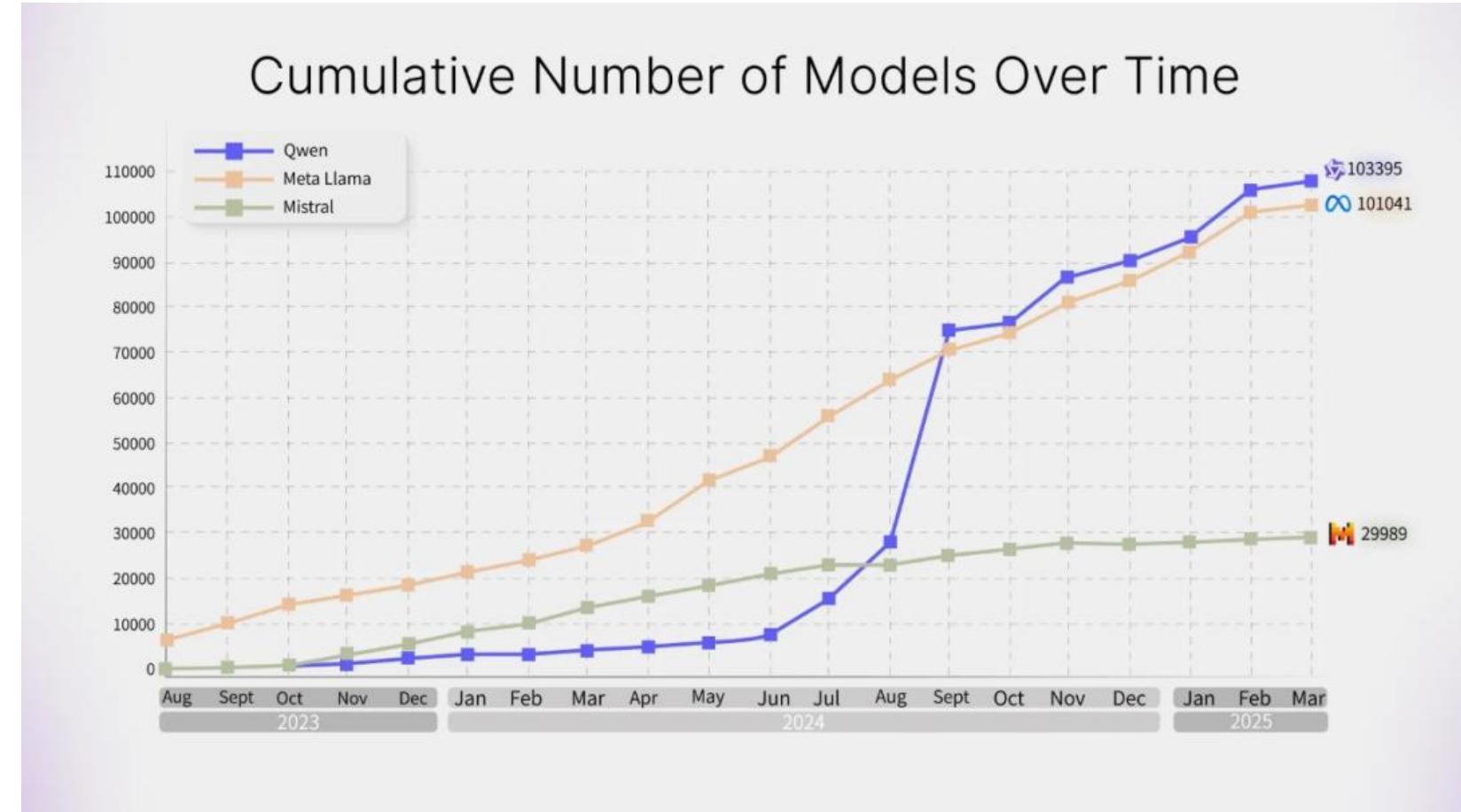
Introduction of Qwen3



Insist on open source

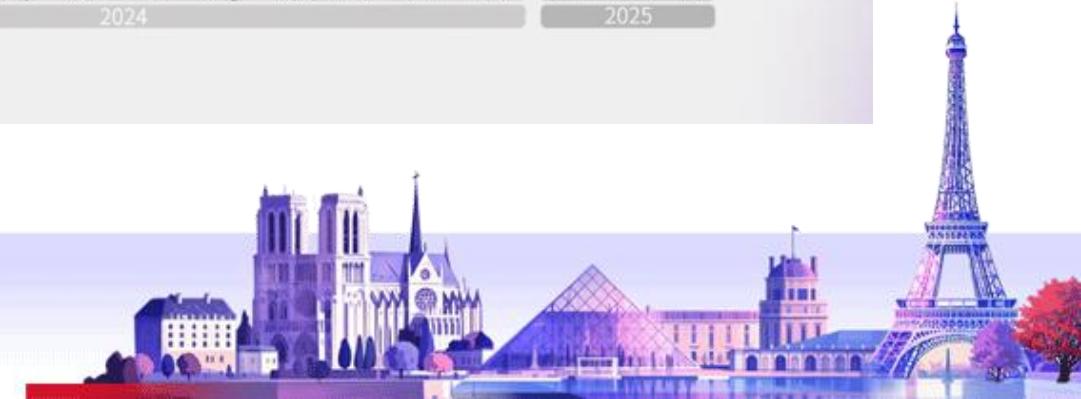
GOSIM

200+ Models
300M+ Downloads
100K+ Derivatives



One of the most popular open-source LLM series.

GOSIM AI Paris 2025



There's always one that's right for you **GOSIM**

- Qwen3-235B-A22B
- Qwen3-32B
- Qwen3-30B-A3B
- Qwen3-14B
- Qwen3-8B
- Qwen3-4B
- Qwen3-1.7B
- Qwen3-0.6B

Models	Layers	Heads (Q / KV)	Tie Embedding	Context Length
Qwen3-0.6B	28	16 / 8	Yes	32K
Qwen3-1.7B	28	16 / 8	Yes	32K
Qwen3-4B	36	32 / 8	Yes	32K
Qwen3-8B	36	32 / 8	No	128K
Qwen3-14B	40	40 / 8	No	128K
Qwen3-32B	64	64 / 8	No	128K

Models	Layers	Heads (Q / KV)	# Experts (Total / Activated)	Context Length
Qwen3-30B-A3B	48	32 / 4	128 / 8	128K
Qwen3-235B-A22B	94	64 / 4	128 / 8	128K

Open-sourcing of Qwen3 will significantly advance the research and development of large foundation models.



Qwen3: Think Deeper, Act Faster

GOSIM

- Qwen3-235B-A22B
- Qwen3-32B
- Qwen3-30B-A3B
- Qwen3-14B
- Qwen3-8B
- Qwen3-4B
- Qwen3-1.7B
- Qwen3-0.6B

	Qwen3-235B-A22B MoE	Qwen3-32B Dense	OpenAI-o1 2024-12-17	Deepseek-R1	Grok 3 Beta Think	Gemini2.5-Pro	OpenAI-o3-mini Medium
ArenaHard	95.6	93.8	92.1	93.2	-	96.4	89.0
AIME'24	85.7	81.4	74.3	79.8	83.9	92.0	79.6
AIME'25	81.5	72.9	79.2	70.0	77.3	86.7	74.8
LiveCodeBench v5, 2024.10-2025.02	70.7	65.7	63.9	64.3	70.6	70.4	66.3
CodeForces Elo Rating	2056	1977	1891	2029	-	2001	2036
Aider Pass@2	61.8	50.2	61.7	56.9	53.3	72.9	53.8
LiveBench 2024-11-25	77.1	74.9	75.7	71.6	-	82.4	70.0
BFCL v3	70.8	70.3	67.8	56.9	-	62.9	64.6
MultilF 8 Languages	71.9	73.0	48.8	67.7	-	77.8	48.4

1. AIME 24/25: We sample 64 times for each query and report the average of the accuracy. AIME'25 consists of Part I and Part II, with a total of 30 questions.

2. Aider: We didn't activate the think mode of Qwen3 to balance efficiency and effectiveness.

3. BFCL: The Qwen3 models are evaluated using the FC format, while the baseline models are assessed using the highest scores obtained from either the FC or prompt formats.

The flagship model, **Qwen3-235B-A22B**, achieves competitive results in benchmark evaluations of coding, math, general capabilities, etc.



Qwen3: Think Deeper, Act Faster

GOSIM

- Qwen3-235B-A22B
- Qwen3-32B
- Qwen3-30B-A3B
- Qwen3-14B
- Qwen3-8B
- Qwen3-4B
- Qwen3-1.7B
- Qwen3-0.6B

		Qwen3-30B-A3B MoE	QwQ-32B	Qwen3-4B Dense	Qwen2.5-72B-Instruct	Gemma3-27B-IT	DeepSeek-V3	GPT-4o 2024-11-20
	ArenaHard	91.0	89.5	76.6	81.2	86.8	85.5	85.3
	AIME'24	80.4	79.5	73.8	18.9	32.6	39.2	11.1
	AIME'25	70.9	69.5	65.6	15.0	24.0	28.8	7.6
	LiveCodeBench v5, 2024.10-2025.02	62.6	62.7	54.2	30.7	26.9	33.1	32.7
	CodeForces Elo Rating	1974	1982	1671	859	1063	1134	864
	GPQA	65.8	65.6	55.9	49.0	42.4	59.1	46.0
	LiveBench 2024-11-25	74.3	72.0	63.6	51.4	49.2	60.5	52.2
	BFCL v3	69.1	66.4	65.9	63.4	59.1	57.6	72.5
	MultilF 8 Languages	72.2	68.3	66.3	65.3	69.8	55.6	65.6

1. AIME 24/25: We sample 64 times for each query and report the average of the accuracy. AIME'25 consists of Part I and Part II, with a total of 30 questions.

2. Aider: We didn't activate the think mode of Qwen3 to balance efficiency and effectiveness.

3. BFCL: The Qwen3 models are evaluated using the FC format, while the baseline models are assessed using the highest scores obtained from either the FC or prompt formats.

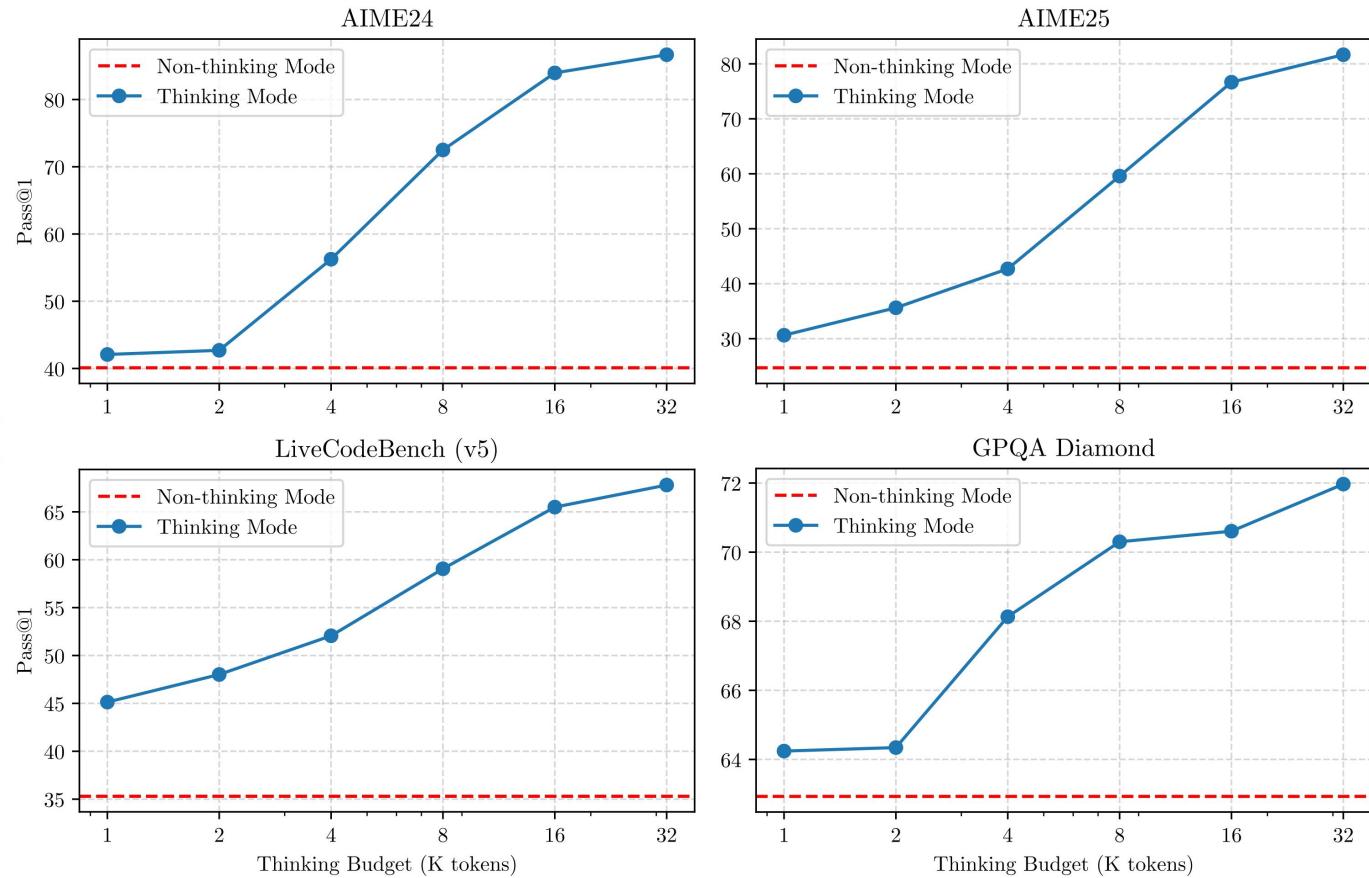
Qwen3-4B yields comparable performance to Qwen2.5-72B.



Key Features

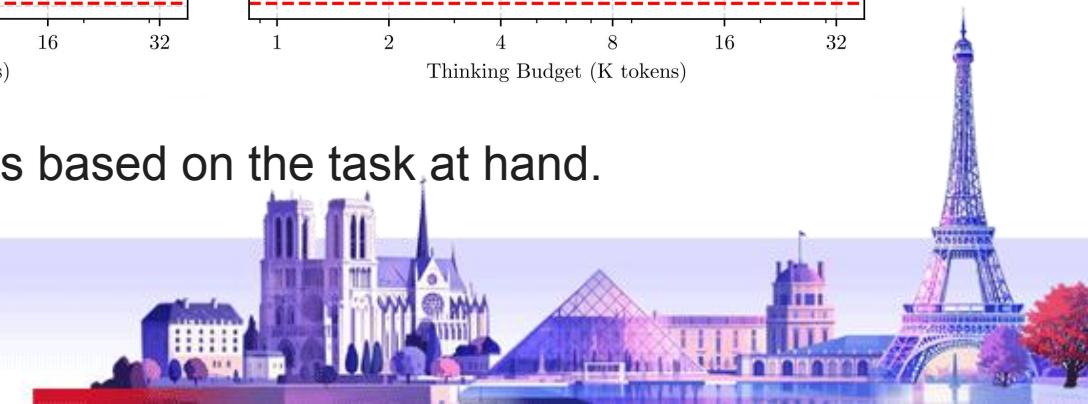
GOSIM

Key Feature 1: Hybrid Thinking Modes



* The Pass@1 scores are averaged over 64 (AIME24 and AIME25) or 16 (LCB and GPQA) samplings.

GOSIM AI Paris 2025



Key Features

GOSIM

Key Feature 1: Hybrid Thinking Modes

The default mode

Reasoning step by step before delivering the final answer. This is ideal for complex problems that require deeper thought.

GOSIM AI Paris 2025

Qwen3-235B-A22B
已完成思考 · 38,912 tokens 预算 ^

Okay, the user is asking again how many 'r's are in the word "strawberries." Let me check my previous answer to make sure I was correct. Last time, I spelled out the word as S-T-R-A-W-B-E-R-R-I-E-S and identified three 'r's. Let me verify that.

First, let's write the word out letter by letter: S-T-R-A-W-B-E-R-R-I-E-S. Breaking it down:

S, T, R (1st R), A, W, B, E, R (2nd R), R (3rd R), I, E, S. Yep, that's three R's.

Wait, maybe I should count again to be sure. Sometimes when words have double letters, it's easy to miscount. Let's go step by step:

1. S - no
2. T - no
3. R - first R
4. A - no
5. W - no
6. B - no
7. E - no
8. R - second R
9. R - third R
10. I - no
11. E - no
12. S - no

Yes, exactly three R's. So my previous answer was correct. The user might be testing me or just wants to confirm. Alternatively, maybe they heard a different answer elsewhere. But according to standard spelling, "strawberries" has two R's in "berries" part. Wait, hold on. Wait, "strawberries" is "straw" + "berries". So "berries" is B-E-R-R-I-E-S. So in "berries", there are two R's. Then in "straw", there's one R. So total of three R's. That's correct. So the answer is three.

Alternatively, could there be a trick here? Like, sometimes people might mispronounce or misspell the word. But the standard spelling is definitely with three R's. So the answer remains three. I should confirm this again to avoid any mistake. Maybe the user made a typo, but the question seems straightforward. Since the user asked the same question again, perhaps they want a concise answer. Let me present the answer clearly, reiterate the steps briefly, and confirm the count. Also, maybe the user is learning English or teaching someone, so reinforcing the correct count is important. Alright, proceed to explain again with clarity.

Let's count the letter 'r' in "strawberries":

1. Strawberries = S-T-R-A-W-B-E-R-R-I-E-S
2. Locate the 'r's :
 - 1st 'r' after 'T' (in straw)
 - 2nd 'r' and 3rd 'r' in berries

✓ Total: 3 r's .
The word "strawberries" has three 'r's in standard spelling.



Key Features

GOSIM

Key Feature 1: Hybrid Thinking Modes

Just type: /no_think

how many r's in blueberries? /no_think

 Qwen3-235B-A22B 20:51

Let's count the number of 'r's in "blueberries" step by step:

Step 1: Spell out the word

Blueberries = B-L-U-E-B-E-R-R-I-E-S

Step 2: Identify the 'r's

- After "bluebe", we get rri → there are two 'r's next to each other.

Step 3: Count them

- First 'r'
- Second 'r'

 **Final Answer:**

There are 2 r's in "blueberries".

Providing quick, near-instant responses, suitable for simpler questions where speed is more important than depth.



Key Features

GOSIM

Key Feature 2: Agentic Capabilities

The screenshot shows a GOSIM interface window. At the top right is the GOSIM logo. Below it is a sidebar with a purple hexagonal icon labeled "Qwen3-32B" and the text "I am Qwen3-32B!". To the right of the sidebar is a list of checkboxes under the heading "插件": "time-get_current_time" (checked), "time-convert_time" (checked), "fetch-fetch" (unchecked), and "code_interpreter" (checked). The main area of the window features large text "Qwen3+MCP" and below it, the instruction "Fetch GitHub stars and plot a bar chart". At the bottom, there is a text input field containing the URL "https://github.com/orgs/QwenLM/repositories Extract markdown content of this page, then draw a bar chart to display the number of stars.", a "点击上传" (Click to upload) button, and a "提交" (Submit) button. A small note at the bottom right says "通过 API 使用" followed by three icons.

Optimized coding and agentic capabilities, strengthened the support of MCP

Key Features

GOSIM

Language Family	Languages & Dialects
Indo-European	English, French, Portuguese, German, Romanian, Swedish, Danish, Bulgarian, Russian, Czech, Greek, Ukrainian, Spanish, Dutch, Slovak, Croatian, Polish, Lithuanian, Norwegian Bokmål, Norwegian Nynorsk, Persian, Slovenian, Gujarati, Latvian, Italian, Occitan, Nepali, Marathi, Belarusian, Serbian, Luxembourgish, Venetian, Assamese, Welsh, Silesian, Asturian, Chhattisgarhi, Awadhi, Maithili, Bhojpuri, Sindhi, Irish, Faroese, Hindi, Punjabi, Bengali, Oriya, Tajik, Eastern Yiddish, Lombard, Ligurian, Sicilian, Friulian, Sardinian, Galician, Catalan, Icelandic, Tosk Albanian, Limburgish, Dari, Afrikaans, Macedonian, Sinhala, Urdu, Magahi, Bosnian, Armenian
Sino-Tibetan	Chinese (Simplified Chinese, Traditional Chinese, Cantonese), Burmese
Afro-Asiatic	Arabic (Standard, Najdi, Levantine, Egyptian, Moroccan, Mesopotamian, Ta'izzi-Adeni, Tunisian), Hebrew, Maltese
Austronesian	Indonesian, Malay, Tagalog, Cebuano, Javanese, Sundanese, Minangkabau, Balinese, Banjar, Pangasinan, Iloko, Waray (Philippines)
Dravidian	Tamil, Telugu, Kannada, Malayalam
Turkic	Turkish, North Azerbaijani, Northern Uzbek, Kazakh, Bashkir, Tatar
Tai-Kadai	Thai, Lao
Uralic	Finnish, Estonian, Hungarian
Austroasiatic	Vietnamese, Khmer
Other	Japanese, Korean, Georgian, Basque, Haitian, Papiamento, Kabuverdianu, Tok Pisin, Swahili

Key Feature 3: Multilingual Support

Qwen3 models are supporting **119 languages and dialects**.

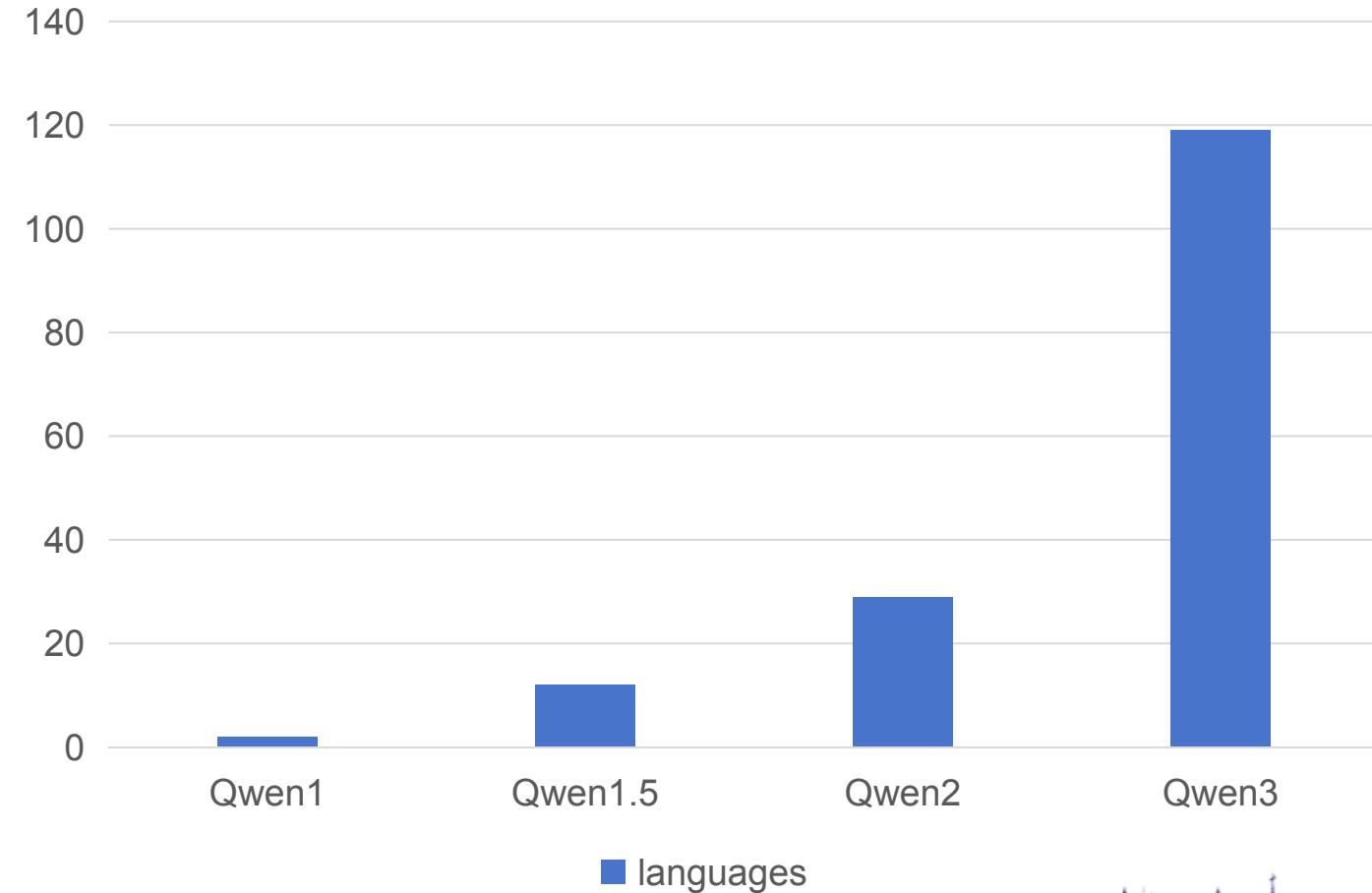
GOSIM AI Paris 2025



Multilingualism of Qwen

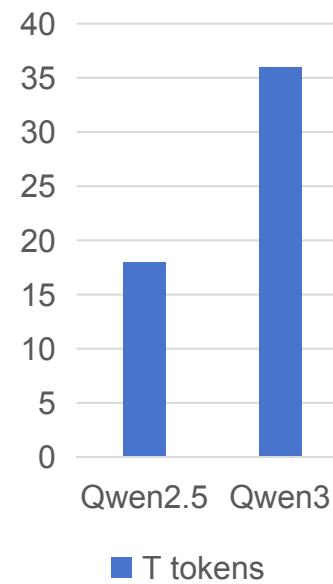
GOSIM

The development of Qwen's multilingualism



Qwen3: Pretraining

GOSIM

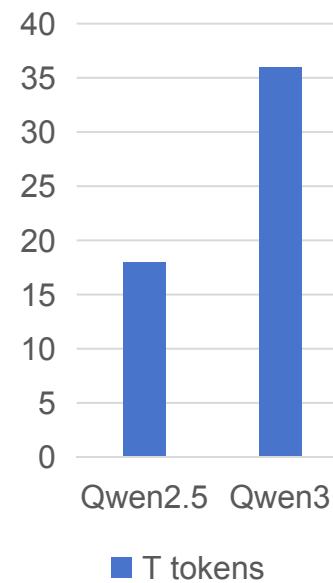


	Qwen2.5-72B Base	Qwen2.5-Plus Base	LLaMA-4-Maverick Base	DeepSeek-V3 Base	Qwen3-235B-A22B Base
# Architecture	Dense	MoE	MoE	MoE	MoE
# Total Params	72B	271B	402B	671B	235B
# Activated Params	72B	37B	17B	37B	22B
<i>General Tasks</i>					
MMLU	86.06	85.02	85.16	<u>87.19</u>	87.81
MMLU-Redux	83.91	82.69	84.05	<u>86.14</u>	87.40
MMLU-Pro	58.07	63.52	<u>63.91</u>	<u>59.84</u>	68.18
SuperGPQA	36.20	37.18	40.85	<u>41.53</u>	44.06
BBH	<u>86.30</u>	85.60	83.62	86.22	88.87
<i>Mathematics & Science Tasks</i>					
GPQA	<u>45.88</u>	41.92	43.94	41.92	47.47
GSM8K	91.50	<u>91.89</u>	87.72	87.57	94.39
MATH	62.12	62.78	<u>63.32</u>	62.62	71.84
<i>Multilingual tasks</i>					
MGSM	82.40	82.21	79.69	<u>82.68</u>	83.53
MMMLU	84.40	83.49	83.09	<u>85.88</u>	86.70
INCLUDE	69.05	66.97	<u>73.47</u>	<u>75.17</u>	73.46
<i>Code tasks</i>					
EvalPlus	65.93	61.43	<u>68.38</u>	63.75	77.60
MultiPL-E	58.70	62.16	<u>57.28</u>	<u>62.26</u>	65.94
MBPP	<u>76.00</u>	74.60	75.40	74.20	81.40
CRUX-O	66.20	68.50	<u>77.00</u>	76.60	79.00

The overall performance of Qwen3 base models matches that of Qwen2.5 base models with more parameters.

Qwen3: Pretraining

GOSIM



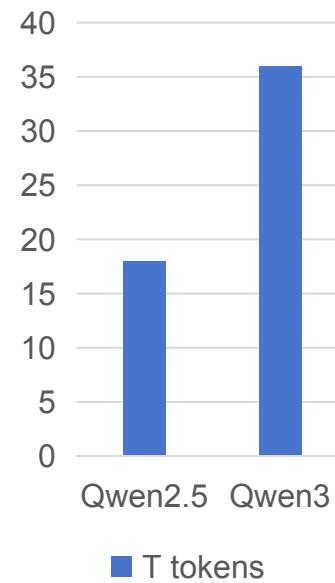
	Qwen2.5-72B Base	Qwen2.5-Plus Base	LLaMA-4-Maverick Base	DeepSeek-V3 Base	Qwen3-235B-A22B Base
# Architecture	Dense	MoE	MoE	MoE	MoE
# Total Params	72B	271B	402B	671B	235B
# Activated Params	72B	37B	17B	37B	22B
<i>General Tasks</i>					
MMLU	86.06	85.02	85.16	87.19	87.81
MMLU-Redux	83.91	82.69	84.05	86.14	87.40
MMLU-Pro	58.07	63.52	63.91	59.84	68.18
SuperGPQA	36.20	37.18	40.85	41.53	44.06
BBH	86.30	85.60	83.62	86.22	88.87
<i>Mathematics & Science Tasks</i>					
GPQA	45.88	41.92	43.94	41.92	47.47
GSM8K	91.50	91.89	87.72	87.57	94.39
MATH	62.12	62.78	63.32	62.62	71.84
<i>Multilingual tasks</i>					
MGSM	82.40	82.21	79.69	82.68	83.53
MMMLU	84.40	83.49	83.09	85.88	86.70
INCLUDE	69.05	66.97	73.47	75.17	73.46
<i>Code tasks</i>					
EvalPlus	65.93	61.43	68.38	63.75	77.60
MultiPL-E	58.70	62.16	57.28	62.26	65.94
MBPP	76.00	74.60	75.40	74.20	81.40
CRUX-O	66.20	68.50	77.00	76.60	79.00

Stage1: the model was pretrained on over 30 trillion tokens with a context length of 4K tokens.



Qwen3: Pretraining

GOSIM

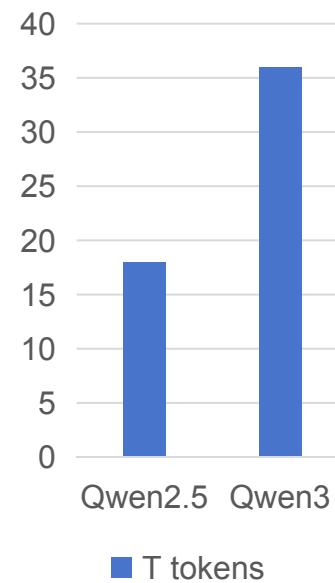


	Qwen2.5-72B Base	Qwen2.5-Plus Base	LLaMA-4-Maverick Base	DeepSeek-V3 Base	Qwen3-235B-A22B Base
# Architecture	Dense	MoE	MoE	MoE	MoE
# Total Params	72B	271B	402B	671B	235B
# Activated Params	72B	37B	17B	37B	22B
<i>General Tasks</i>					
MMLU	86.06	85.02	85.16	<u>87.19</u>	87.81
MMLU-Redux	83.91	82.69	84.05	<u>86.14</u>	87.40
MMLU-Pro	58.07	63.52	<u>63.91</u>	<u>59.84</u>	68.18
SuperGPQA	36.20	37.18	40.85	<u>41.53</u>	44.06
BBH	<u>86.30</u>	85.60	83.62	86.22	88.87
<i>Mathematics & Science Tasks</i>					
GPQA	<u>45.88</u>	41.92	43.94	41.92	47.47
GSM8K	91.50	<u>91.89</u>	87.72	87.57	94.39
MATH	62.12	62.78	<u>63.32</u>	62.62	71.84
<i>Multilingual tasks</i>					
MGSM	82.40	82.21	79.69	<u>82.68</u>	83.53
MMMLU	84.40	83.49	83.09	<u>85.88</u>	86.70
INCLUDE	69.05	66.97	<u>73.47</u>	<u>75.17</u>	73.46
<i>Code tasks</i>					
EvalPlus	65.93	61.43	<u>68.38</u>	63.75	77.60
MultiPL-E	58.70	62.16	<u>57.28</u>	<u>62.26</u>	65.94
MBPP	<u>76.00</u>	74.60	75.40	74.20	81.40
CRUX-O	66.20	68.50	<u>77.00</u>	76.60	79.00

Stage2: improves the dataset by increasing the proportion of knowledge-intensive data, such as STEM, coding, and reasoning tasks.

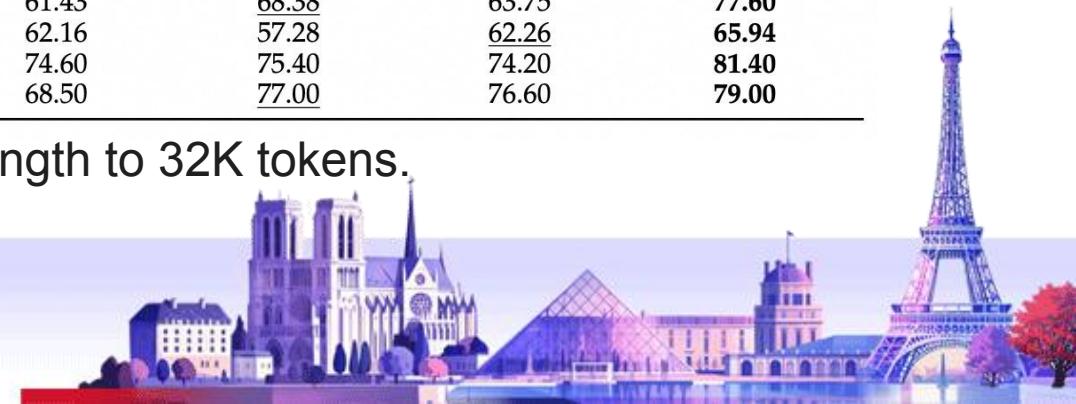
Qwen3: Pretraining

GOSIM



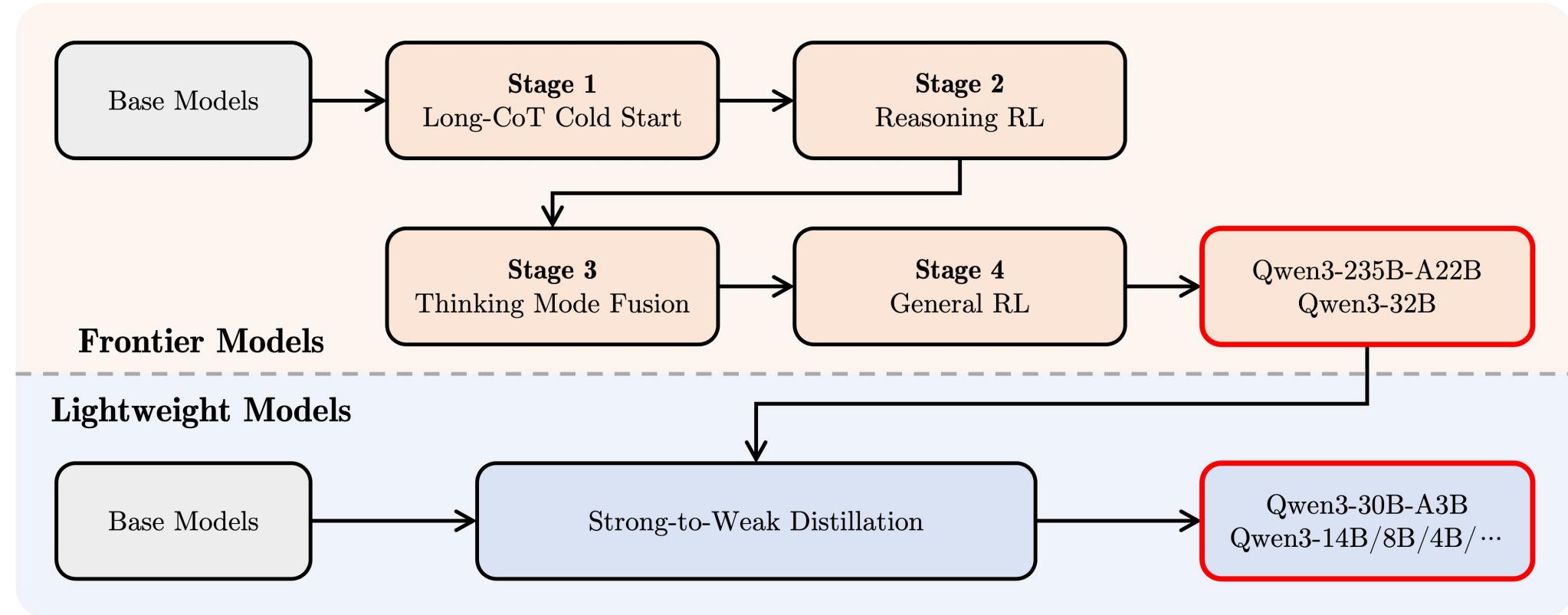
	Qwen2.5-72B Base	Qwen2.5-Plus Base	LLaMA-4-Maverick Base	DeepSeek-V3 Base	Qwen3-235B-A22B Base
# Architecture	Dense	MoE	MoE	MoE	MoE
# Total Params	72B	271B	402B	671B	235B
# Activated Params	72B	37B	17B	37B	22B
<i>General Tasks</i>					
MMLU	86.06	85.02	85.16	<u>87.19</u>	87.81
MMLU-Redux	83.91	82.69	84.05	<u>86.14</u>	87.40
MMLU-Pro	58.07	63.52	<u>63.91</u>	<u>59.84</u>	68.18
SuperGPQA	36.20	37.18	40.85	<u>41.53</u>	44.06
BBH	<u>86.30</u>	85.60	83.62	86.22	88.87
<i>Mathematics & Science Tasks</i>					
GPQA	<u>45.88</u>	41.92	43.94	41.92	47.47
GSM8K	91.50	<u>91.89</u>	87.72	87.57	94.39
MATH	62.12	62.78	<u>63.32</u>	62.62	71.84
<i>Multilingual tasks</i>					
MGSM	82.40	82.21	79.69	<u>82.68</u>	83.53
MMMLU	84.40	83.49	83.09	<u>85.88</u>	86.70
INCLUDE	69.05	66.97	<u>73.47</u>	<u>75.17</u>	73.46
<i>Code tasks</i>					
EvalPlus	65.93	61.43	<u>68.38</u>	63.75	77.60
MultiPL-E	58.70	62.16	<u>57.28</u>	<u>62.26</u>	65.94
MBPP	<u>76.00</u>	74.60	75.40	74.20	81.40
CRUX-O	66.20	68.50	<u>77.00</u>	76.60	79.00

Stage3: uses high-quality long-context data to extend the context length to 32K tokens.



Qwen3: Multi-Stage Post-Training

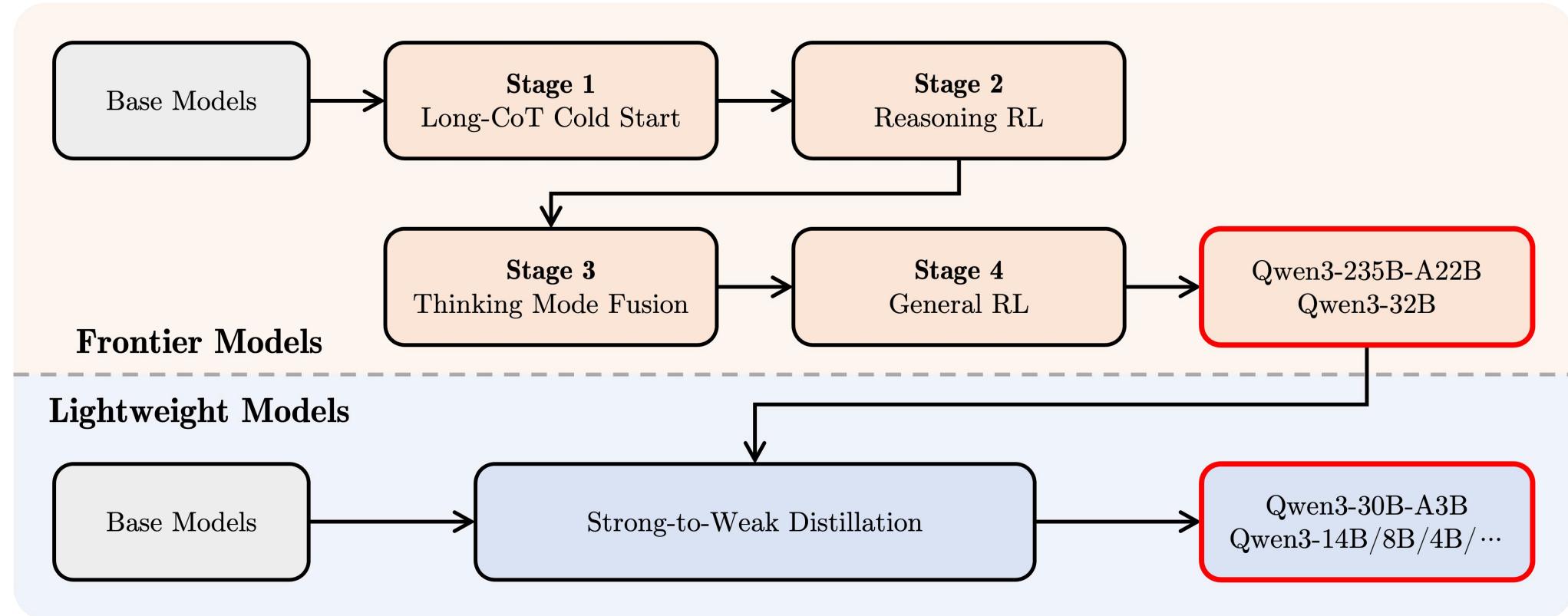
GOSIM



A multiple stage post-training achieves the hybrid model capable of reasoning and rapid responses

Qwen3: Multi-Stage Post-Training

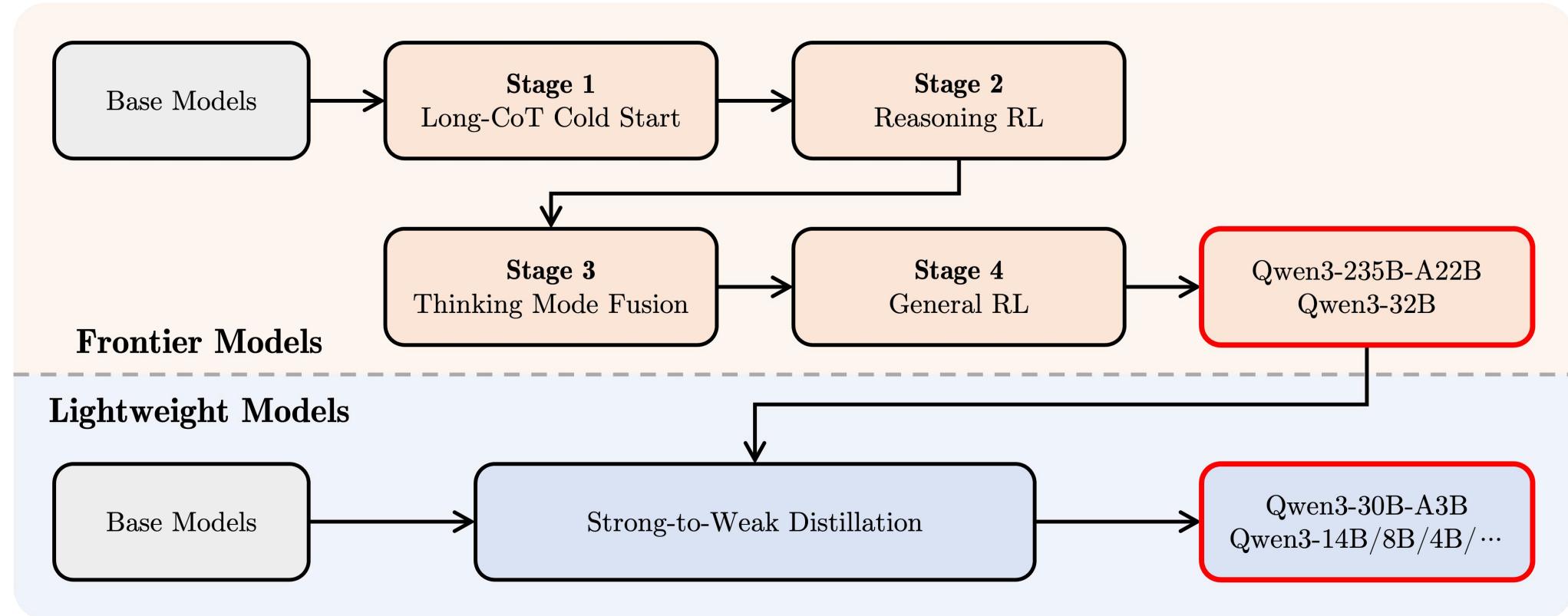
GOSIM



Stage1: diverse long CoT data, covering various tasks and domains such as mathematics, coding, logical reasoning, and STEM problems

Qwen3: Multi-Stage Post-Training

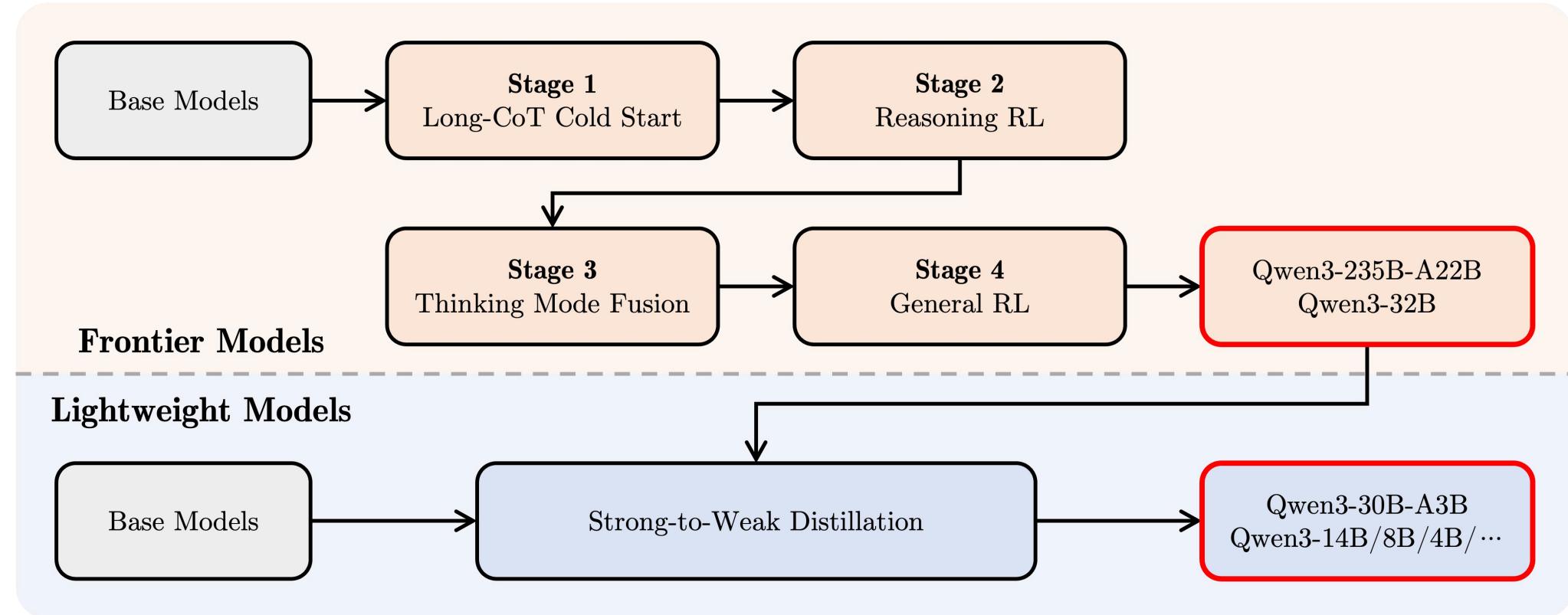
GOSIM



Stage2: focus on scaling up computational resources for RL, utilizing rule-based rewards to enhance the model's exploration and exploitation capabilities.

Qwen3: Multi-Stage Post-Training

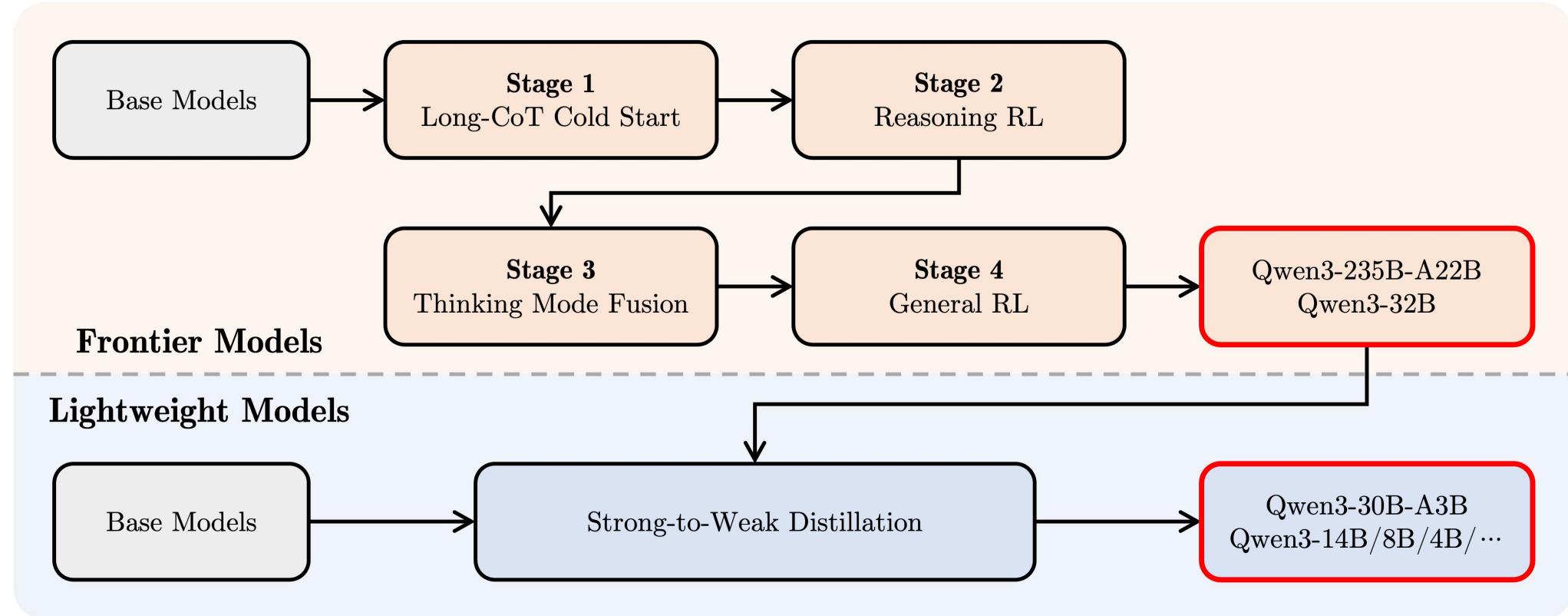
GOSIM



Stage3: integrates non-thinking capabilities into the thinking model by fine-tuning it on a combination of long CoT data and commonly used instruction-tuning data.

Qwen3: Multi-Stage Post-Training

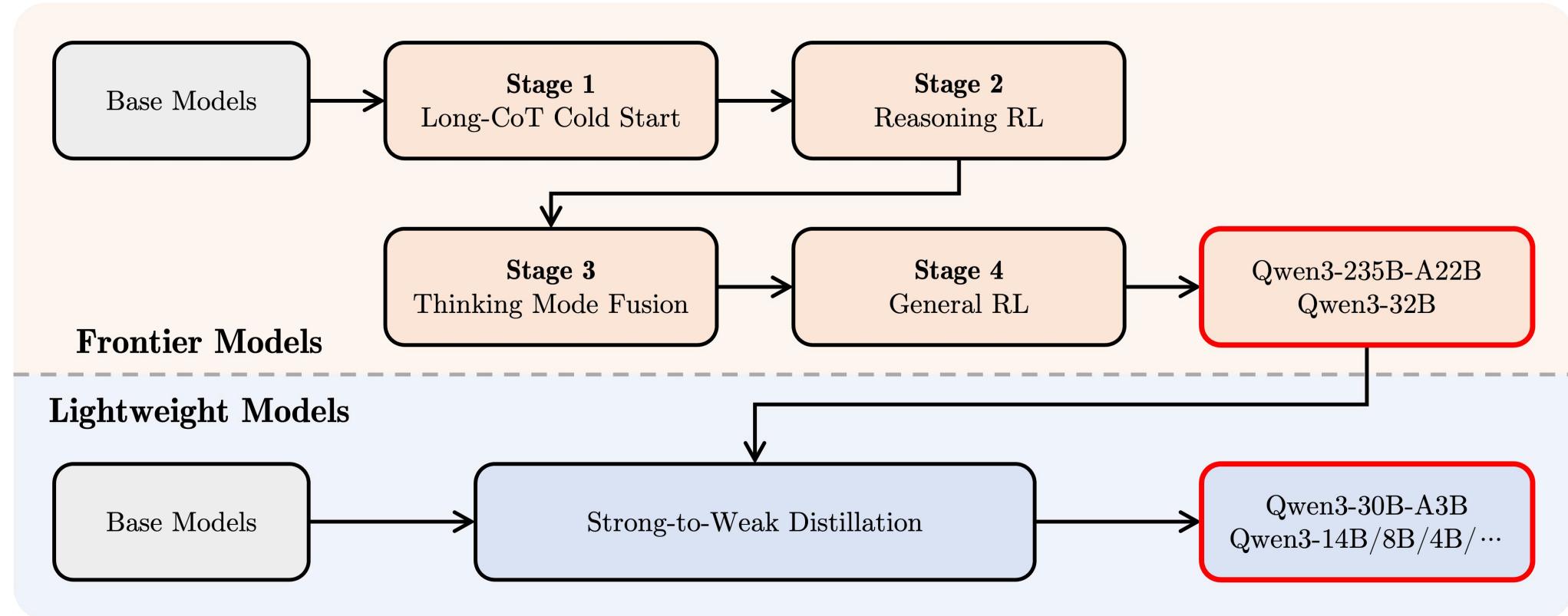
GOSIM



Stage4: applies RL across more than 20 general-domain tasks to further strengthen the model's general capabilities and correct undesired behaviors.

Qwen3: Multi-Stage Post-Training

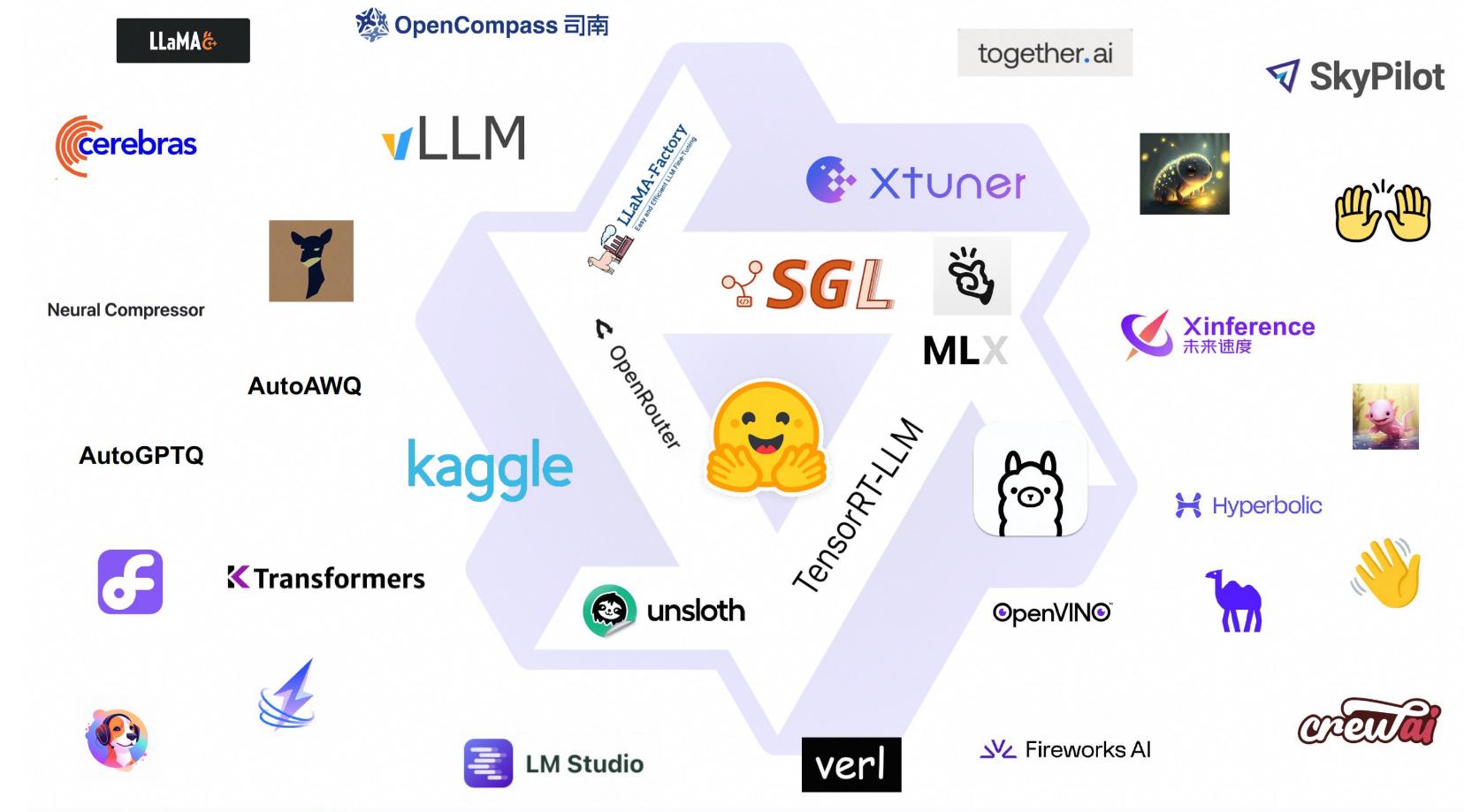
GOSIM



Small-sized models are distilled from large ones.

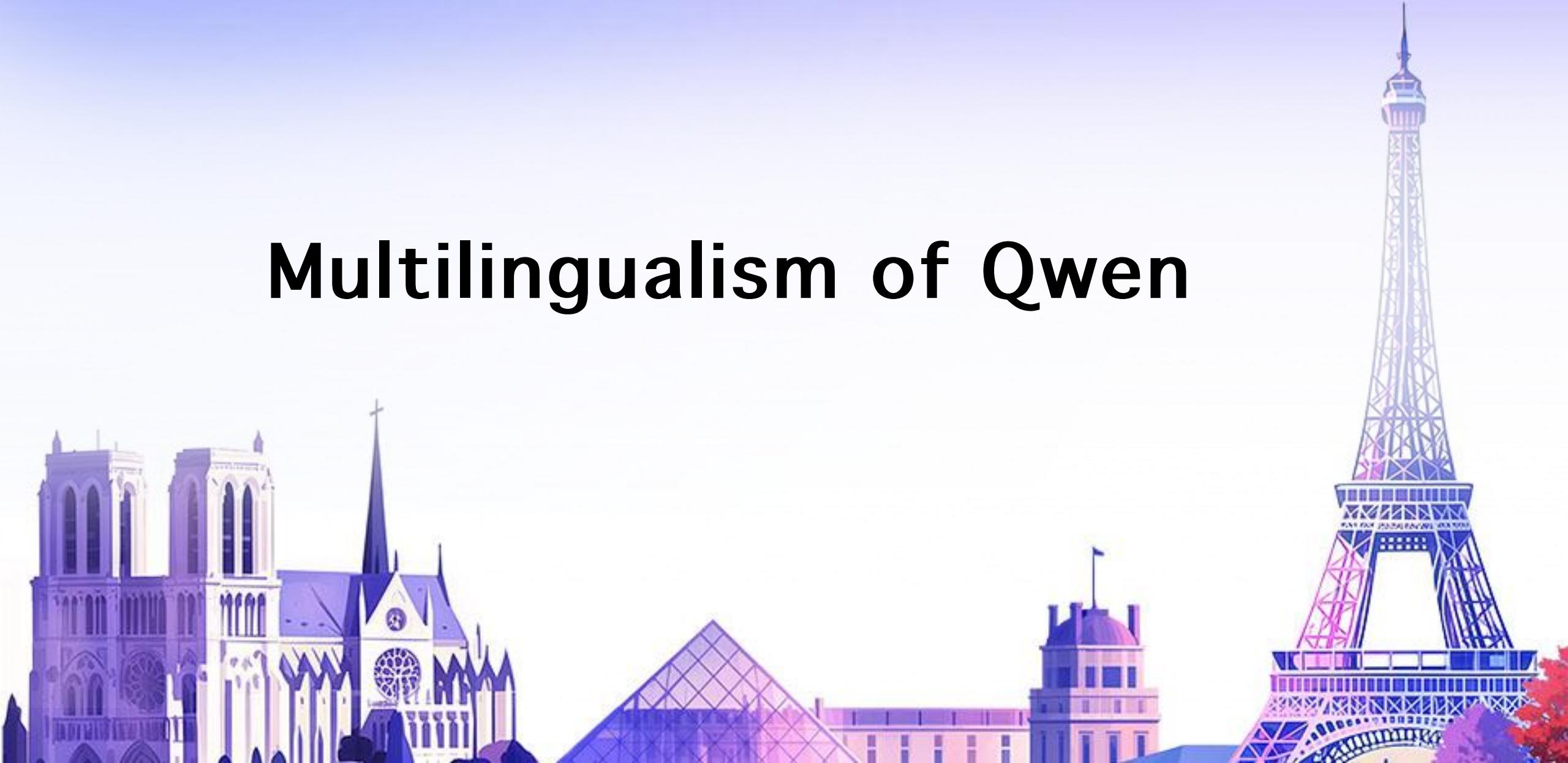
Qwen is nothing without its friends!

GOSIM



Thanks to the support of so many friends!

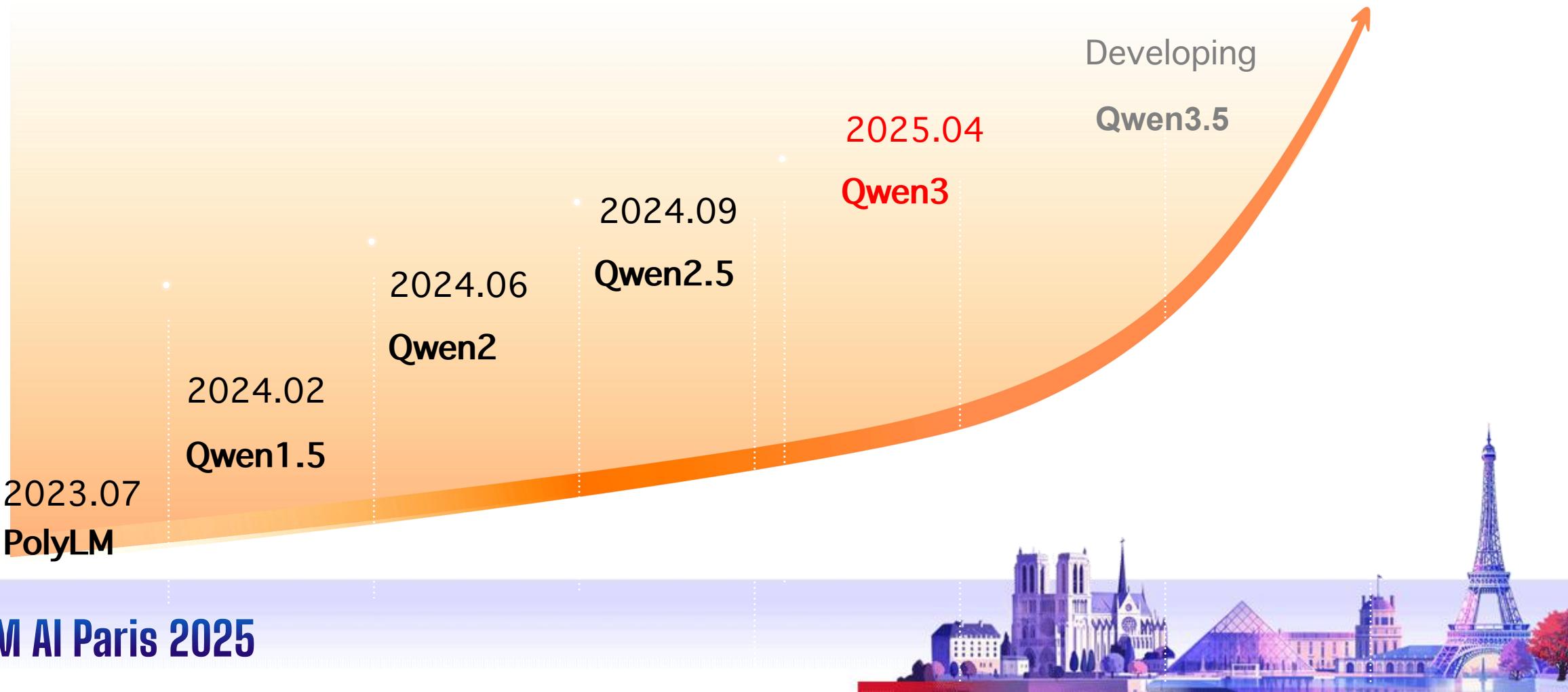
Multilingualism of Qwen



Multilingualism of Qwen

GOSIM

The development of Qwen's multilingualism



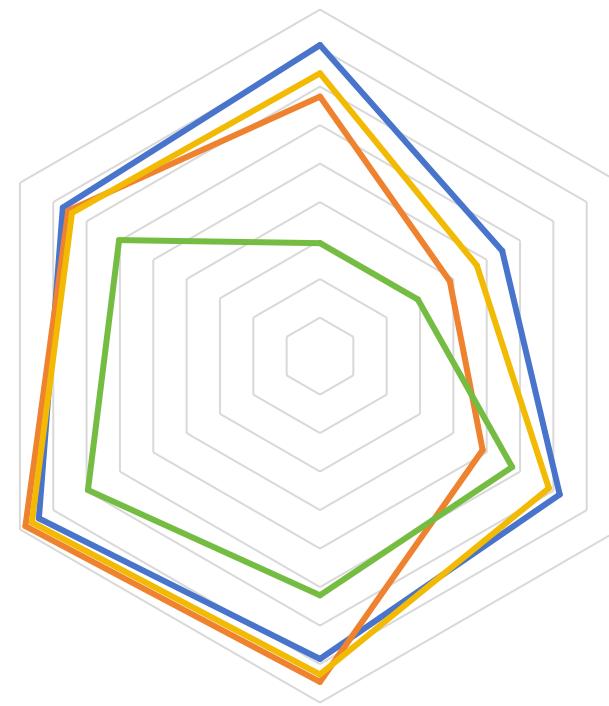
Multilingualism of Qwen

GOSIM

Thinking Mode



Math-AIME24 (55 langs)

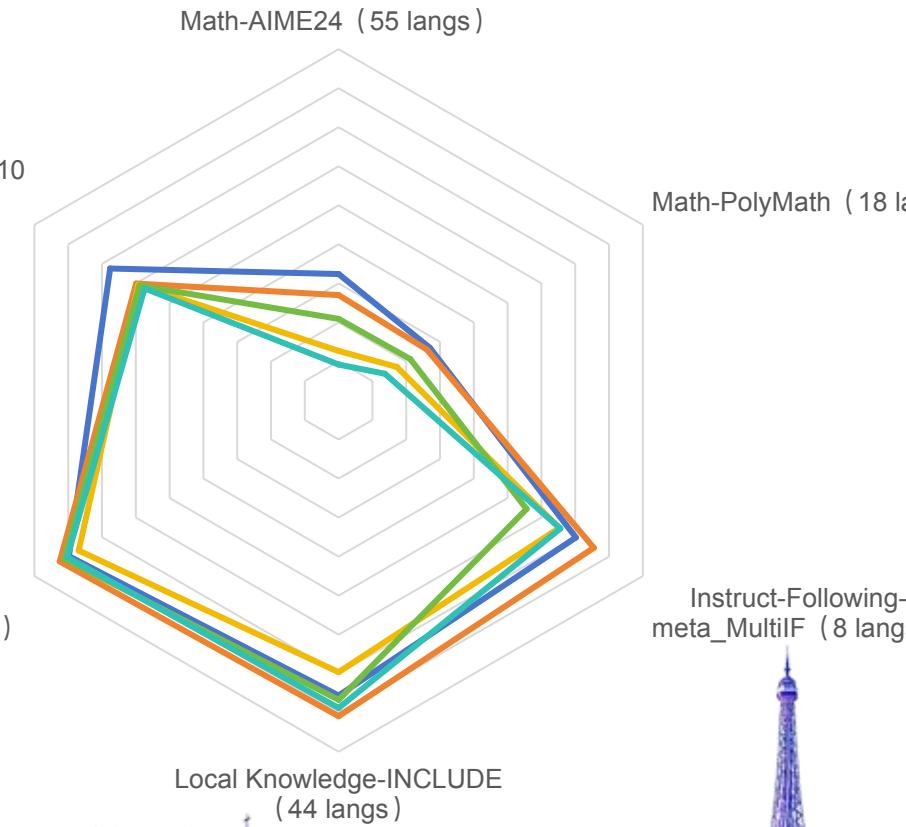


No-thinking Mode



Math-AIME24 (55 langs)

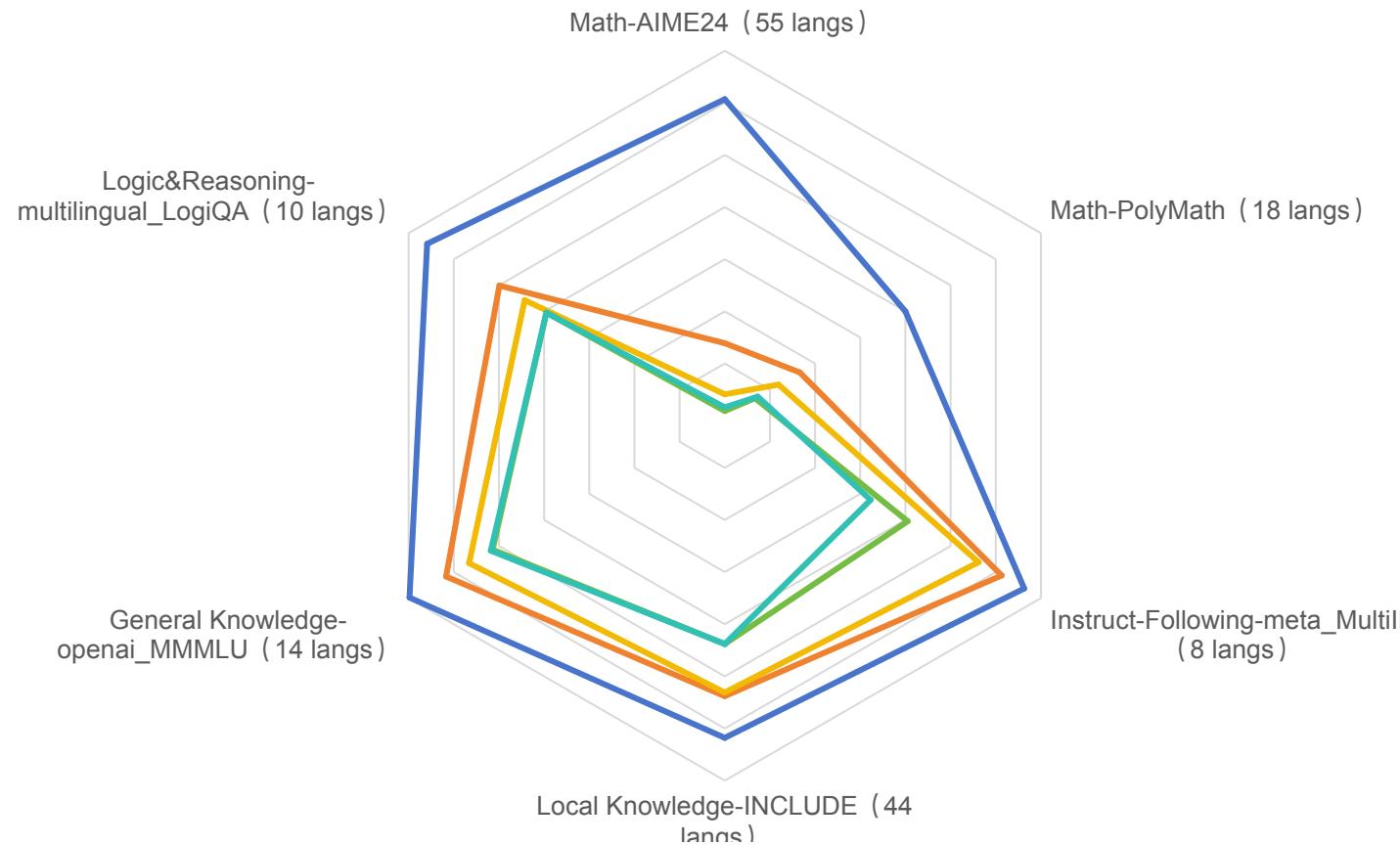
Logic&Reasoning-multilingual_LogiQA (10 langs)



Multilingualism of Qwen

GOSIM

— Qwen3-4B (Thinking Mode) — Qwen3-4B (Nothinking Mode) — gemma3-4B
— phi4-mini — qwen2.5-3B



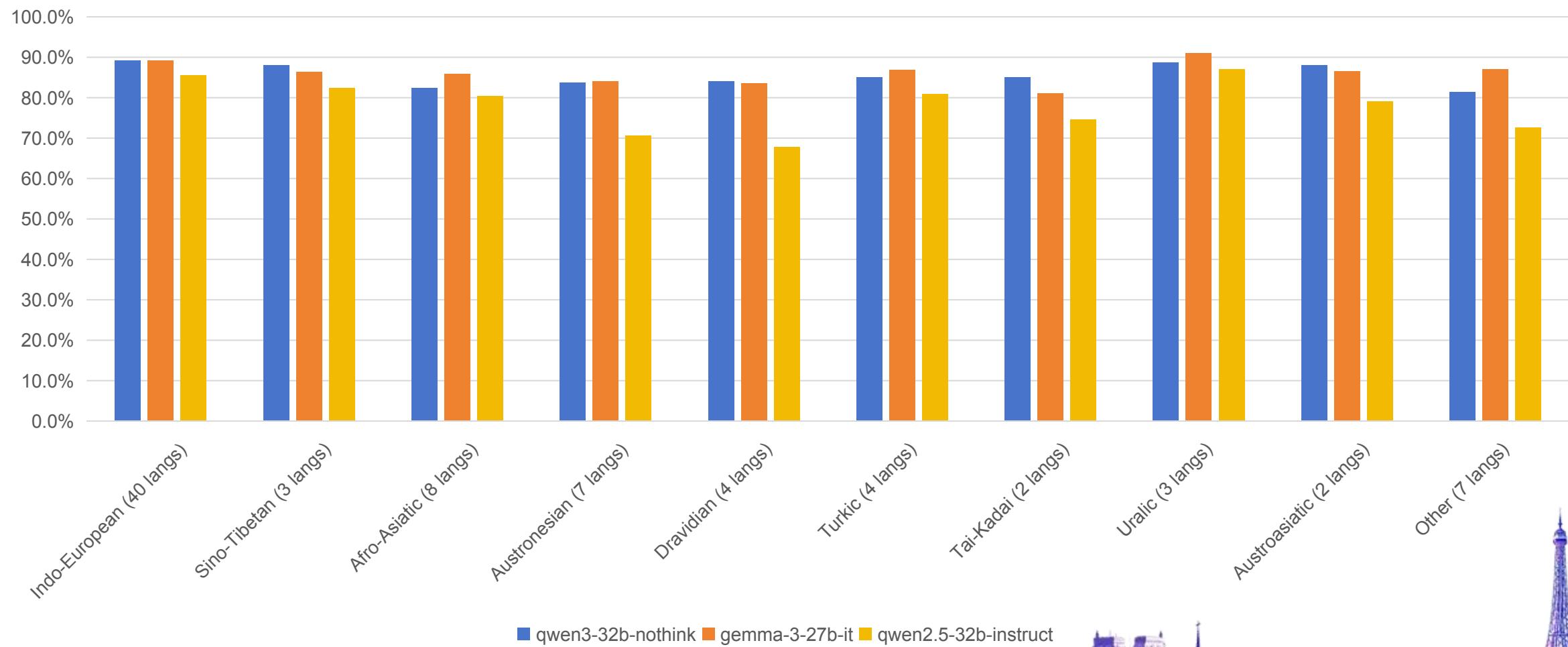
1. Small-sized models can still achieve strong reasoning capabilities.
2. Easier to be deployed on edge devices.



Multilingualism of Qwen

GOSIM

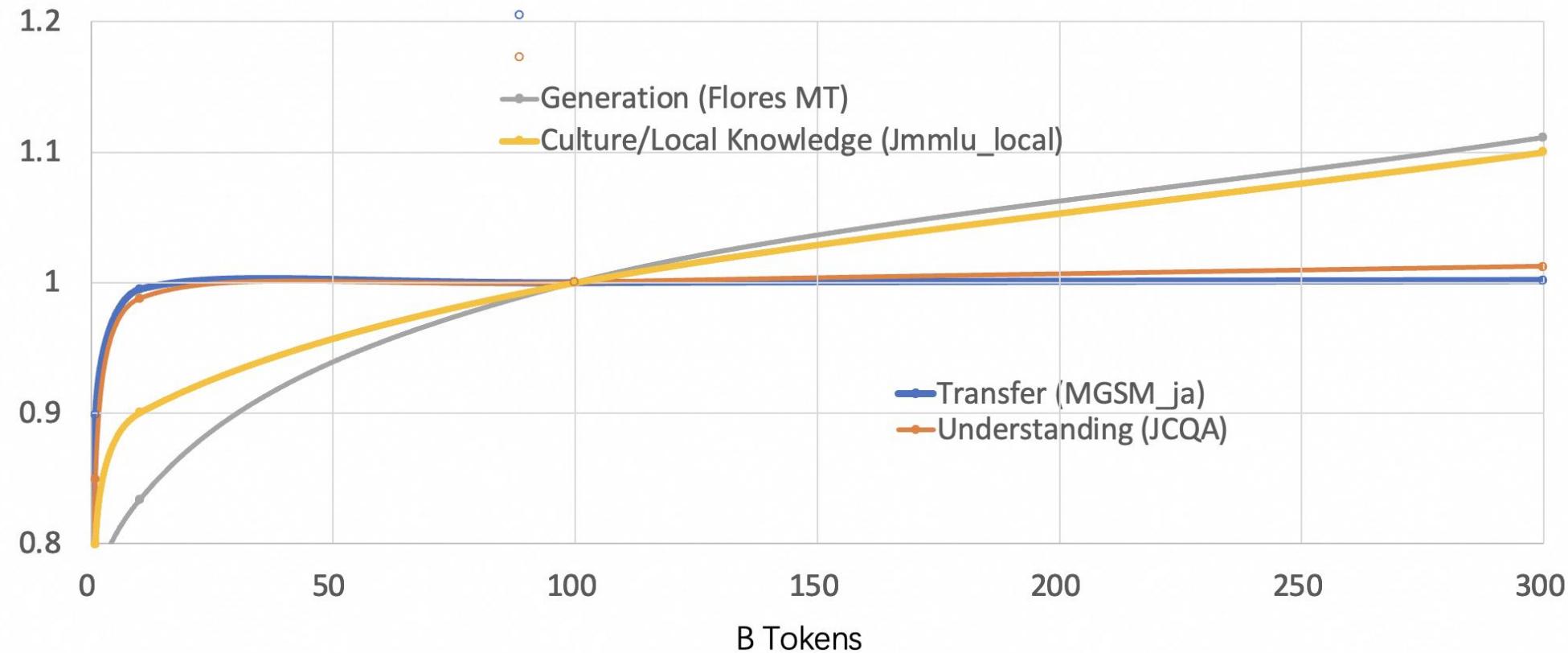
BELEBELE 80 langs.



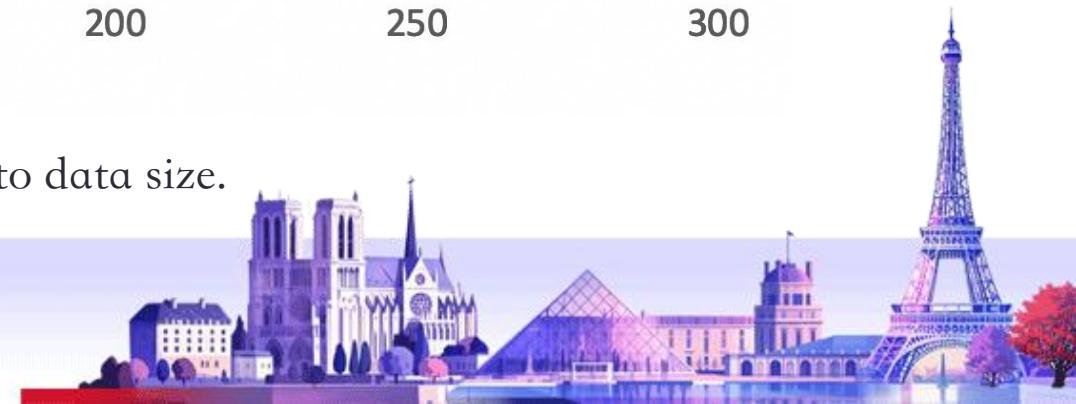
Data Scaling

GOSIM

Normalized performance on different tasks by scaling the Japanese data



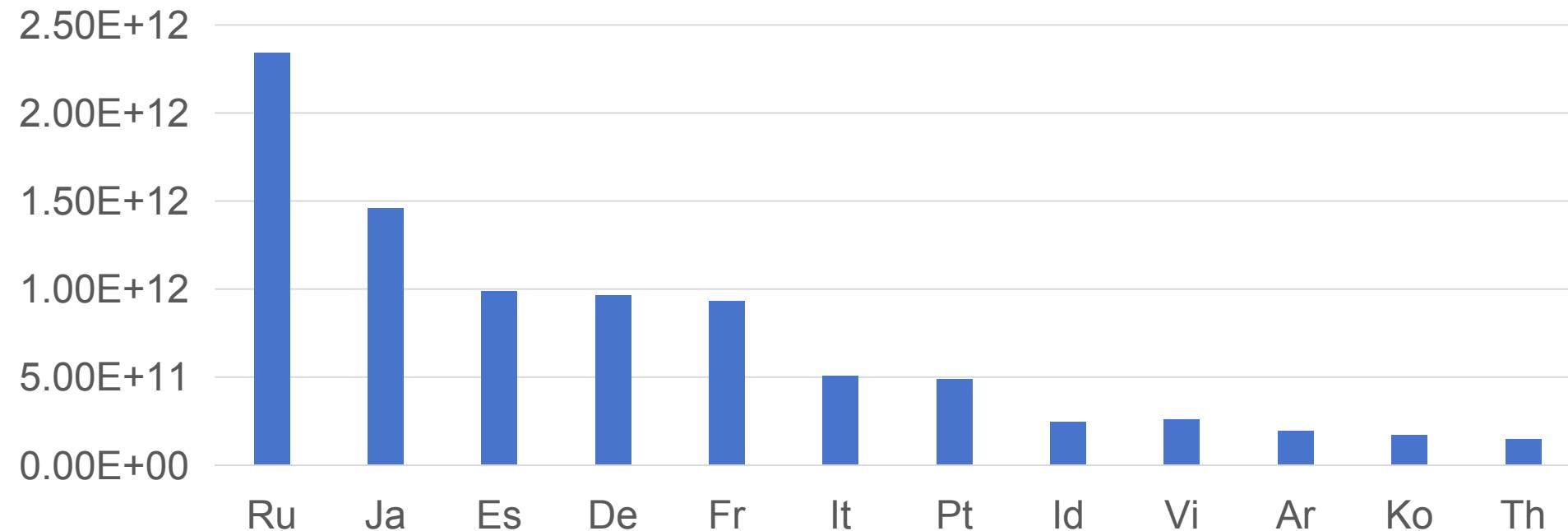
Many capabilities (generation, local knowledge, e.g.) are related to data size.



Data Unbalance

GOSIM

Data Scale on CC

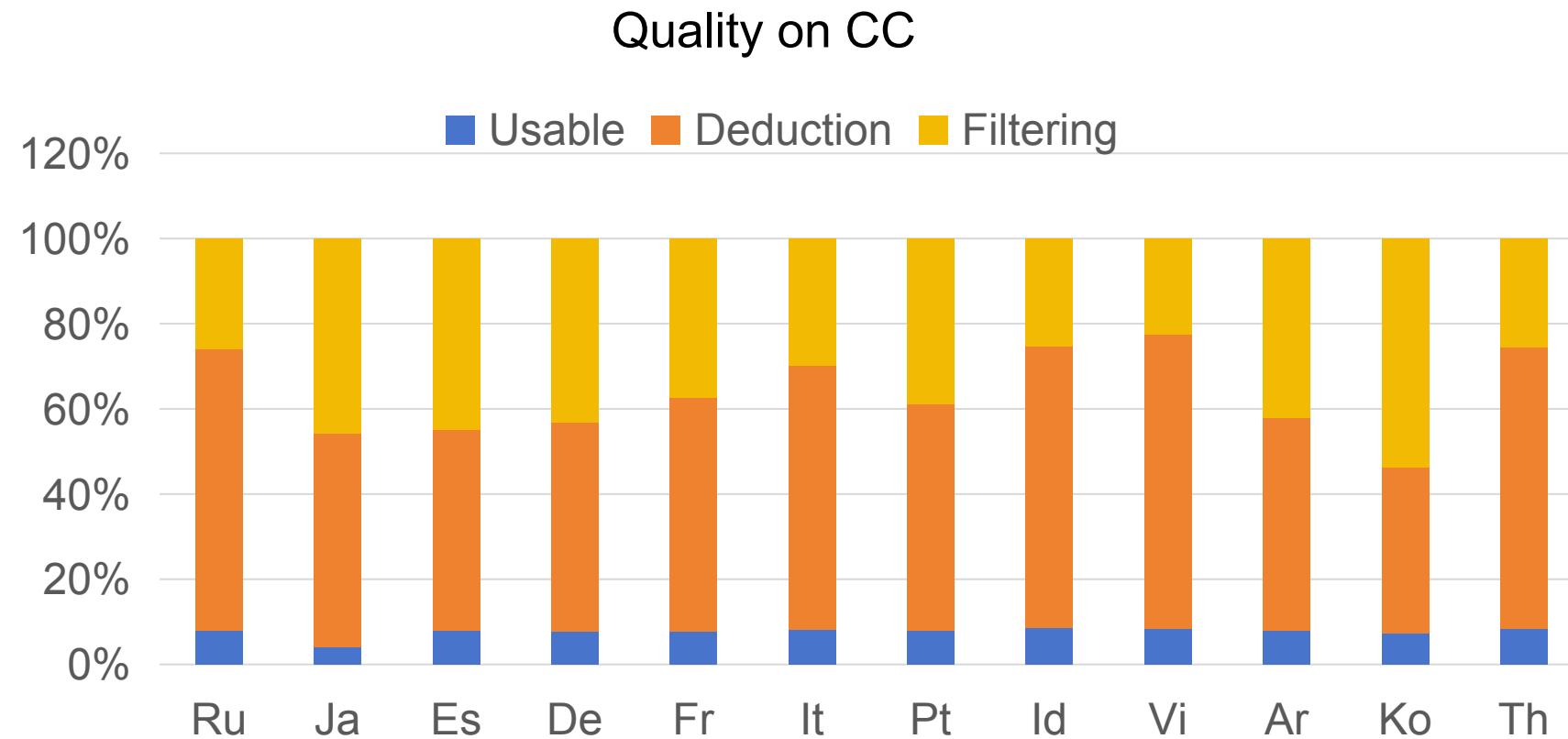


Only 1% languages have more than 100B tokens data

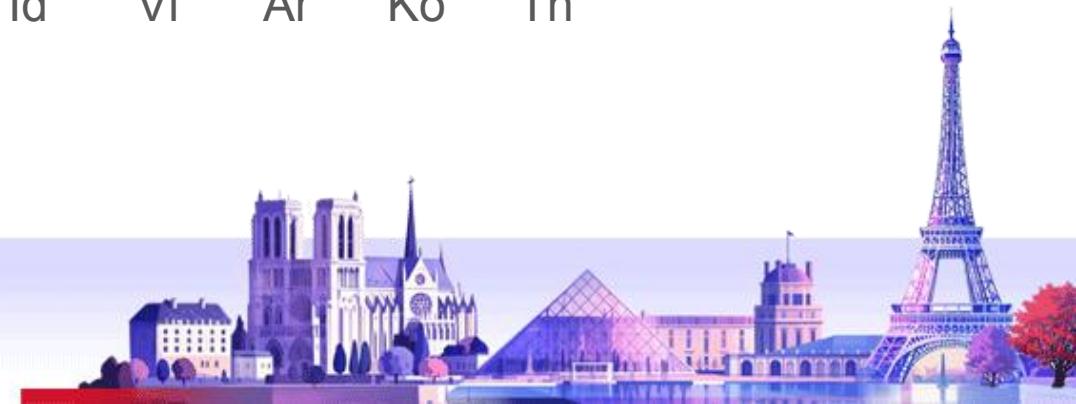


Data Unbalance

GOSIM

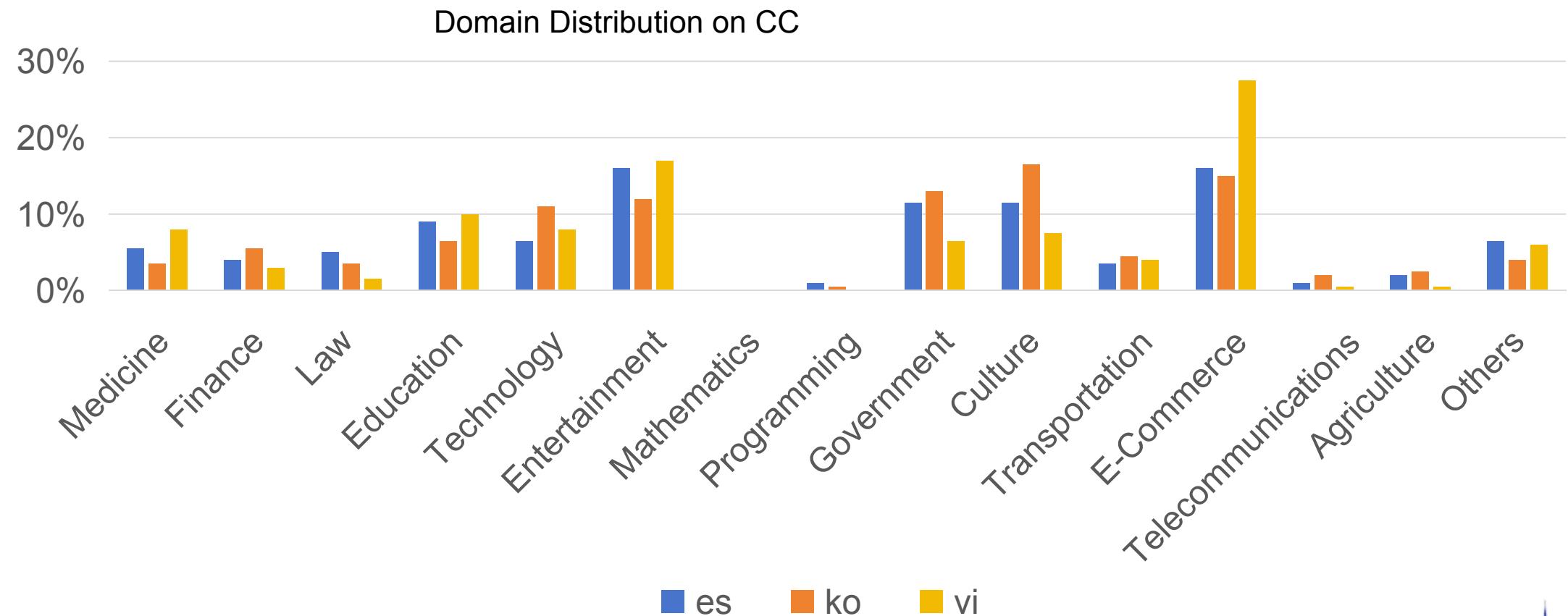


The vast majority of the data is of poor quality.



Data Unbalance

GOSIM

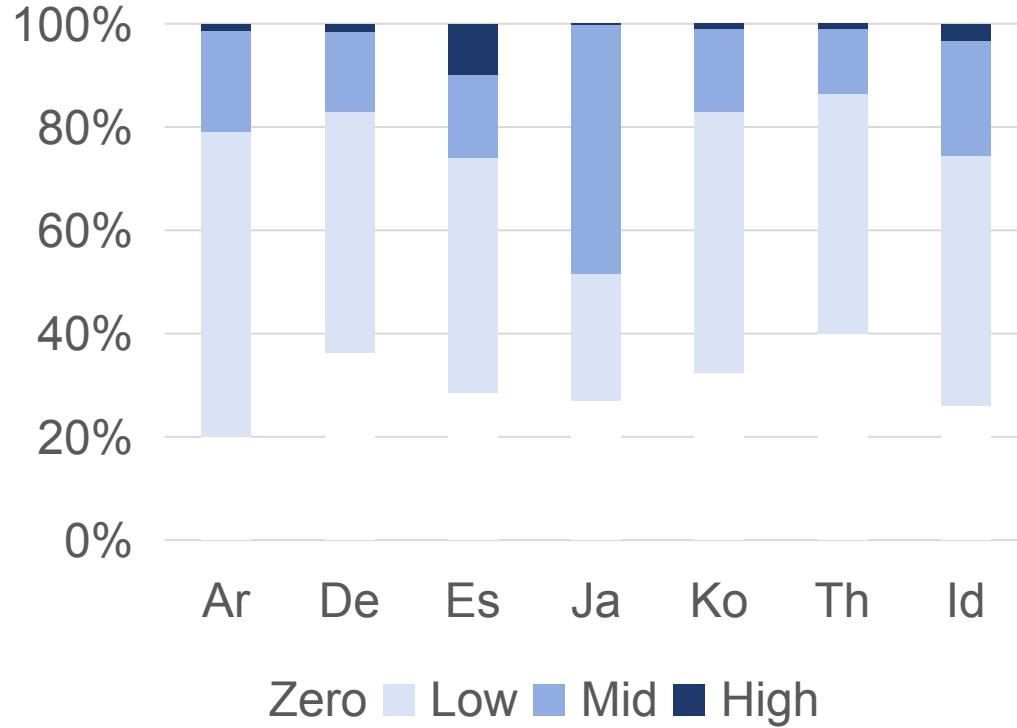


Significant differences in the domain distribution across different languages.

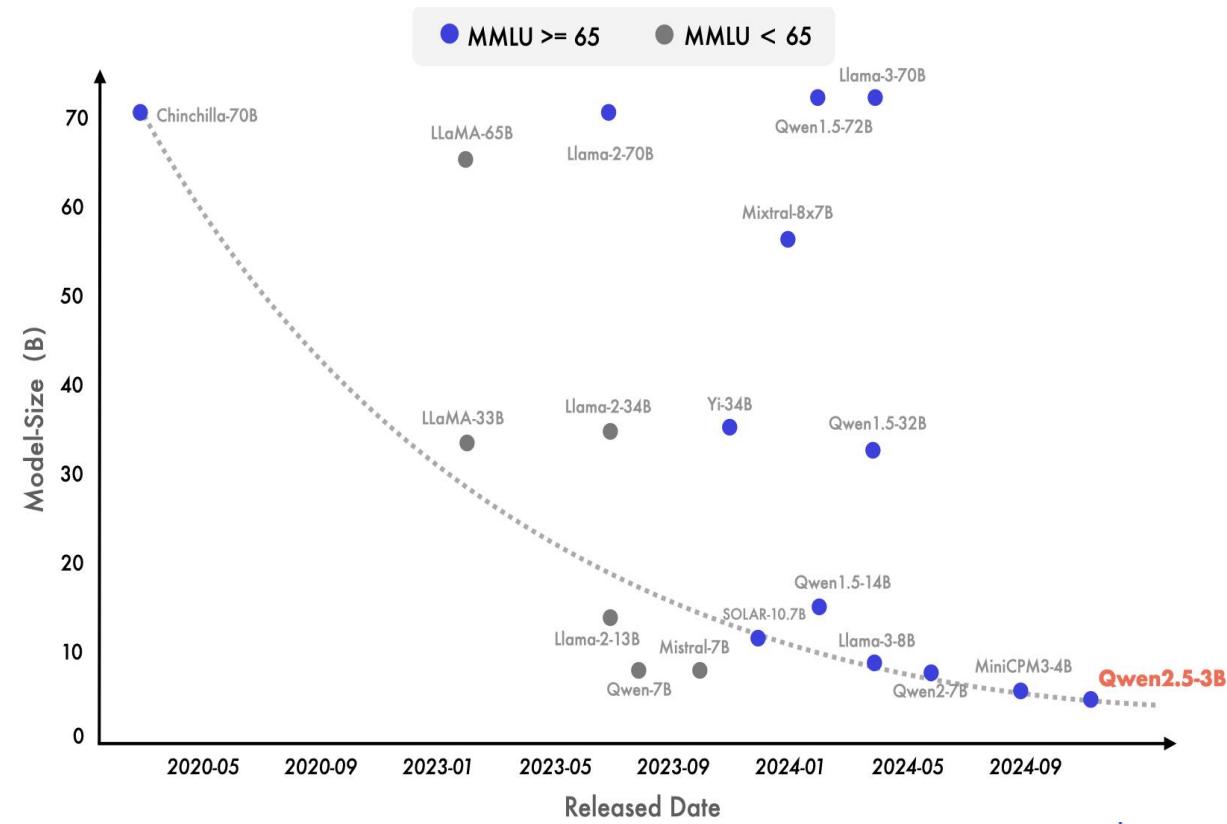
Data Unbalance

GOSIM

Knowledge Density on CC



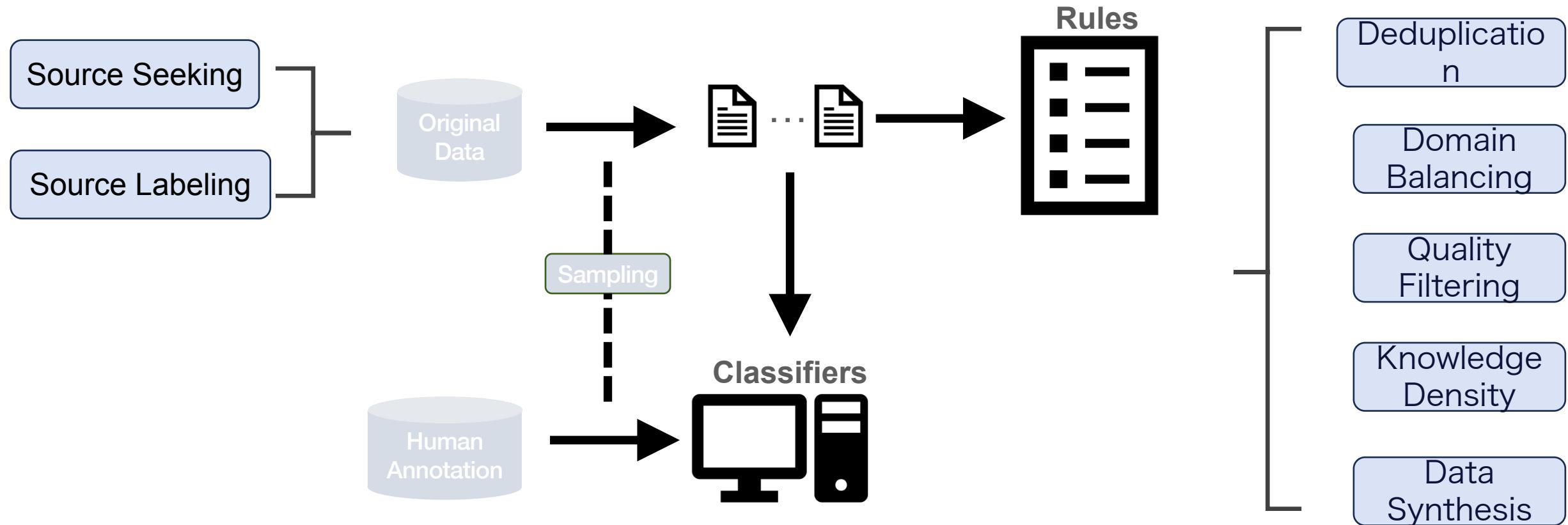
Labeling data with educational score (Penedo et al., 2024)



High knowledge density can improve model training efficiency.

Data Unbalance

GOSIM



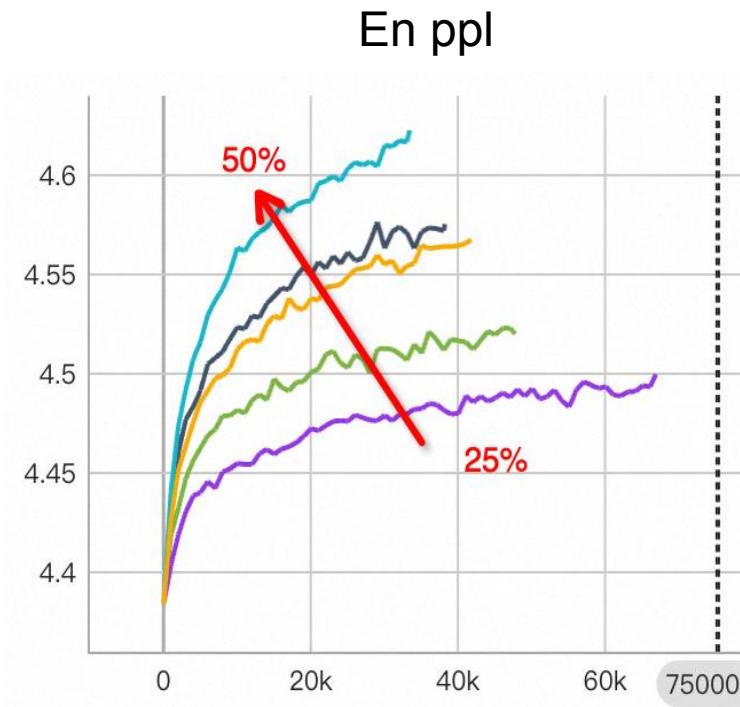
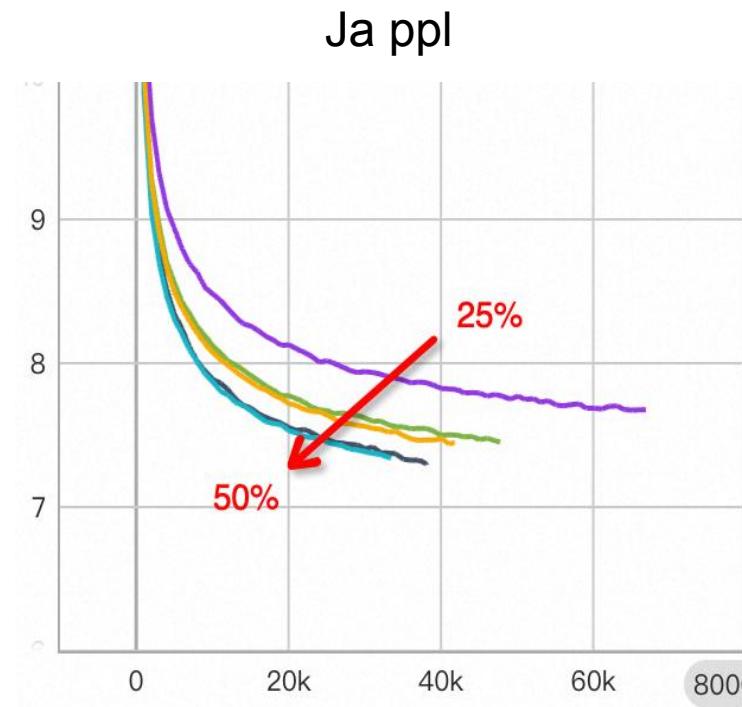
Relying on careful data engineering: data selection, cleaning, balancing, etc.

Predictable Training

GOSIM

data ratio, batch size, learning rate...

Experiments on Japanese-English data ratio:



How to select a better group of hyper-parameters?

Predictable Training

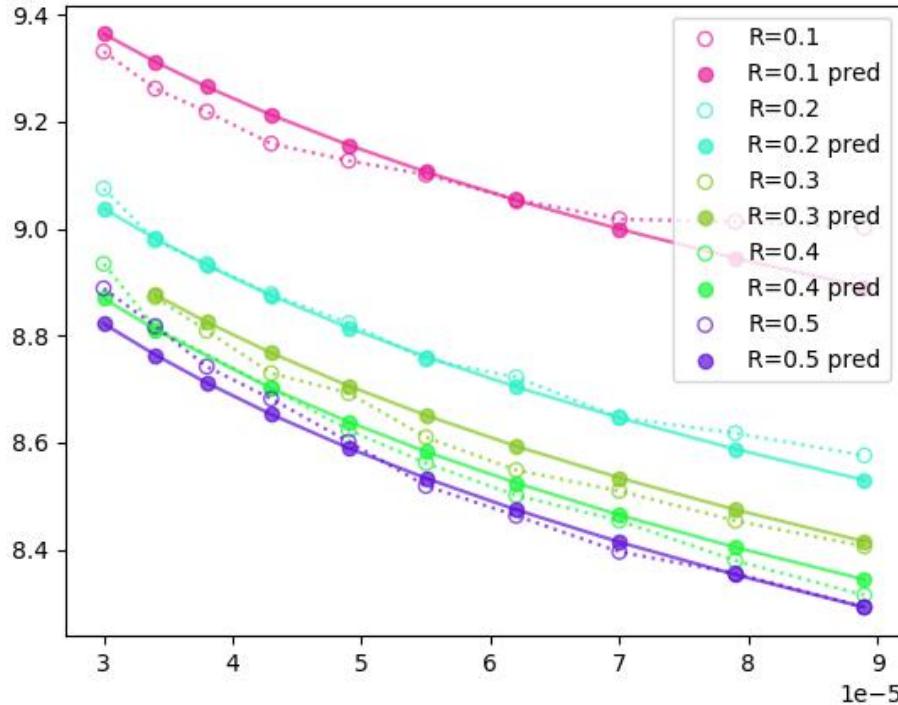
GOSIM

Massive preliminary experiments

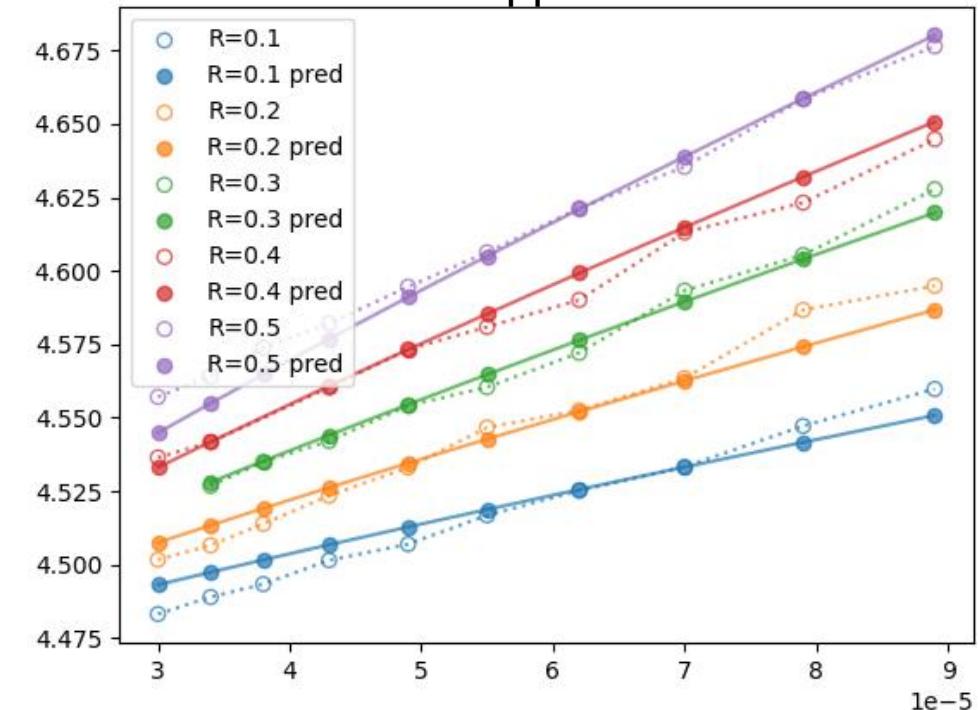


$$F(L, R) = x \cdot L^{\alpha_L} \cdot (y + R)^{\alpha_R}$$

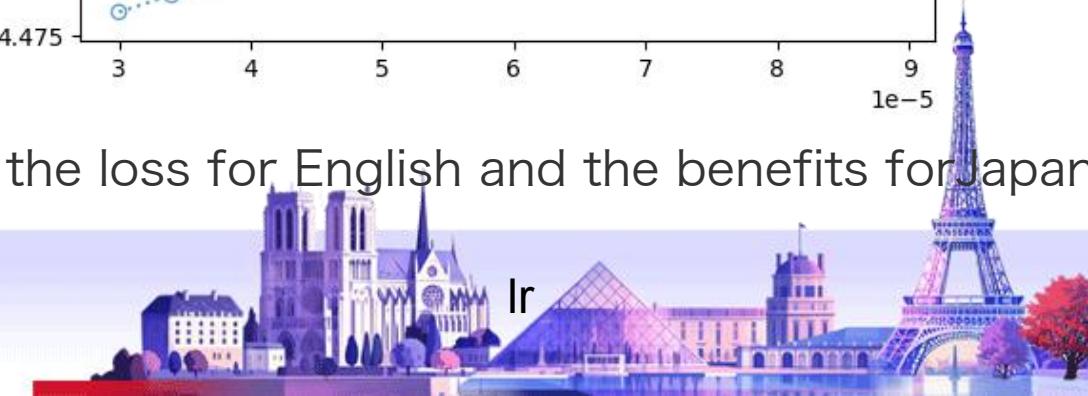
Ja ppl



En ppl



The optimal set of parameters can be found by balancing the loss for English and the benefits for Japan



Evaluation

GOSIM

To alleviate the lack of multilingual benchmarks in our community, we have released *P-MMEval* covering 10 languages, 7 kind of capabilities, all the examples are parallel

Exist:		We extend:	En	Zh	Ar	Es	Ja	Ko	Th	Fr	Pt	Vi
Generation	XL-SUM (Hasan et al., 2021)											
Translation	Flores (Costa-jussà et al., 2022)											
Understanding	XNLI Conneau et al., 2018)											
	XHellaswag (Zellers et al., 2019)											
Code	HumanEval-XL (Peng et al., 2024)											
Math	MGSM (Shi et al., 2023)											
Reasoning	LogiQA (Liu et al., 2020)											
Knowledge	MMMLU-hard (Hendrycks et al., 2021a)											

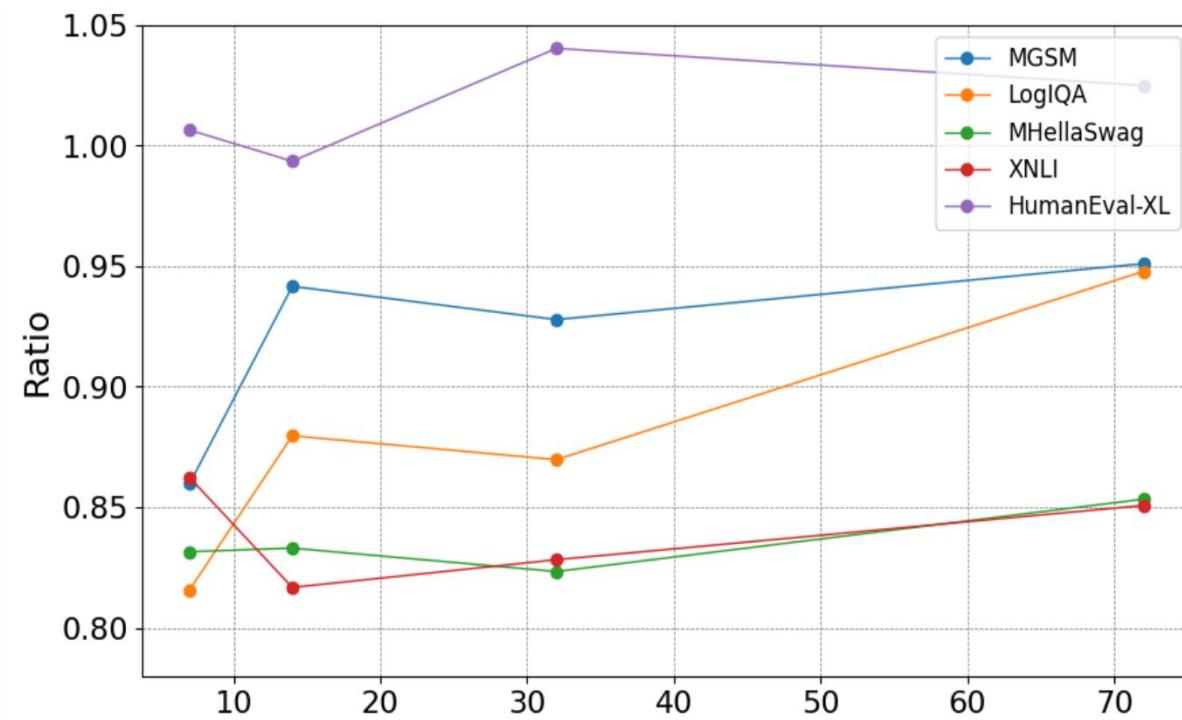
Zhang et al., [P-MMEval: A Parallel Multilingual Multitask Benchmark for Consistent Evaluation of LLMs](#). Arxiv 2024.



Evaluation

GOSIM

$$\text{Ratio} = \text{Score}_{\text{non-En}} / \text{Score}_{\text{En}}$$

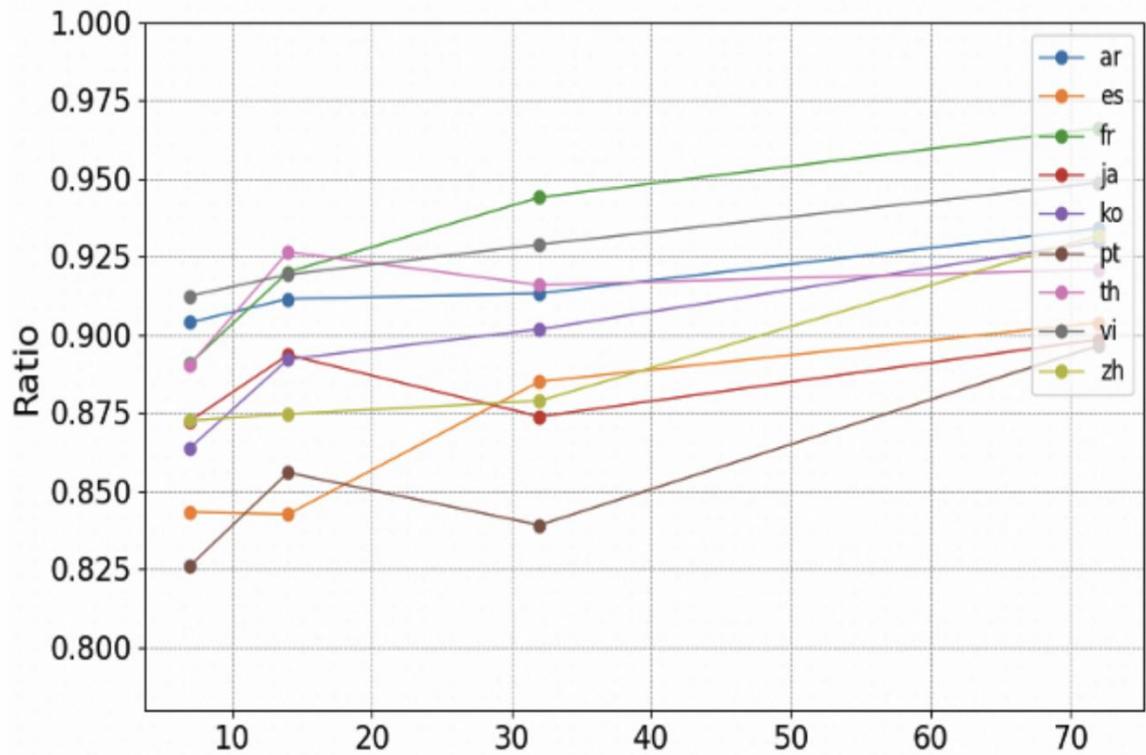


Qwen2.5

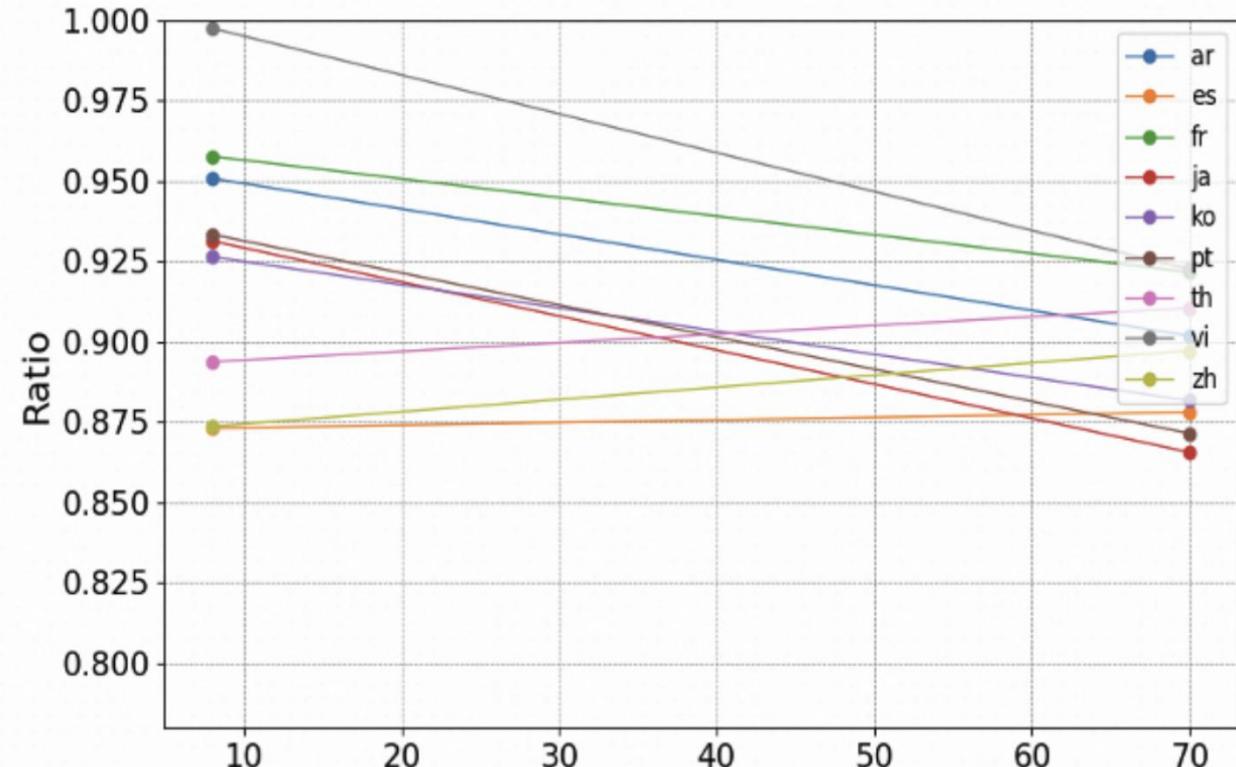
The transfer effects differ for different capabilities.

Evaluation

GOSIM



Qwen2.5



Llama3.1

Scaling can enhance the capability transfer for most of LLMs, except for Llama.



Evaluation

GOSIM

Lacking a challenging multilingual mathematical reasoning benchmark

Difficulty Level	Dimension		Problem Type
	Thought Depth	Knowledge Breadth	
Low	★	★	<ul style="list-style-type: none"> K-12 Mathematics: Basic algebra, geometry, and probability & statistics, primarily presented as Math Word Problems (MWP).
Medium	★★	★★	<ul style="list-style-type: none"> Exercises and Exams (High School & University): Post-class exercises from various math branches and authoritative entrance exams. Competitions (Low Difficulty): Publicly accessible competitions that are slightly more challenging than standard in-class exam problems.
High	★★★	★★	<ul style="list-style-type: none"> Competitions (Mid-to-High Difficulty): Problems that require critical thinking but do not demand deep theoretical knowledge. In comparison to the competitions at the medium level, the participants in these contests have already undergone preliminary selection.
Top	★★★★ → ∞	★★★ → ∞	<ul style="list-style-type: none"> Competitions (Top Olympiad): The highest-tier international/national/regional mathematics Olympiads, representing the upper limits of human IQ. Frontier Mathematics: Advanced mathematical disciplines and emerging research areas, approaching the limits of human mathematical systems.

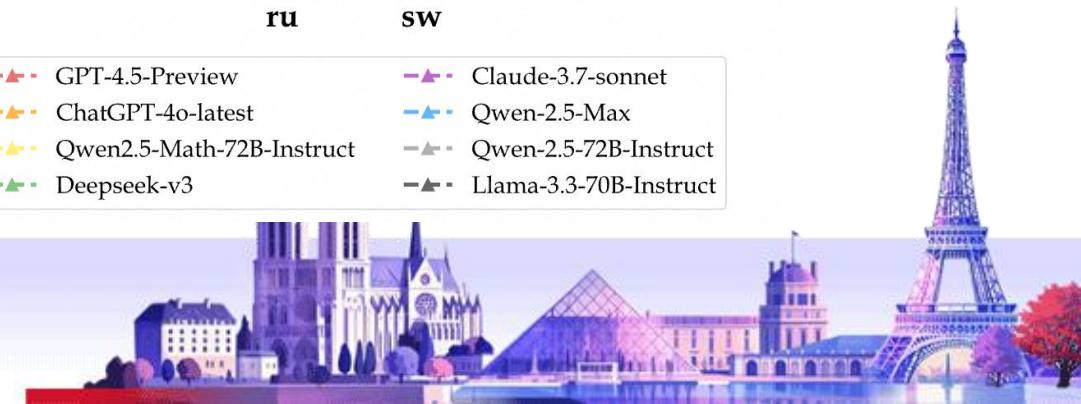
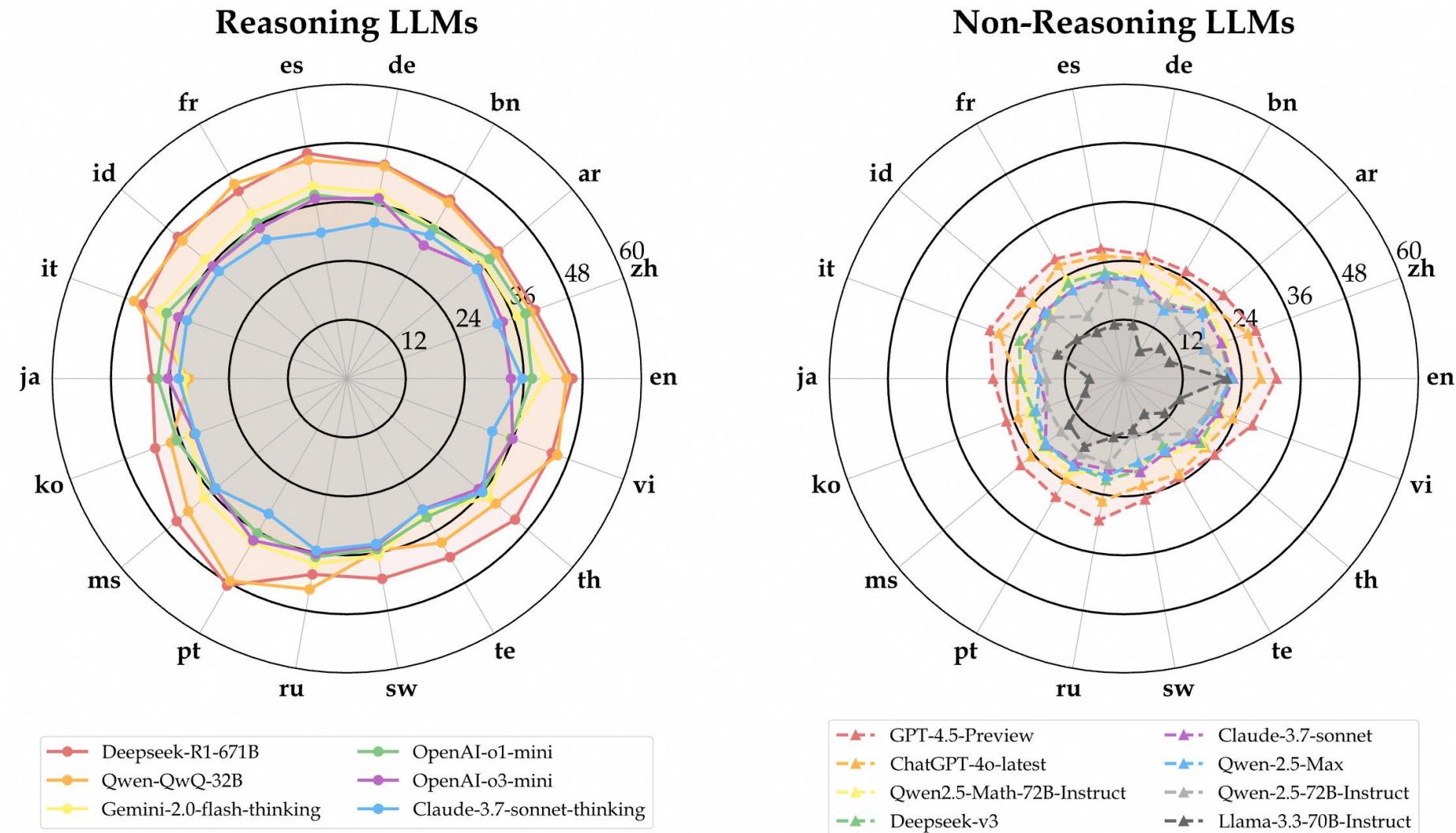
Code	Full Name	Language Family	Native Speakers (M)
en	English	Indo-European	1,500
zh	Chinese	Sino-Tibetan	1,400
es	Spanish	Indo-European	595
ar	Arabic	Afro-Asiatic	400
fr	French	Indo-European	300
bn	Bengali	Indo-European	300
pt	Portuguese	Indo-European	270
ru	Russian	Indo-European	260
id	Indonesian	Austronesian	200
de	German	Indo-European	135
ja	Japanese	Japonic	130
sw	Swahili	Niger-Congo	100
vi	Vietnamese	Austroasiatic	86
it	Italian	Indo-European	85
te	Telugu	Dravidian	81
ko	Korean	Koreanic	80
th	Thai	Kra-Dai	80
ms	Malay	Austronesian	77

WANG ET AL., 2025. POLYMATH: Evaluating Mathematical Reasoning in Multilingual Contexts



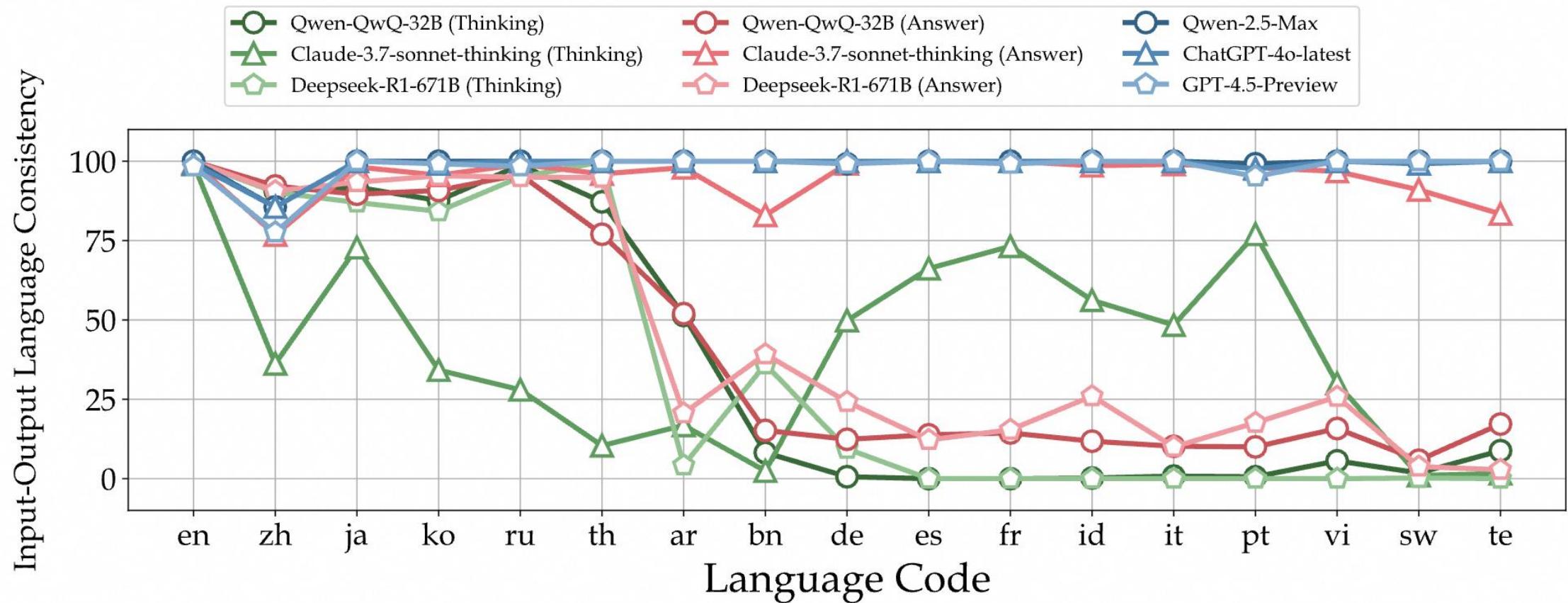
Evaluation

GOSIM

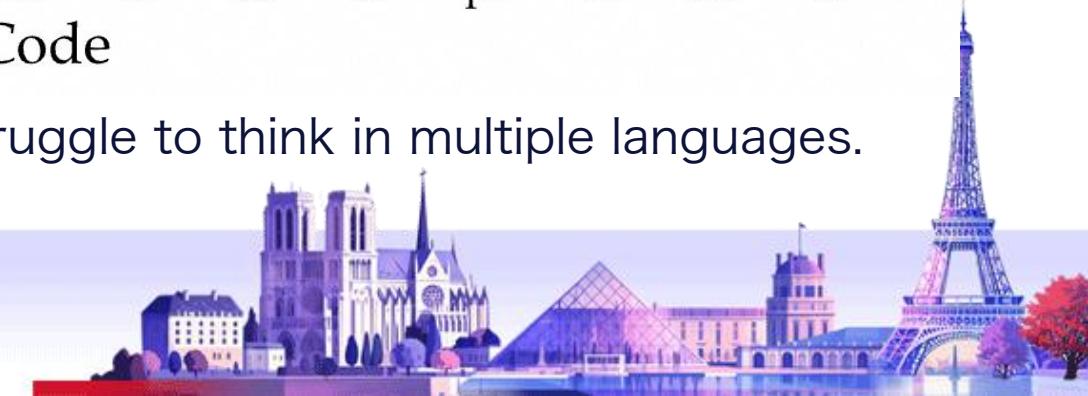


Evaluation

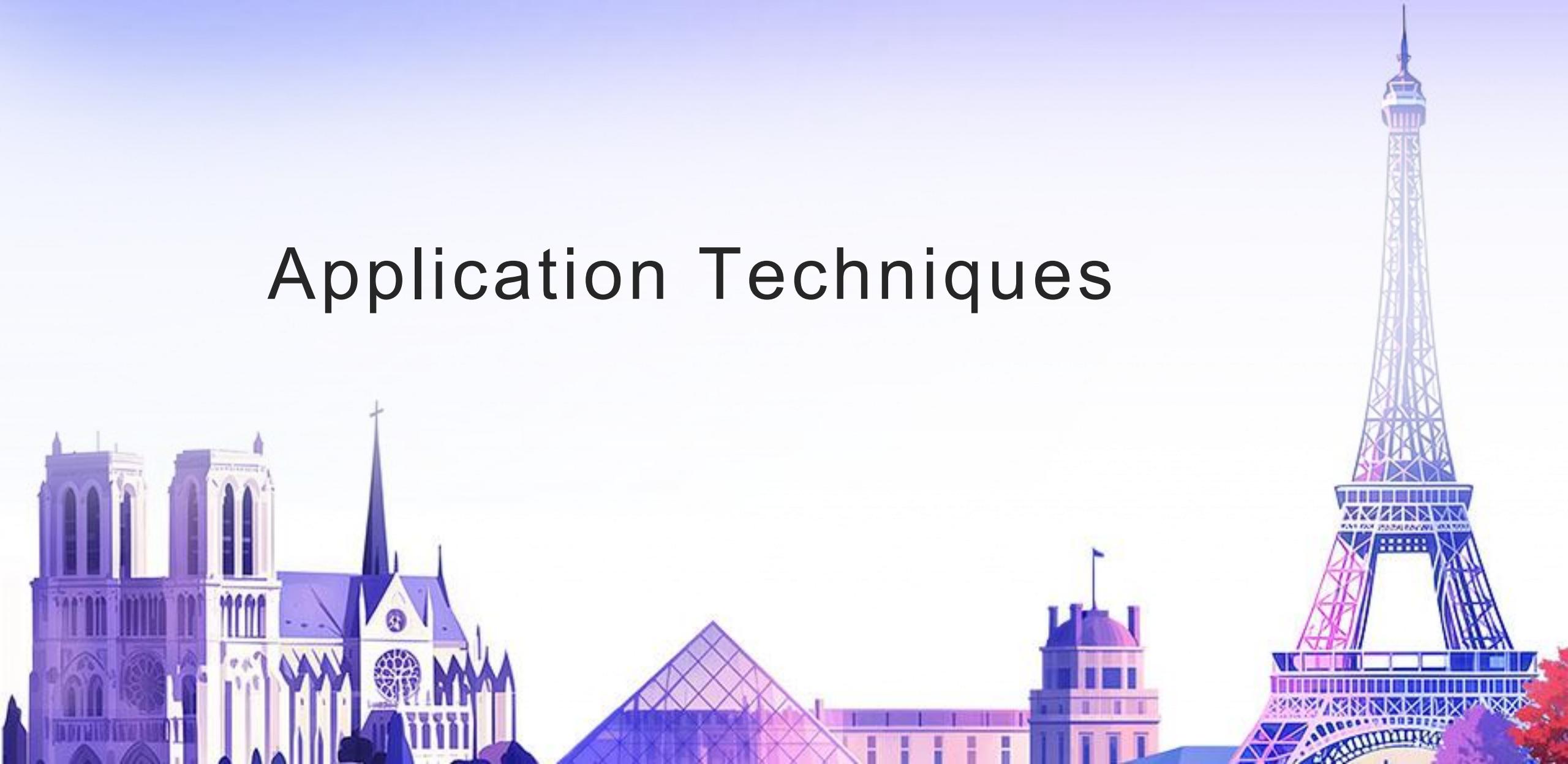
GOSIM



Currently, long reasoning models struggle to think in multiple languages.



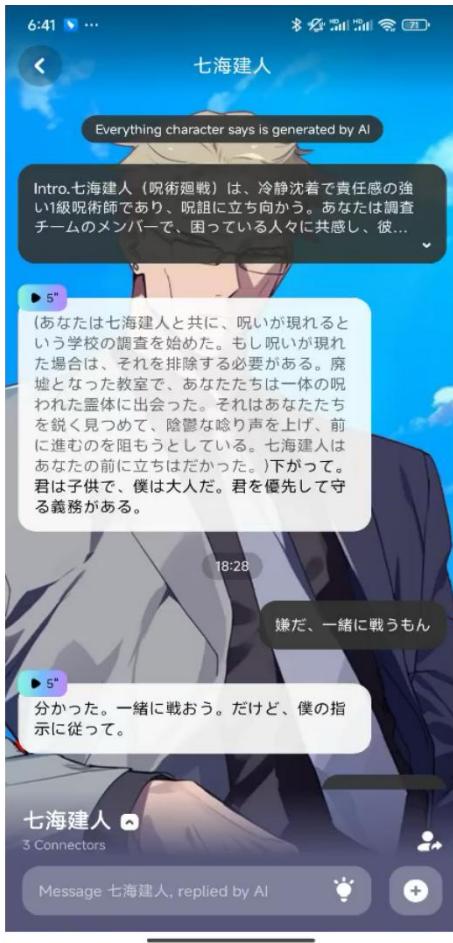
Application Techniques



Applications of Multilingual LLM

GOSIM

Role-Playing



Personal Assistant



Oral Practice Partner



Applications of Multilingual LLM

GOSIM

Efficiency Tools

2311.07919(2)

2311.07919(2) arXiv:2311.07919v2 [eess.AS] 21 Dec 2023

[Code & Demo & Models: https://github.com/QwenLM/Qwen-Audio](https://github.com/QwenLM/Qwen-Audio)

Abstract

Recently, instruction-following audio-language models have received broad attention for audio interaction with humans. However, the absence of pre-trained audio models capable of handling diverse audio types and tasks has hindered progress in this field. Consequently, most existing works have only been able to support a limited range of interaction capabilities. In this paper, we develop the Qwen-Audio model and address this limitation by scaling up audio-language pre-training to cover over 30 tasks and various audio types, such as human speech, natural sounds, music, and songs, to facilitate universal audio understanding abilities. However, directly co-training all tasks and datasets can lead to interference issues, as the textual labels associated with different datasets exhibit considerable variations due to differences in task focus, language, granularity of annotation, and text structure. To overcome the one-to-many interference, we carefully design a multi-task training framework by conditioning on a sequence of hierarchical tags to the decoder for encouraging knowledge sharing and avoiding interference through shared and specified tags respectively. Remarkably, Qwen-Audio achieves state-of-the-art performance across diverse benchmark tasks without requiring any task-specific fine-tuning, surpassing its counterparts. Building upon the capabilities of Qwen-Audio, we further develop Qwen-Audio-Chat, which allows for input from various audio and text inputs, enabling multi-turn dialogues and supporting various audio-centric scenarios.

1 Introduction

Large language models (LLMs) (Brown et al., 2020; OpenAI, 2022, 2023; Chowdhery et al., 2022; Anil et al., 2023; Touvron et al., 2023a; Qwen, 2023) have greatly propelled advancements in the field of general artificial intelligence (AGI) due to their strong knowledge retention, complex reasoning and problem-solving capabilities. However, language models lack the capability to perceive non-textual modalities like images and audio in the same manner as humans do. Speech, as an important modality, provides diverse and complex signals beyond texts such as emotions, tones, and intentions in human voice, train whistle, clock chime and thunder in natural sounds, and melody in music. Enabling LLMs to perceive and comprehend rich audio signals for audio interaction has received broad attention (Huang et al., 2023; Shen et al., 2023; Wang et al., 2023a; Lyu et al., 2023; Wu et al., 2023b; Gong et al., 2023b; Wang et al., 2023c; Shu et al., 2023). Prior works for instruction following mainly inherit the capabilities from large (multimodal) LLMs and adopt light-weight supervised fine-tuning to activate the abilities of the model to align with user intent (Ouyang et al., 2022; Wang et al., 2023a; Gong et al., 2023b). However, most works have been constrained in terms

*Equal contribution. **Corresponding author

1 / 18 - 100% +

Cross-Border E-commerce

The image shows four identical product listing cards for a "Small Y humidifier" displayed in four different languages: English, French, Spanish, and Portuguese. Each card includes a promotional banner at the bottom.

English: C\$ 8.70 € 6.22 SUPER PROMOTION
€ 5.14 off every C\$ 41.44 spent on promo | 5% cashback | Extra 2% Off

French: € 5,29 € 3,24 SUPER PROMOTION
€ 2 tous les € 25 dépensés sur la promo | 2% de réduction bonus

Spanish: € 5,29 € 3,24 SUPER OFERTAS
€ 2 dtos. cada € 25 gastados en la promo | 2% de descuento extra

Portuguese: R\$ 29,99 R\$ 18,72 SUPER PROMOÇÃO
R\$ 15 off a cada R\$ 150 em compras na promo | 5% de Cashback | 2% off extra

Each card also displays the product name, price, promotional details, and a call-to-action button.

Continue Pretraining

GOSIM

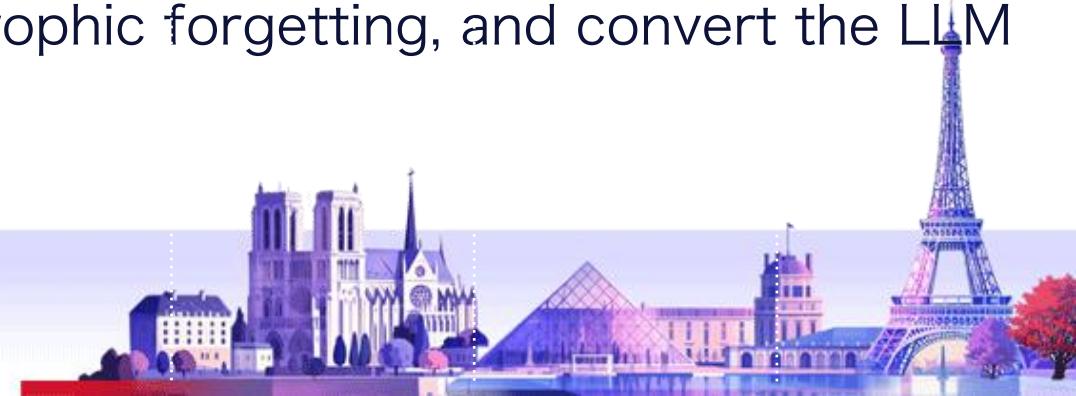
Problem:

High ratio -> catastrophic forgetting

Low ratio -> computational cost

Example: 10 B tokens for the language pending to enhance, using 20% as the ratio of CT, it requires 40 B replay data for training!

Approach: Utilize a mixture of experts model to store multilingual knowledge, freeze the original feedforward layers of the LLM to alleviate catastrophic forgetting, and convert the LLM into a multilingual LLM with no loss or minimal loss.

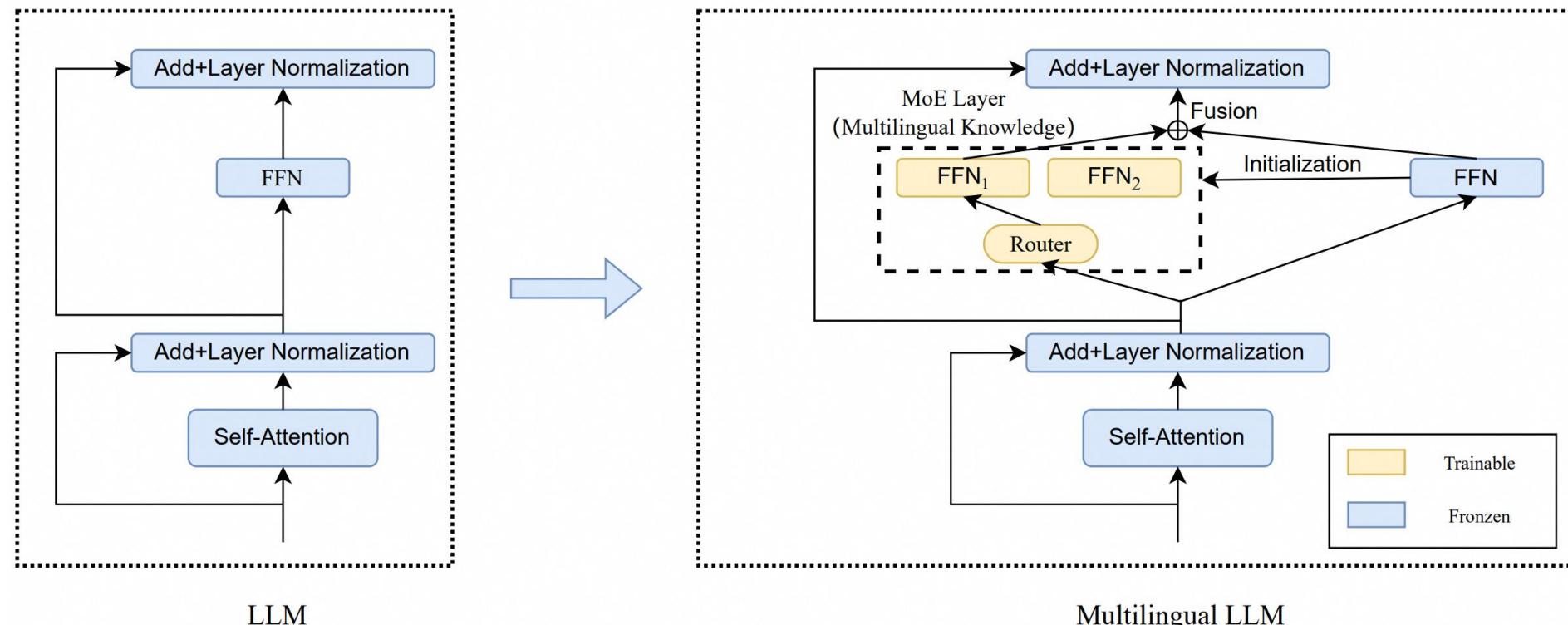


Continue Pretraining

GOSIM

Solution:

- Retaining the original parameters to prevent catastrophic forgetting
- Enhancing multilingual capabilities through Experts
- Initializing parameters with the original FFN layer to inherit existing knowledge



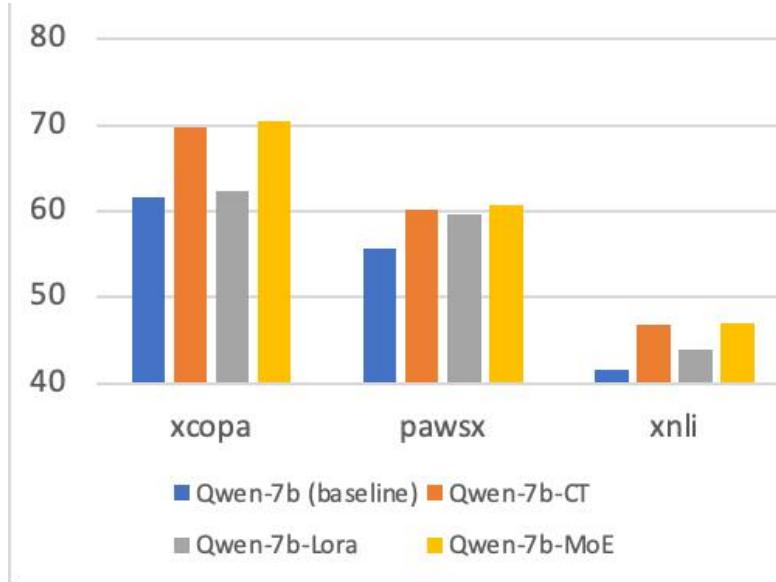
Continue Pretraining

GOSIM

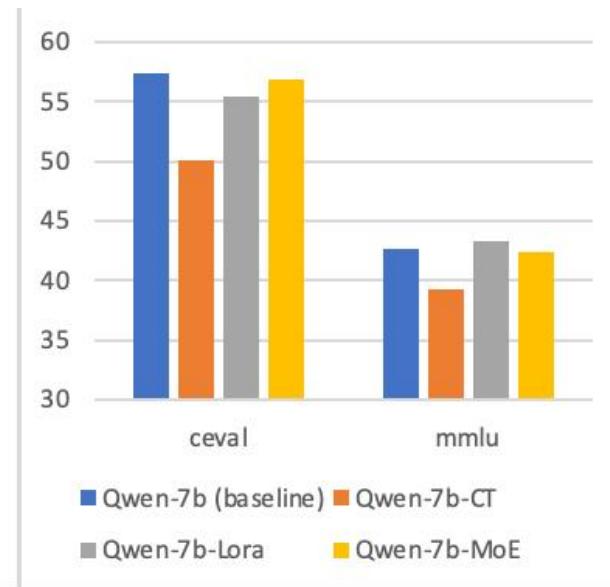
1. Increasing the ratio to 50%, almost no loss in Chinese-English performance, and the multilingual performance is comparable to that of CT.

2. Training only requires 40% of the original training tokens to achieve comparable results!

Multilingual Benchmarks



Conventional Capabilities



Li et al., MMNMT: Modularizing Multilingual Neural Machine Translation with Flexibly Assembled MoE and Dense Blocks. In EMNLP 2023.

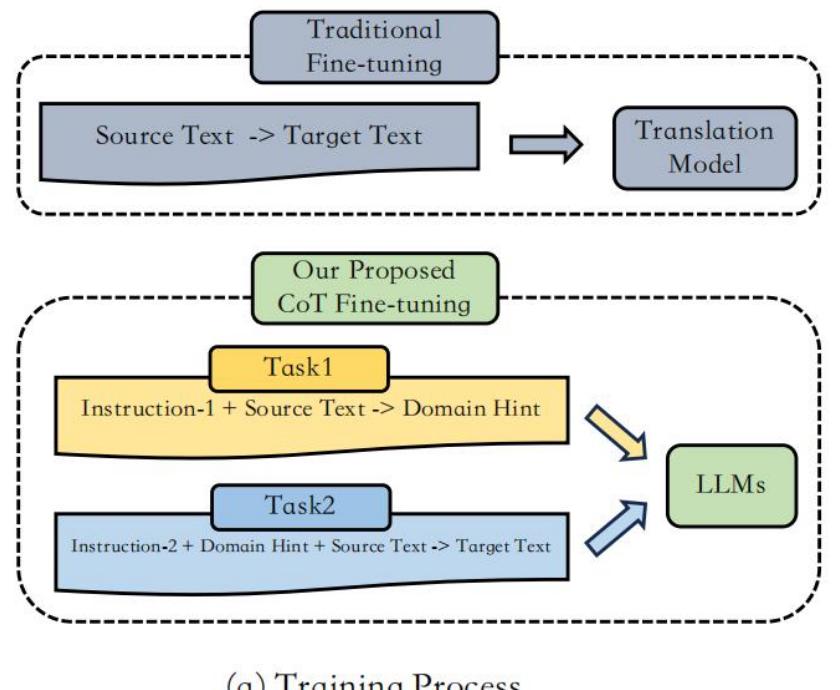
Li et al., MoE-CT: A Novel Approach For Large Language Models Training With Resistance To Catastrophic Forgetting. Arxiv 2024

Post-Training – An example on Machine Translation

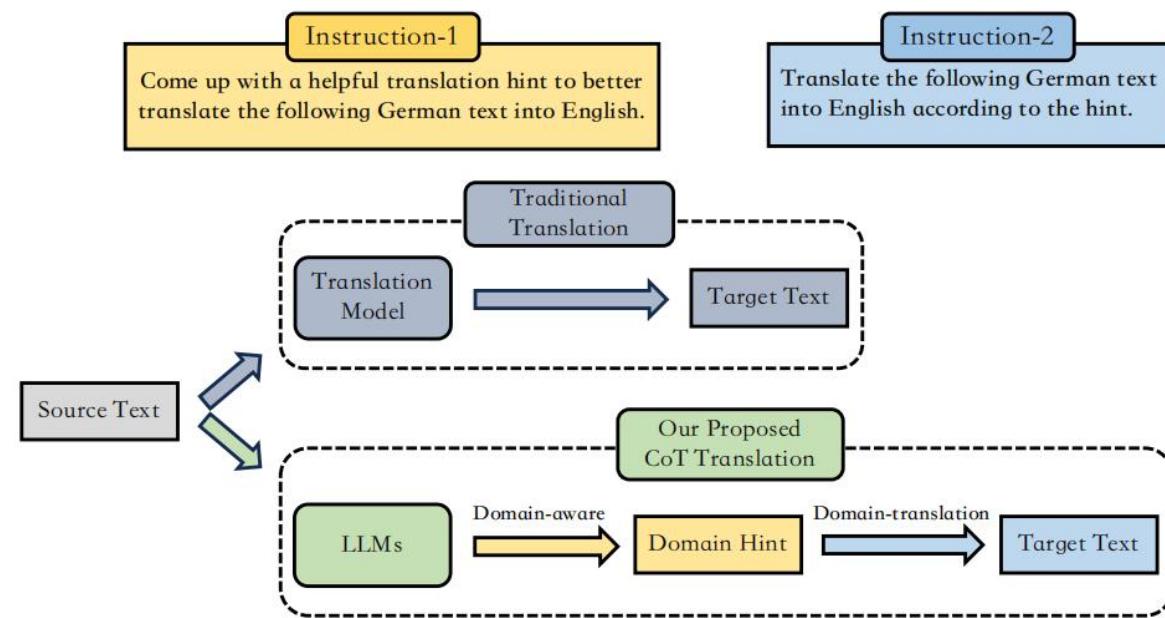
GOSIM

LLMs still face overfitting issues, lack high-quality parallel data across multiple domains, and experience catastrophic forgetting during training.

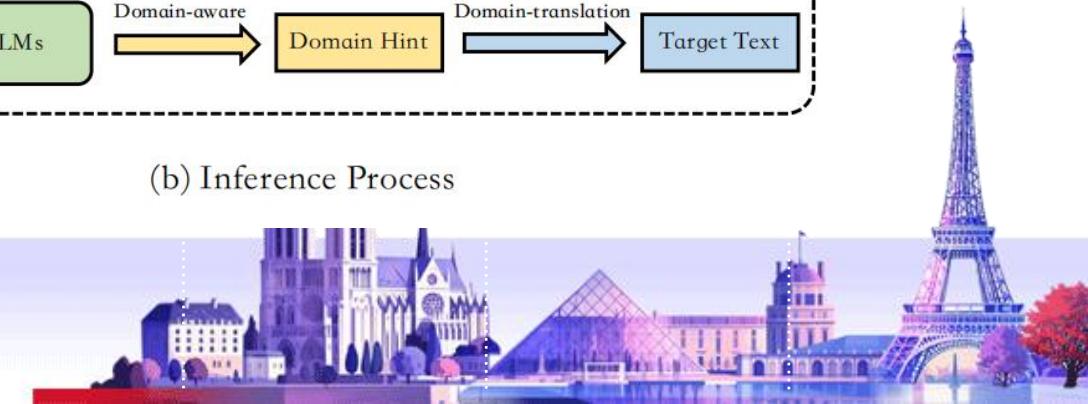
Solution: Enable the model to understand the input sentences and generate corresponding domain prompts to activate the LLM's domain knowledge, thereby producing translations with different domain styles.



(a) Training Process

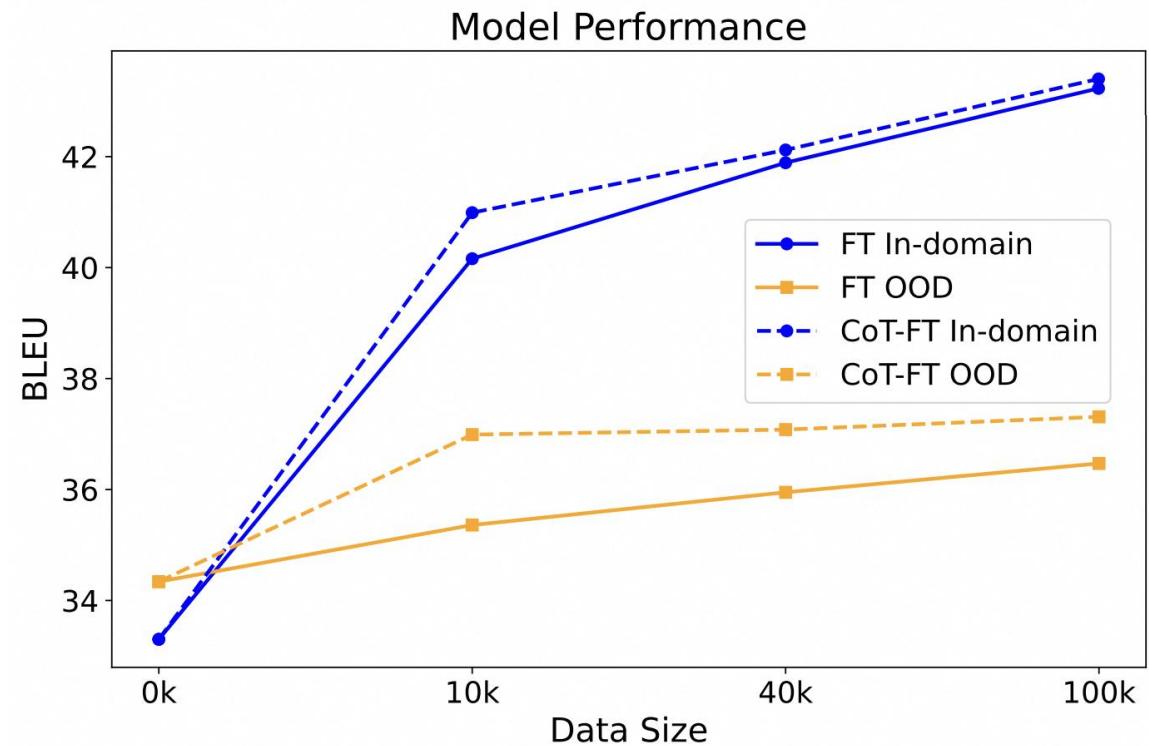
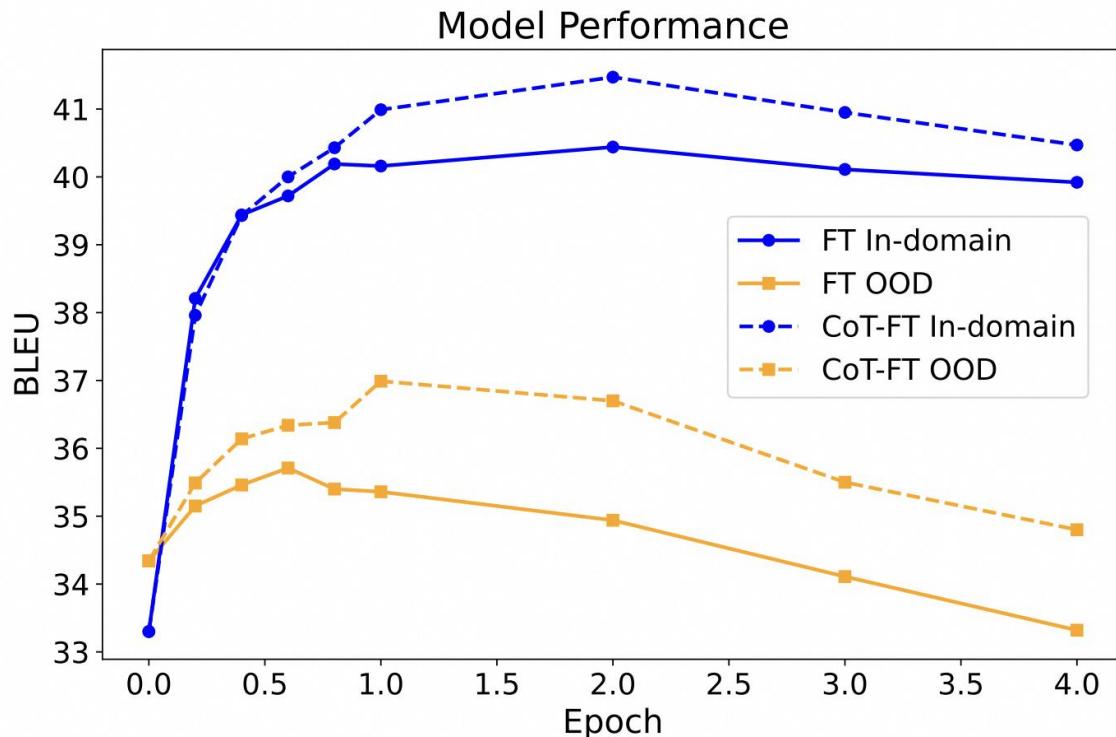


(b) Inference Process



Post-Training – An example on Machine Translation

GOSIM



The CoT-FT strategy effectively alleviates overfitting and shows more significant improvements on OOD data.

Hu et al., Large Language Model for Multi-Domain Translation: Benchmarking and Domain CoT Fine-tuning. In EMNLP 2024.

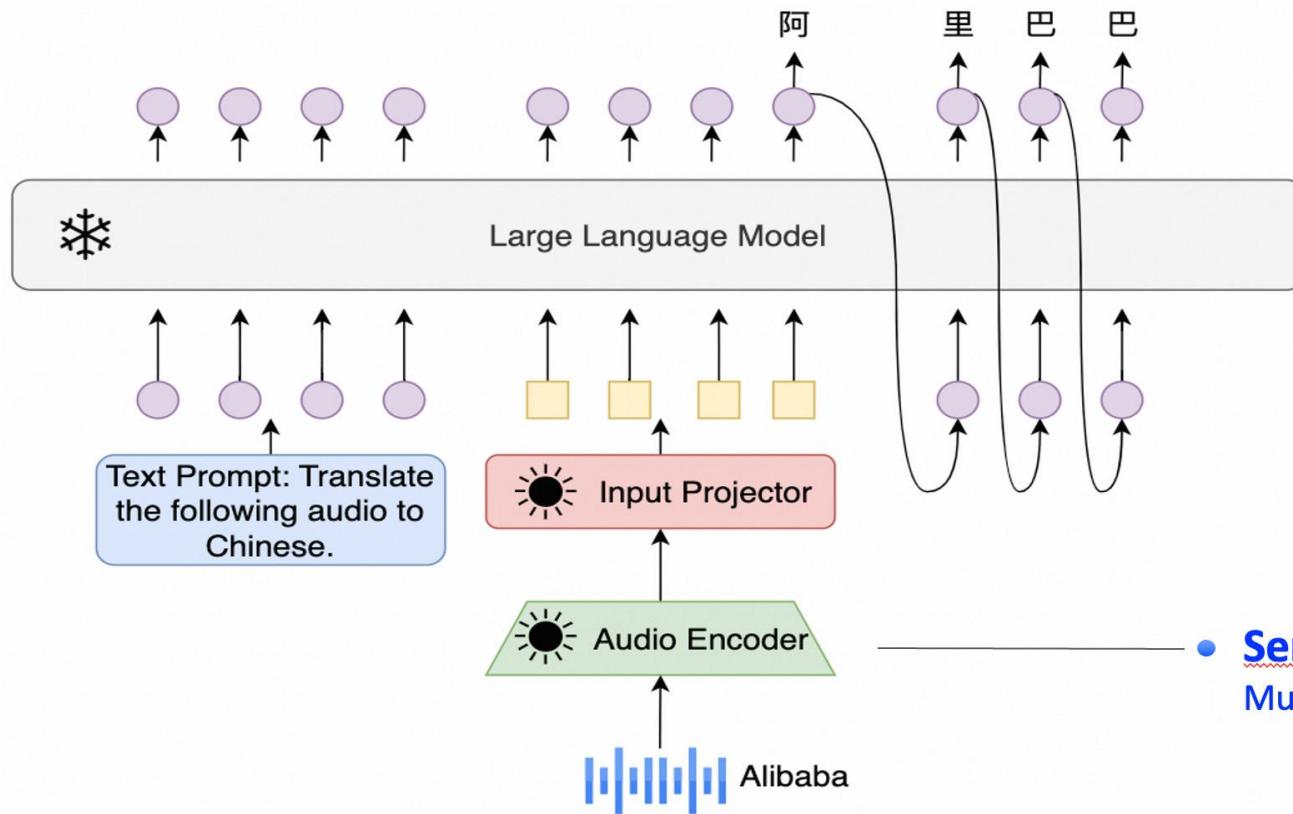
Cascaded Speech Translation

Chaining automatic Speech Recognition (ASR) model, a punctuation model, and a Machine Translation (MT) model



Cross-Lingual & Cross-Modal

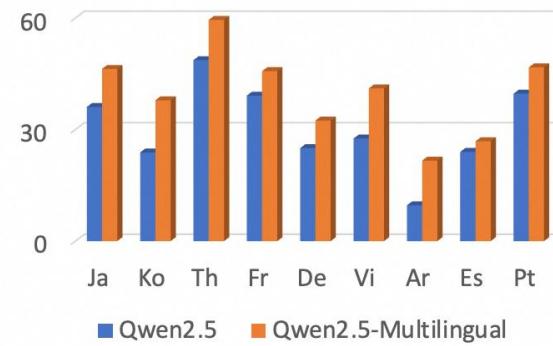
GOSIM



End-to-End Speech Translation

Multilingual Owen

Continually training using multilingual texts



- **SenseVoice (An et al., 2024)**
Multilingual speech foundation model

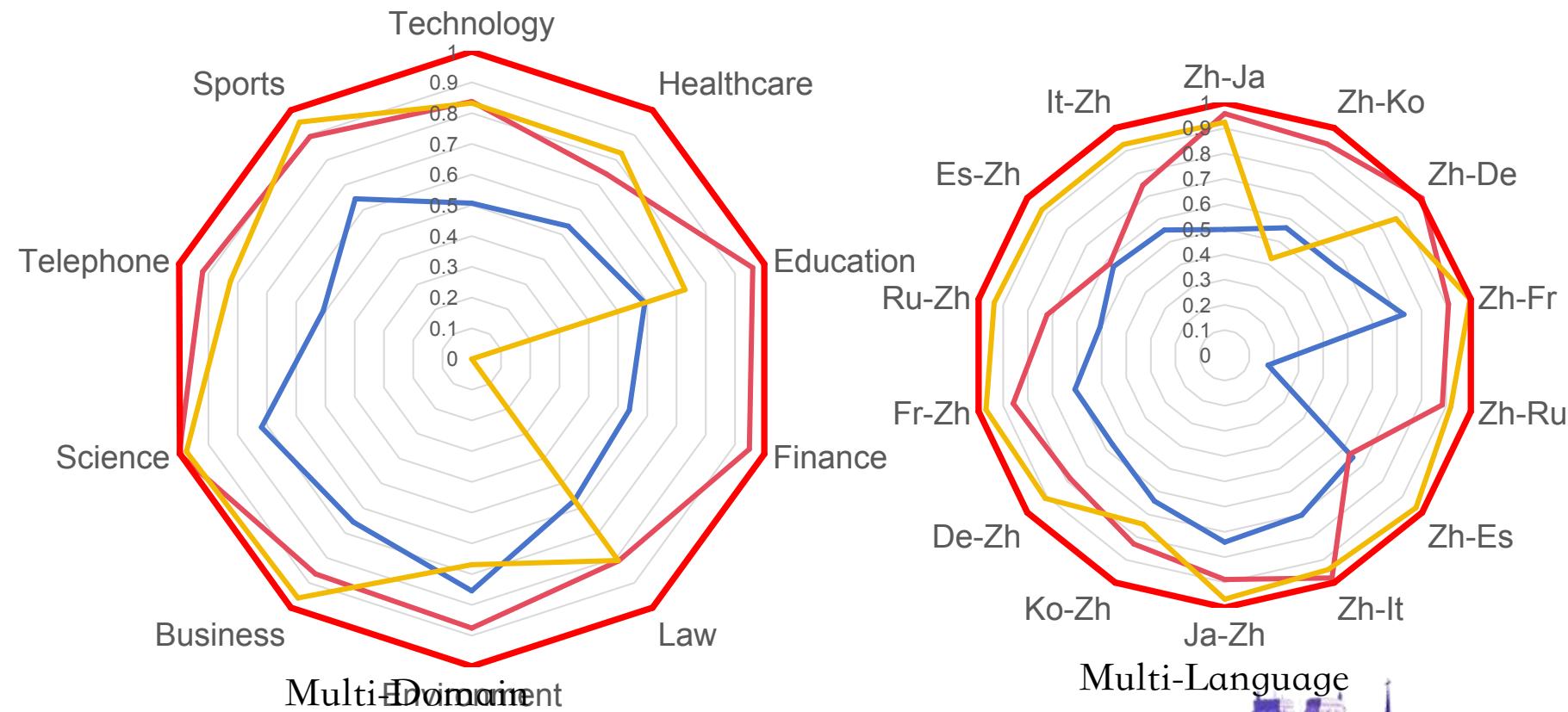


Cross-Lingual & Cross-Modal

GOSIM

Normalized Translation Quality (BLEU)

— Seamless (Meta) — Commercial API — Ours-Cascaded Engine — Ours-Gummy1.0



Cross-Lingual & Cross-Modal

GOSIM

Waiting for long dependency reordering

Input

区块链是今年



区块链是今年最火



区块链是今年最火热的科技趋势了



Output

Blockchain is

Blockchain is the most trendy

Blockchain is the most trendy technology of the year.

Prediction according to contexts

Input

所有草莓的基因都由几组含七条染色体的



所有草莓的基因都由几组含七条染色体的染色体组构造

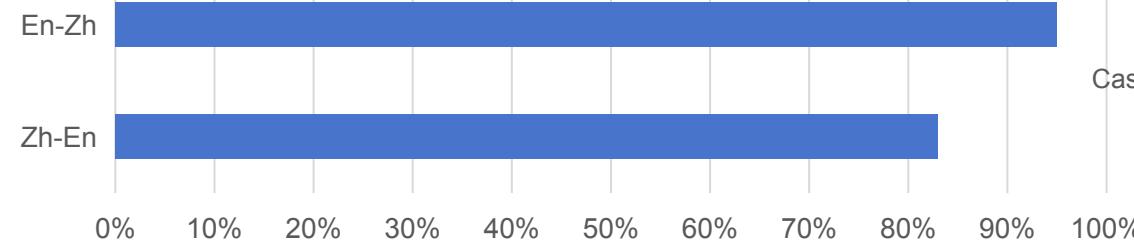


Output

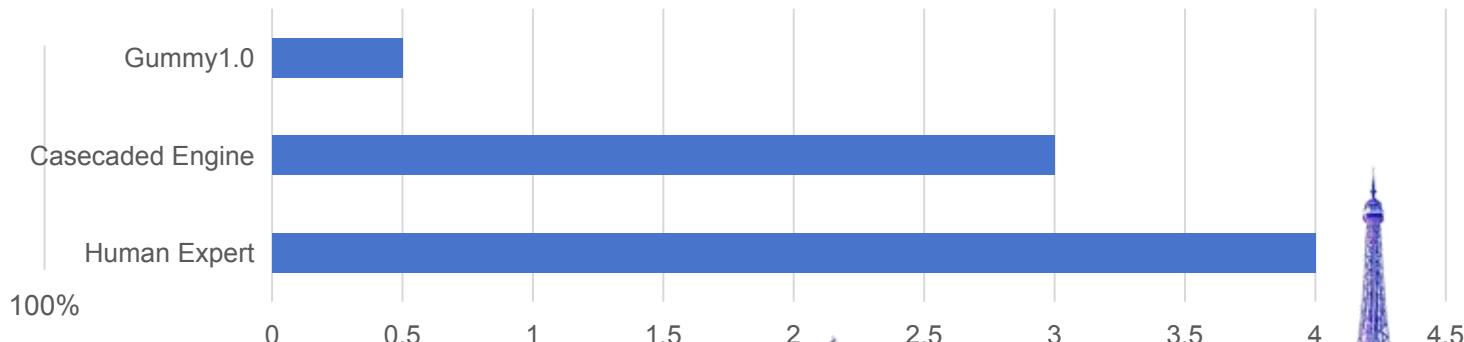
All genes of strawberry are consist of

All genes of strawberry are consist of genomes containing seven chromosomes.

Quality Retention Rate in Streaming Translation

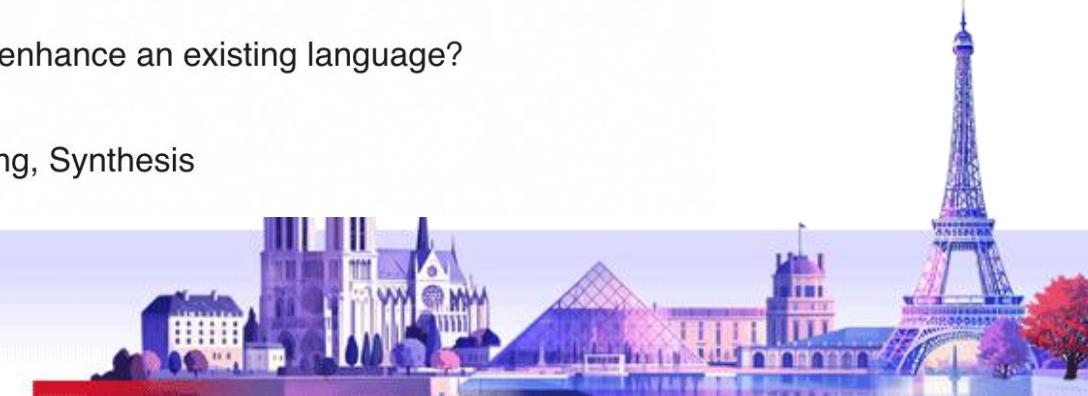
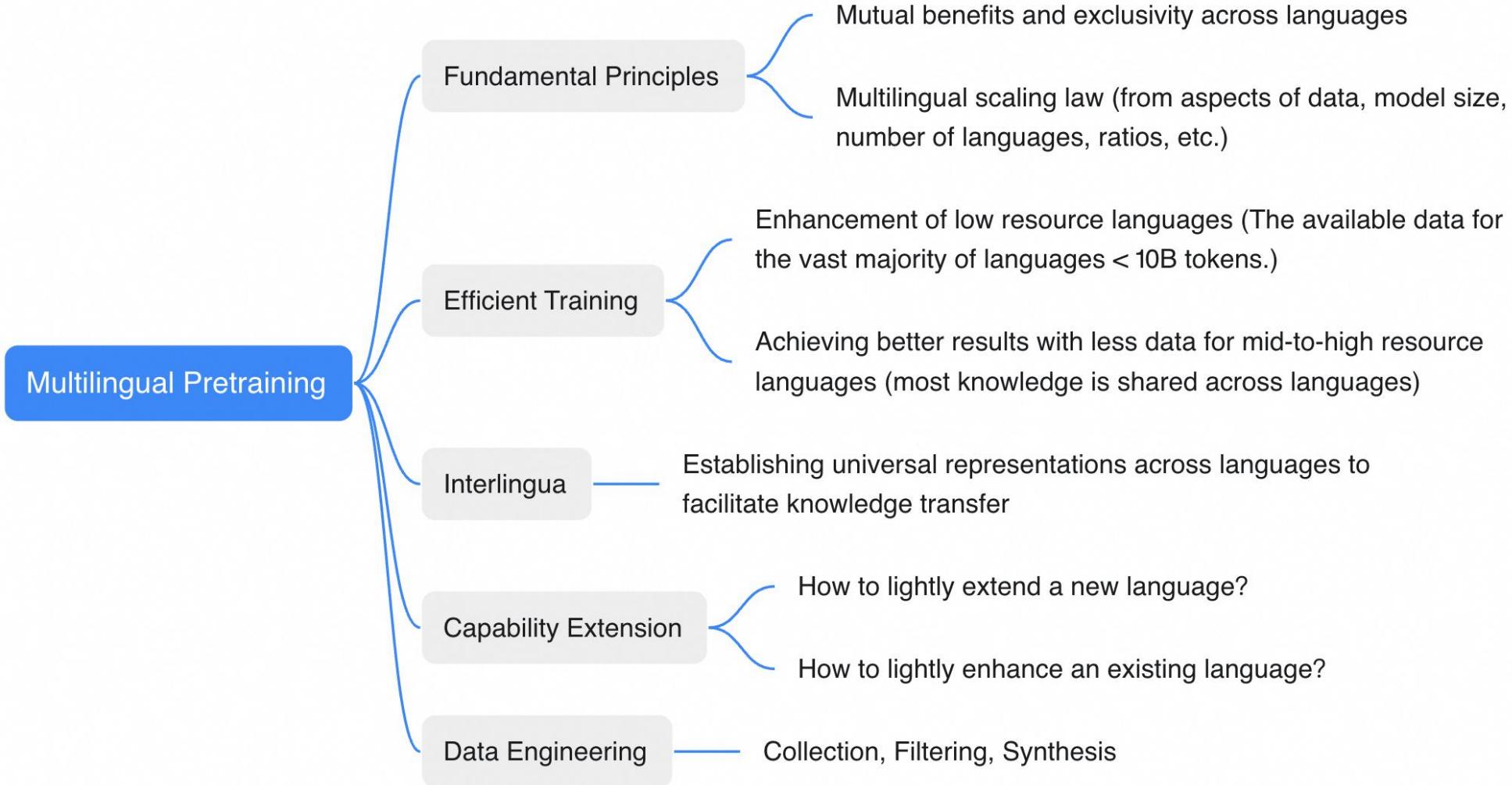


Stream Translation Latency (s)



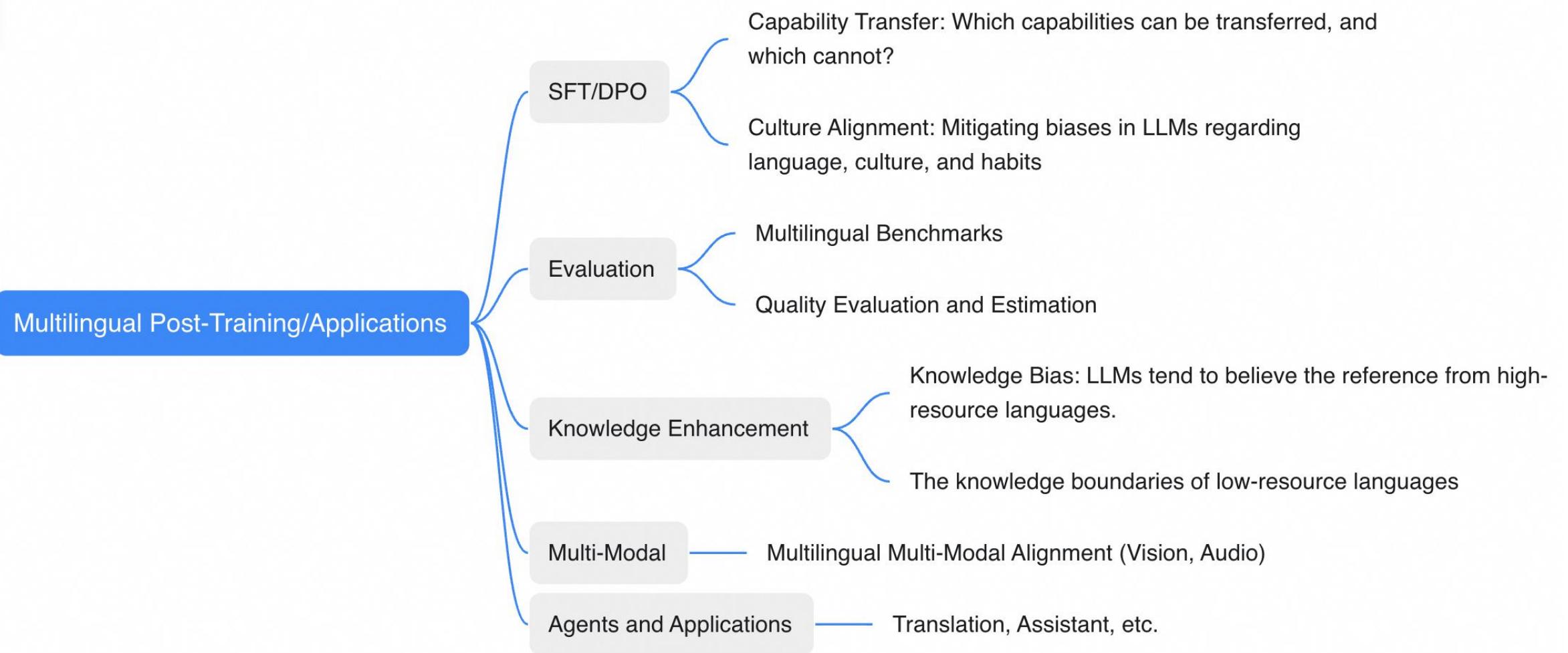
Potential Research Directions in Multilingualism of LLMs

GOSIM



Potential Research Directions in Multilingualism of LLMs

GOSIM

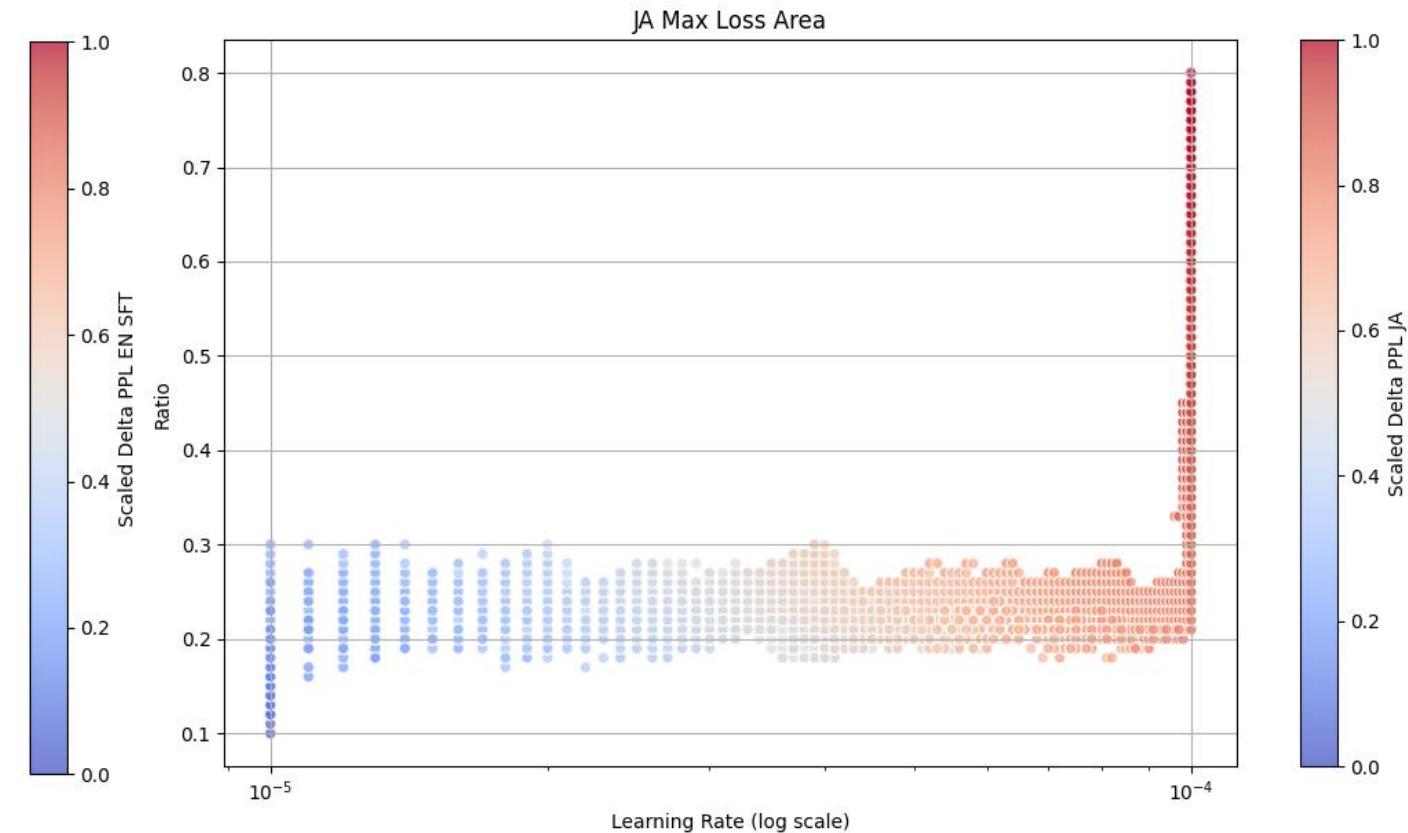
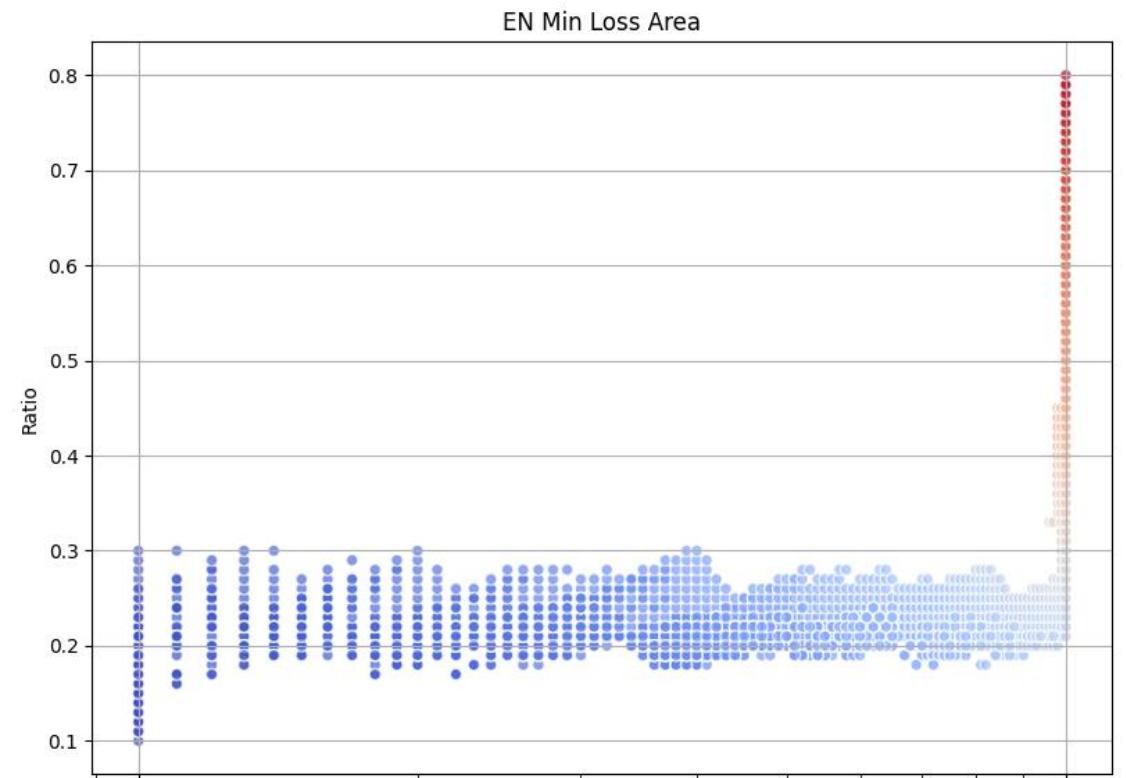


THANK YOU



Predictable Training

GOSIM



How to guide parameter selection?

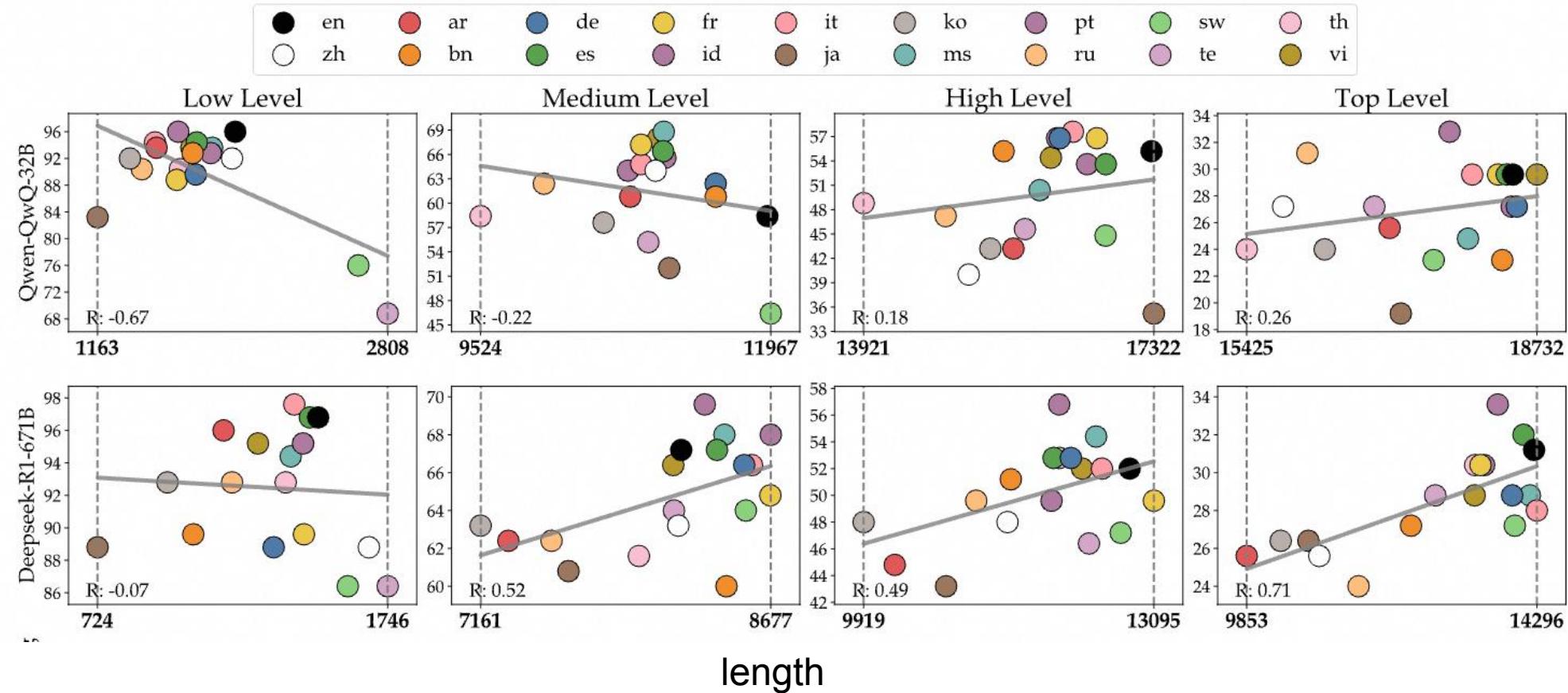
- Goals: keeping the English ppl, reducing Japanese ppl as much as possible.
- A ratio between 20% and 30% appears to be optimal, and a learning rate of 5e-5 or higher is considered a favorable area.



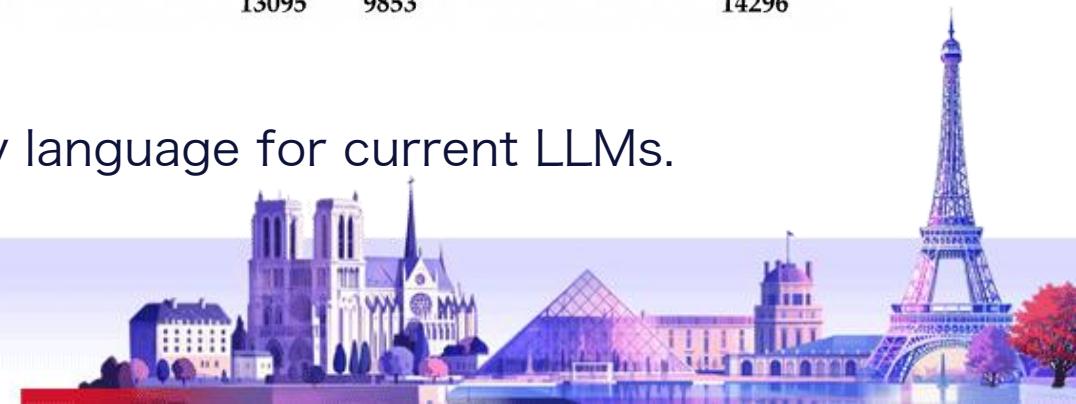
Evaluation

GOSIM

Accuracy



The thinking length differs significantly by language for current LLMs.



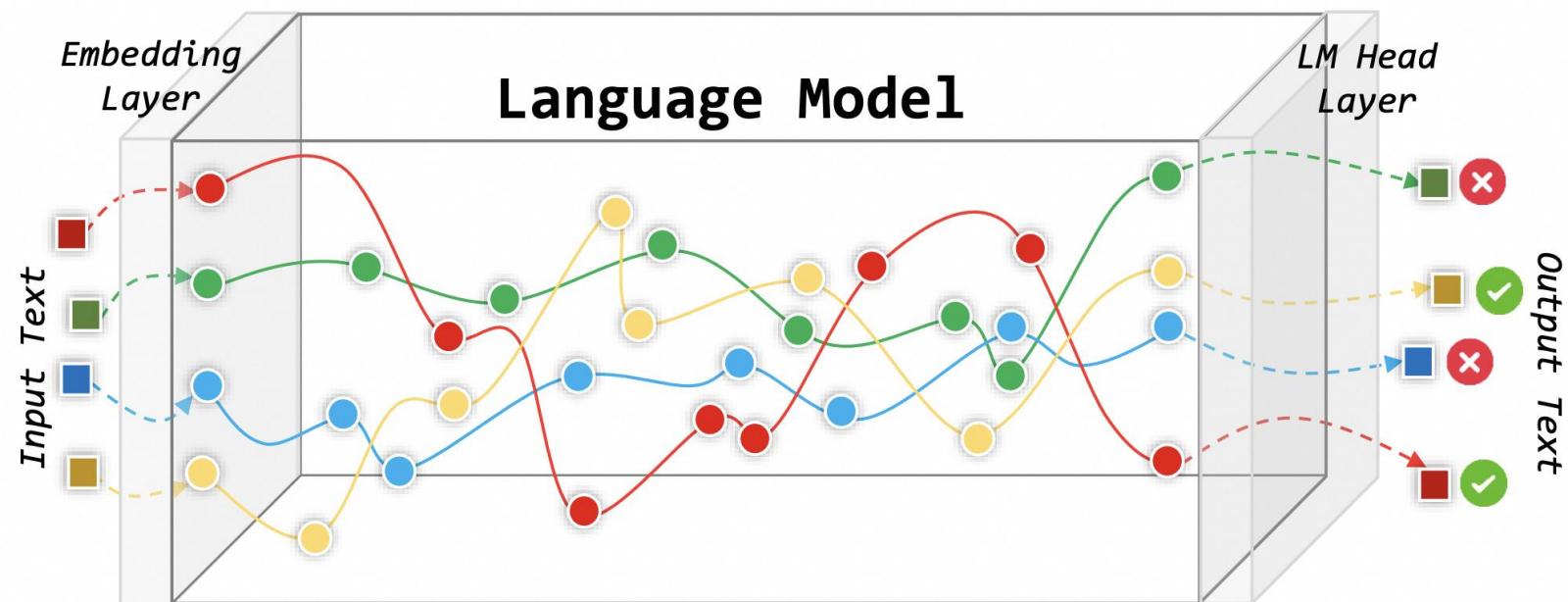
Evaluation

GOSIM

With multiple languages and tasks, it becomes difficult to evaluate the effectiveness → Self-Estimation

Traditional methods: Monte Carlo sampling to calculate uncertainty, probabilities, etc. → overconfidence.

Human: Intuition (prone to errors) and deliberation (often correct) (Evans, 2003; Kahneman, 2010).



Do LLMs exhibit certain pathways
of thought within their hidden layers?

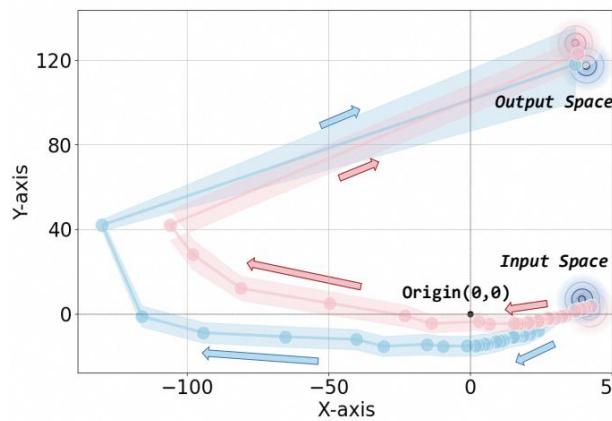
Evaluation

GOSIM

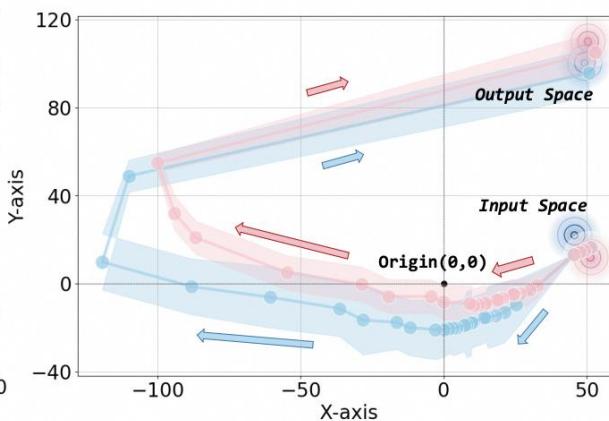
Measure the complexity of pathways based on the differences between layers:

$$M(\mathbf{h}_l, \mathbf{h}_{l+1}) = \|\mathbf{h}_{l+1} - \mathbf{h}_l\|_2, \quad A(\mathbf{h}_l, \mathbf{h}_{l+1}) = \arccos\left(\frac{\mathbf{h}_{l+1}^\top \mathbf{h}_l}{\|\mathbf{h}_{l+1}\|_2 \cdot \|\mathbf{h}_l\|_2}\right)$$

Distance

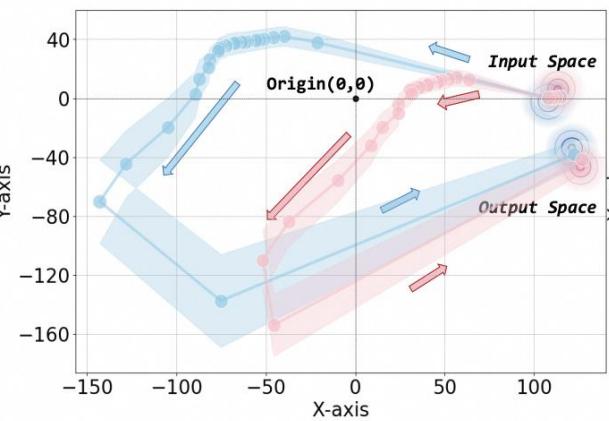


(a) Mathematics

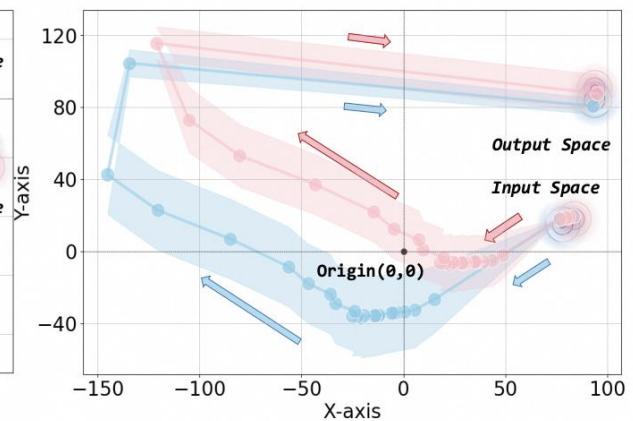


(b) Reasoning

Angle



(c) Knowledge



(d) Understanding

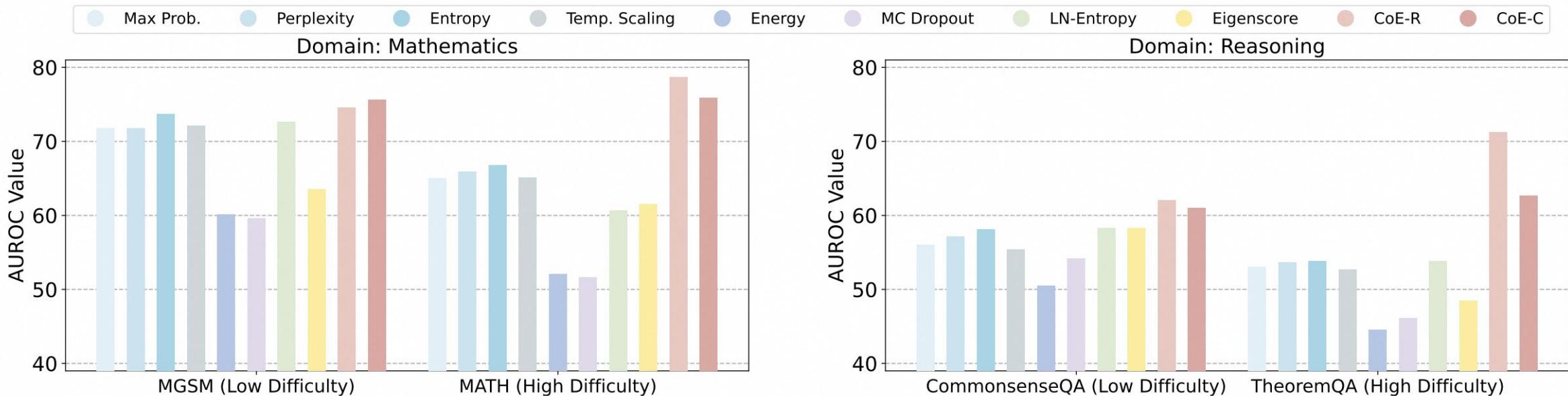
Interesting finding: Correct cases often have more complex pathways (in blue).



Evaluation

GOSIM

Use this property to assess the reliability of model outputs:



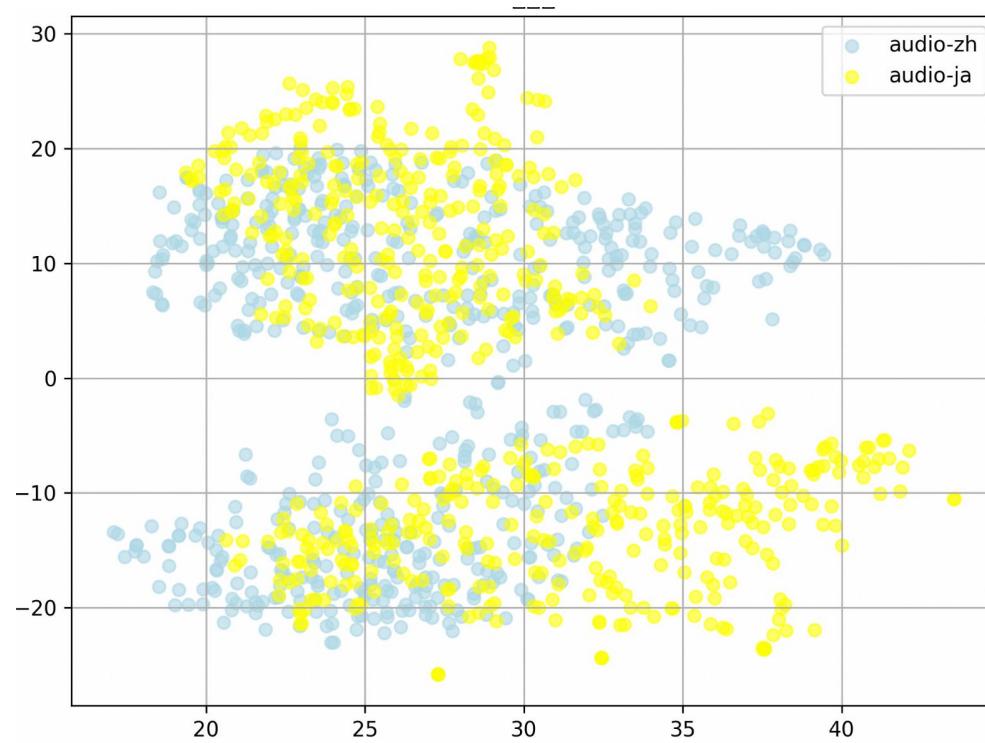
Wang et al., Trajectory Volatility for Out-of-Distribution Detection in Mathematical Reasoning. *NeurIPS 2024*.

Wang et al., Latent Space Chain-of-Embedding Enables Output-free LLM Self-Evaluation. *ICLR 2025*



Existing Solution: Multi-Task Alignment (*Llama3.1*, *CLASI*, etc.)

Training projector and audio encoder with cross-lingual tasks



Besides
Relying on massive cross-lingual
audio-text parallel data



Post-Training – An example on Machine Translation

GOSIM

CoT SFT

[Task 1: domain hint generation task]

Instruction:

Come up with a helpful translation hint to better translate the following German text into English.

Input:

Eine Vollremission (VR) wurde definiert als das Fehlen sichtbarer Leukämiezellen im Knochenmark und die periphere Regeneration von Thrombozyten

Response:

The sentence is from a text related to medicine and pharmacy, translate into a medical domain style.

[Task 2: domain translation task]

Instruction:

Translate the following German text into English according to the hint.

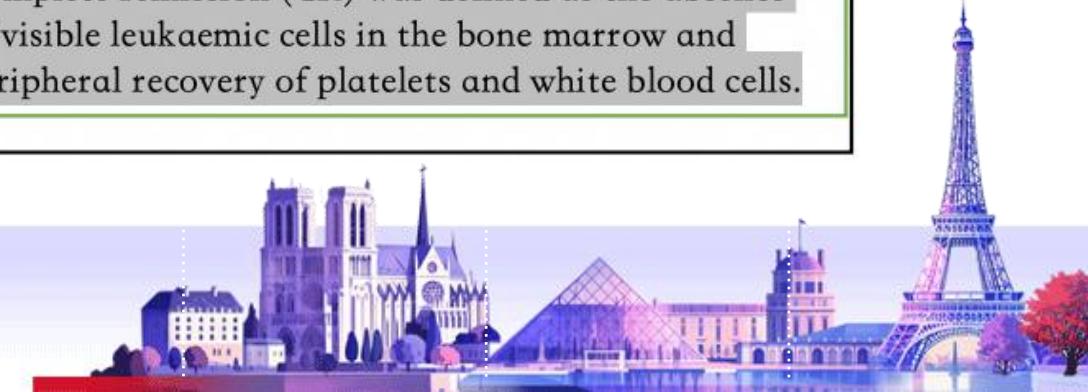
Hint: The sentence is from a text related to medicine and pharmacy, translate into a medical domain style.

Input:

Eine Vollremission (VR) wurde definiert als das Fehlen sichtbarer Leukämiezellen im Knochenmark und die periphere Regeneration von Thrombozyten und Leukozyten.

Response:

Complete remission (CR) was defined as the absence of visible leukaemic cells in the bone marrow and peripheral recovery of platelets and white blood cells.

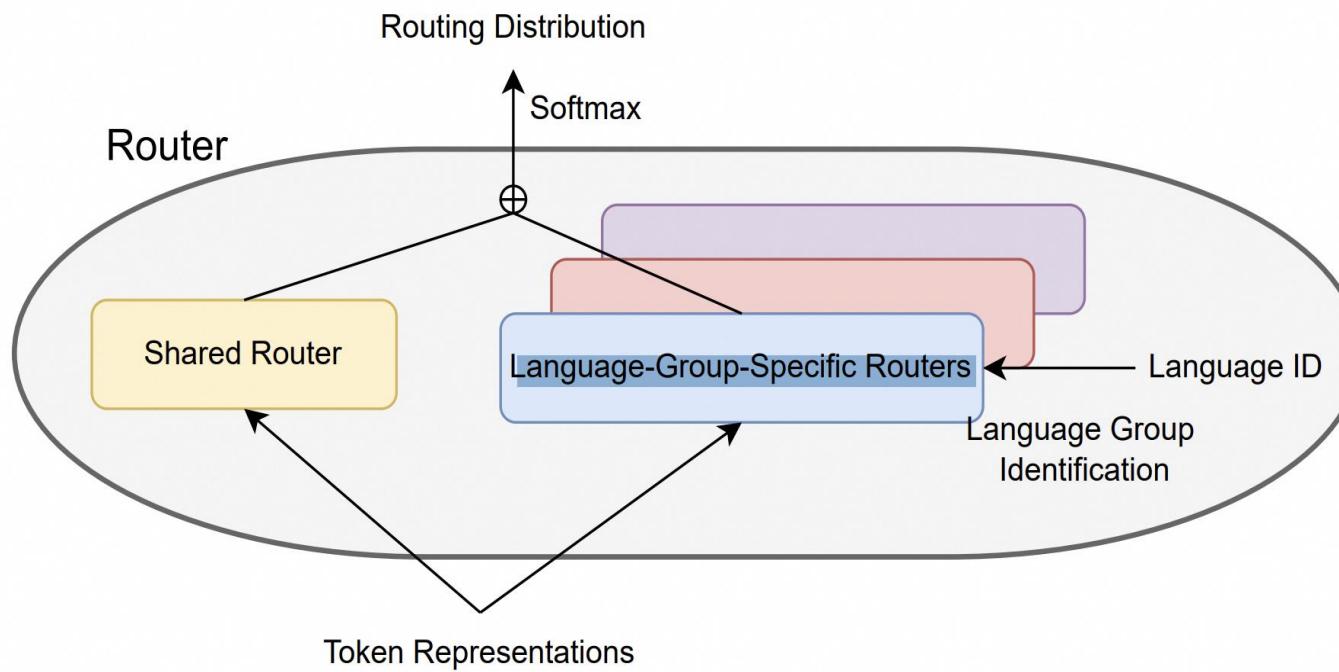


Continue Pretraining

GOSIM

MoE-LGR (Language-Group-Specific Routers)

Step1 : Initialize 22 groups according to Ethologue



$$\mathbf{E}_G = \sum_{i \in G} \frac{T_i}{T_G} \mathbf{E}_i$$

Step2: Merge the two most similar groups until the size of the groups exceeds M.

$$\mathbf{E}_{\text{new}} = \frac{\mathbf{T}_a * \mathbf{E}_a + \mathbf{T}_b * \mathbf{E}_b}{\mathbf{T}_a + \mathbf{T}_b}$$

Finally K Groups

