EuroLLM and FinLLM

Stories from the trenches

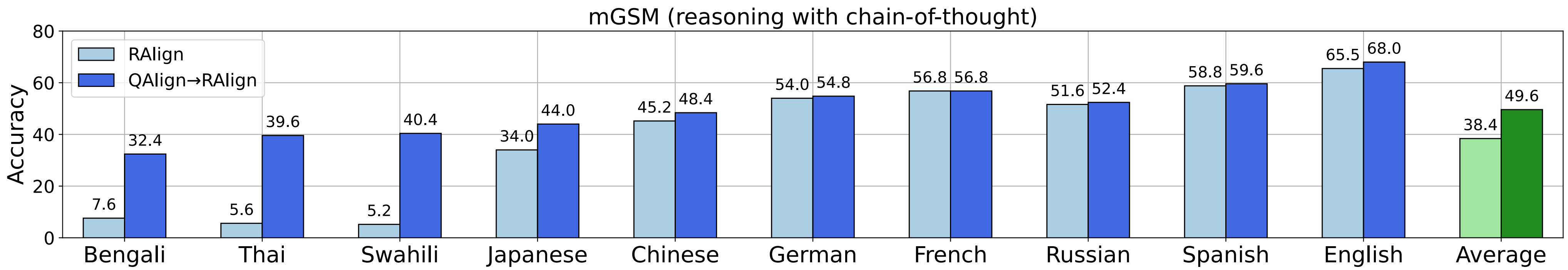
Alexandra Birch

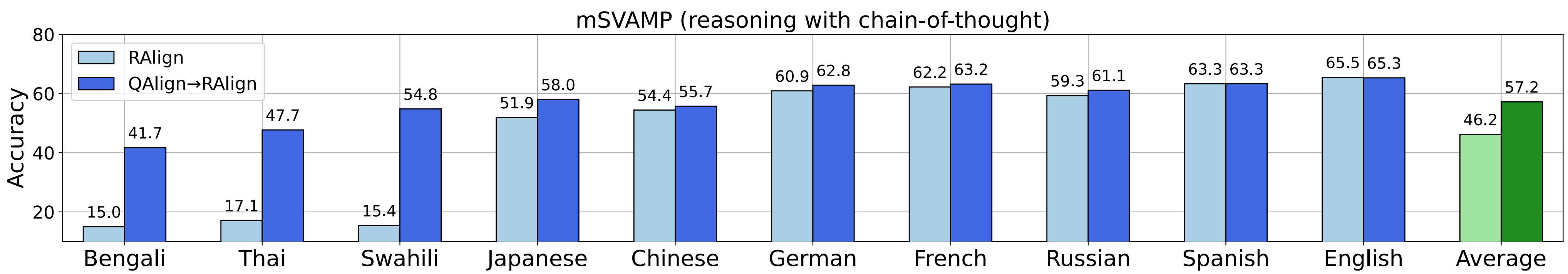
THE UNIVERSITY OF EDINBURGH

Do LLMs serve Europe?

- Top LLMs are primarily trained on English, or English-Chinese
- Commercial models language mix is not often disclosed eg. Google's Gema 2 9B trained on 8 trillion tokens of web data primarily in English, code, maths
- Data mix not given: Aya-Expanse 8B (Dang et al. 2024) covers 23 languages - not focussed on Europe
- Initial efforts in bilingual (CroissantLLM, FinLLM) or on a language family (VikingLLM)
- TowerLLM covered 10 languages and instruction for translation related tasks based on Llama2
- Tueken, Salamander came out in parallel

Multilingual Performance





CodeLLama7B

The Power of Question Translation Training in Multilingual Reasoning: Broadened Scope and Deepened Insights
Wenhao Zhu, Shujian Huang, Fei Yuan, Cheng Chen, Jiajun Chen, Alexandra Birch

EuroLLM Aims

- Multilingual Support all official EU languages plus selected major world languages. Pretrain from scratch with best tokenisation!
- High Performance Competitive with similar sized open-weights models.
- Open Source No usage restrictions, code and data made available.

The Team

Pedro Martins

João Alves

Patrick Fernandes

Nuno Guerreiro

Ricardo Rei

Duarte Alves

Jose Pombal

Amin Farajian

Manuel Fayasse

Mateusz Klimaszewski

Pierre Colombo

Françoise Yvan

Barry Haddow

José de Sousa

André Martins

EuroLLM

- EuroHPC Extreme Call
- Applied for 1.5M Node hours in May 2023
- Approved 420k node hours (4xH100) on October 2023 for Barcelona Super Computer
- Got access to MareNostrum5 on 1st May 2024 for1 year
- Informed 1 August divide quota by 2.2 - got this reversed - now on a low priority queue
- Been selected as one of the best 15 Extreme call projects for JUREAP and 220k node hours on JUPITER May-October 2025

Language Choice

- 24 Official European Languages:

Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, Swedish

- 11 other strategic languages:

Arabic, Catalan, Chinese, Galician, Hindi, Japanese, Korean, Norwegian, Russian, Turkish, and Ukrainian

EuroLLM Plan

- Scaling experiments
- 1.7B base and instruct - 6 August 2024 - 60k downloads

"EuroLLM: Multilingual Language Models for Europe" P. Martins, P. Fernandes, J. Alves, N. Guerreiro, R. Rei, D. Alves, J. Pombal, A. Farajian, M. Faysse, M. Klimaszewski, P. Colombo, B. Haddow, J. Souza, A. Birch, A. Martins https://arxiv.org/abs/2409.16235

- 9B base and instruct - 2 December 2024 - 90k downloads
- 22B going to start next week

Best European Model

Spaces | openGPT-X/european-llm-leaderboard | like 88 | Running on CPU UPGRADE

European LLM Leaderboard

Select languages to average over

- [x] BG
- [x] CZ
- [x] DK
- [x] DE
- [x] EL
- [x] EN
- [x] ES
- [x] ET
- [x] FI
- [x] FR
- [x] HU
- [x] IT
- [x] LT
- [x] LV
- [x] NL
- [x] PL
- [x] PT
- [x] RO
- [x] SK
- [x] SL
- [x] SV

Deselect all languages

Select all languages

Select tasks to show

- [x] ARC
- [x] GSM8K
- [x] HellaSwag
- [x] MMLU
- [x] TruthfulQA

Deselect all tasks

Select all tasks

Type	Model_Name	Average	ARC	GSM8K	HellaSwag	MMLU
	Meta-Llama-3.1-70B-Instruct	0.71	0.71	0.75	0.73	0.77
	Gemma-2-27b-Instruct	0.70	0.75	0.75	0.71	0.68
	Mistral-Nemo-Instruct-12.2B_2407	0.60	0.62	0.57	0.62	0.59
	Mixtral-8x7B-Instruct-v0.1	0.59	0.62	0.48	0.64	0.61
	Gemma-2-9b-Instruct	0.58	0.67	0.45	0.61	0.59
	$\boxed{\text{EuroLLM-9B-Instruct}}$	0.58	0.68	0.45	0.68	0.57
	Mistral-Nemo-Base-12.2B_2407	0.56	0.61	0.44	0.64	0.60
	Meta-Llama-3.1-8B-Instruct	0.56	0.56	0.56	0.58	0.58

Best European Model

Select languages to average over

BG CZ DK DE EL EN ES ET FI
FR HU IT LT LV NL PL PT RO
SK SL SV

Deselect all languages

Select all languages

Select tasks to show

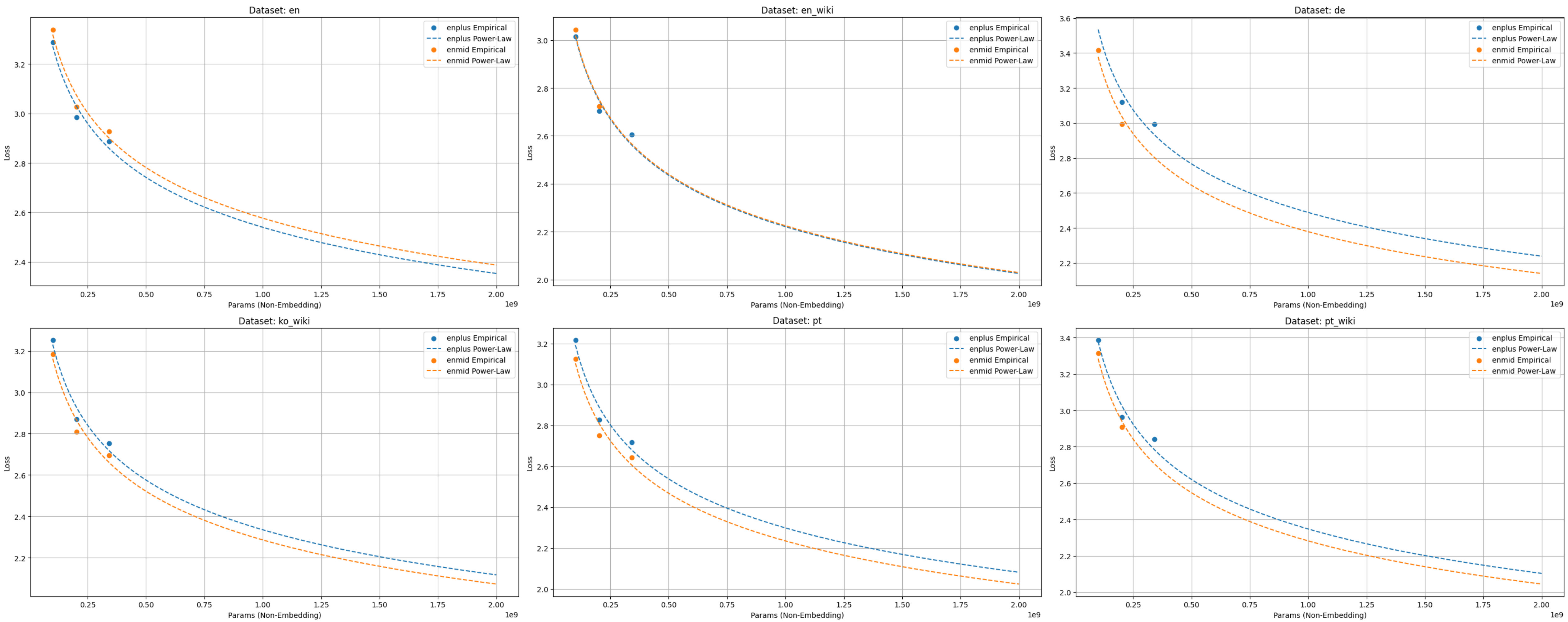
ARC GSM8K HellaSwag MMLU TruthfulQA

Deselect all tasks

Select all tasks

Type	Model_Name	Average	ARC	HellaSwag	MMLU	Truthful
	Meta-Llama-3.1-70B-Instruct	0.70	0.71	0.73	0.77	0.57
	Gemma-2-27b-Instruct	0.69	0.75	0.71	0.68	0.60
	Mixtral-8x7B-Instruct-v0.1	0.62	0.62	0.64	0.61	0.60
	Gemma-2-9b-Instruct	0.61	0.67	0.61	0.59	0.59
	EuroLLM-9B-Instruct	0.61	0.68	0.68	0.57	0.51
	Mistral-Nemo-Instruct-12.2B_2407	0.60	0.62	0.62	0.59	0.58
	EuroLLM-9B-4T	0.60	0.66	0.67	0.56	0.52
	Mistral-Nemo-Base-12.2B_2407	0.59	0.61	0.64	0.60	0.51
	Mixtral-8x7B-v0.1	0.59	0.61	0.64	0.61	0.49
	c4ai-command-r-35B-v01	0.59	0.59	0.65	0.56	0.54
	Teuken-7B-sigma-v05	0.57	0.60	0.67	0.45	0.58

Data Mix

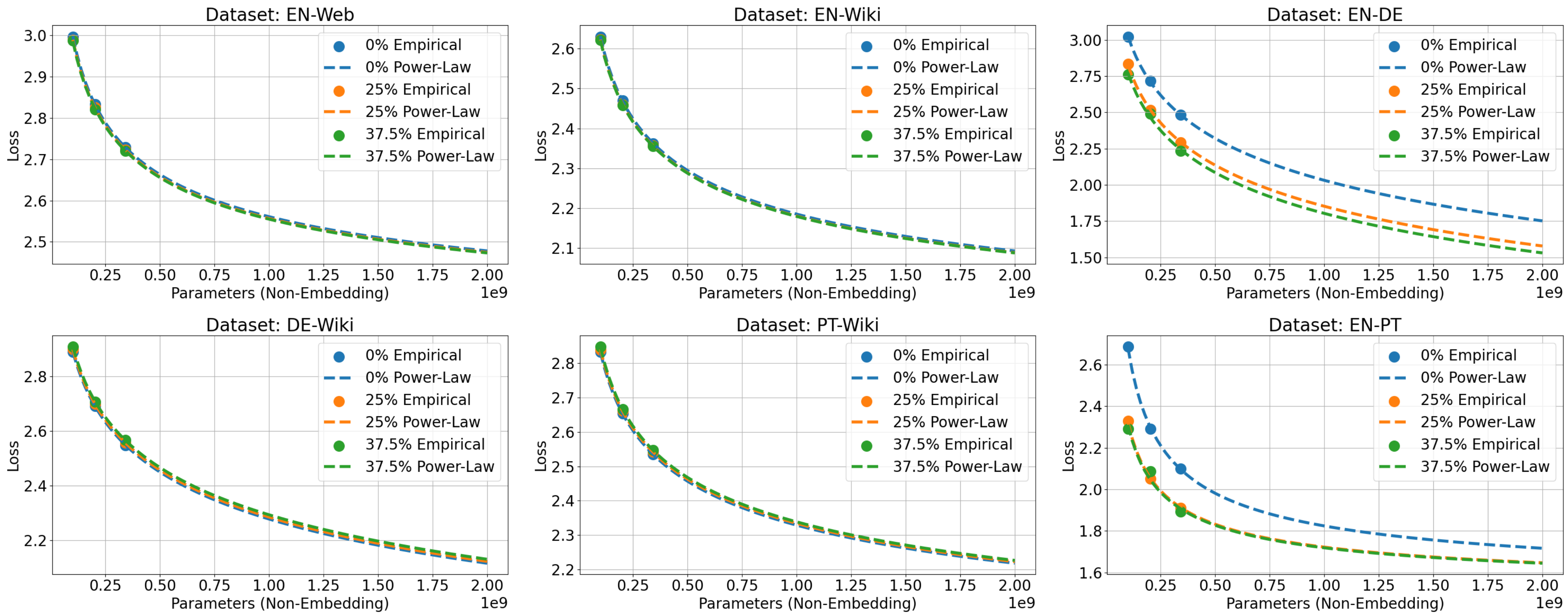


Scaling law: How much English?

enmid - 33%

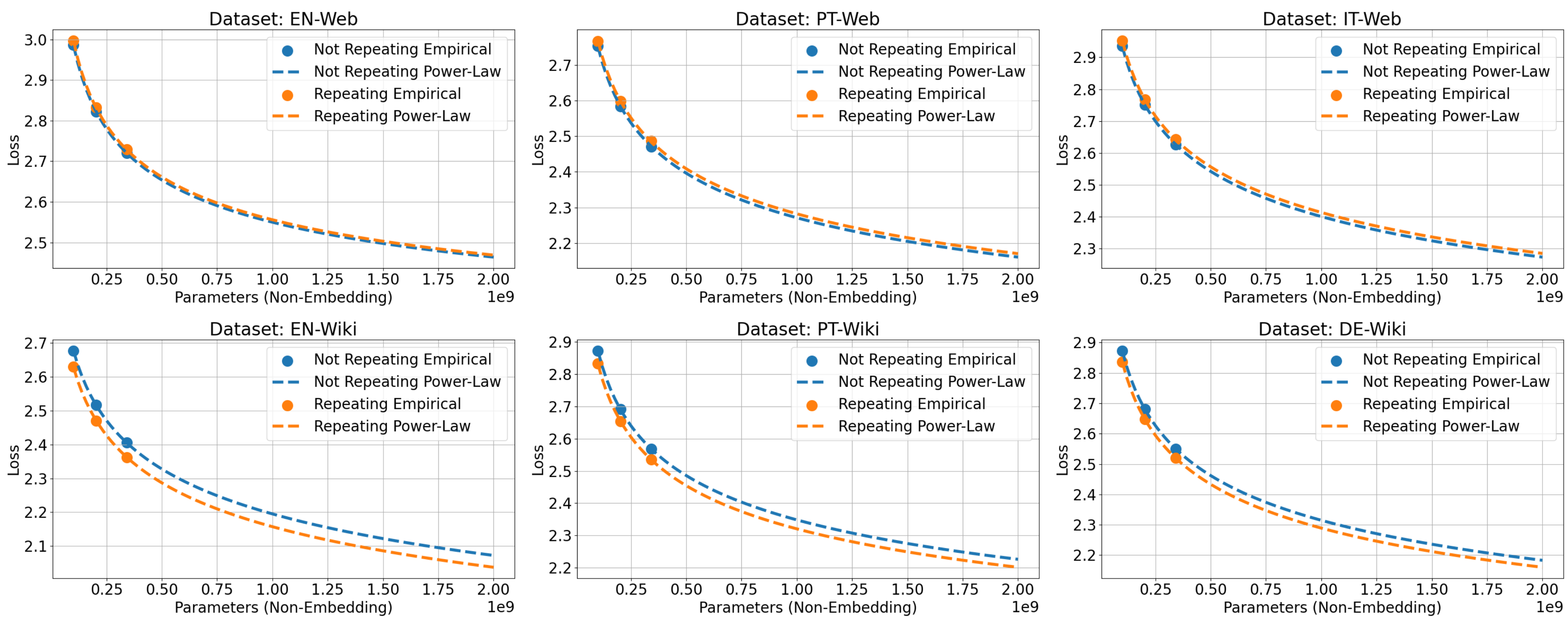
enplus - 50%

Data Mix



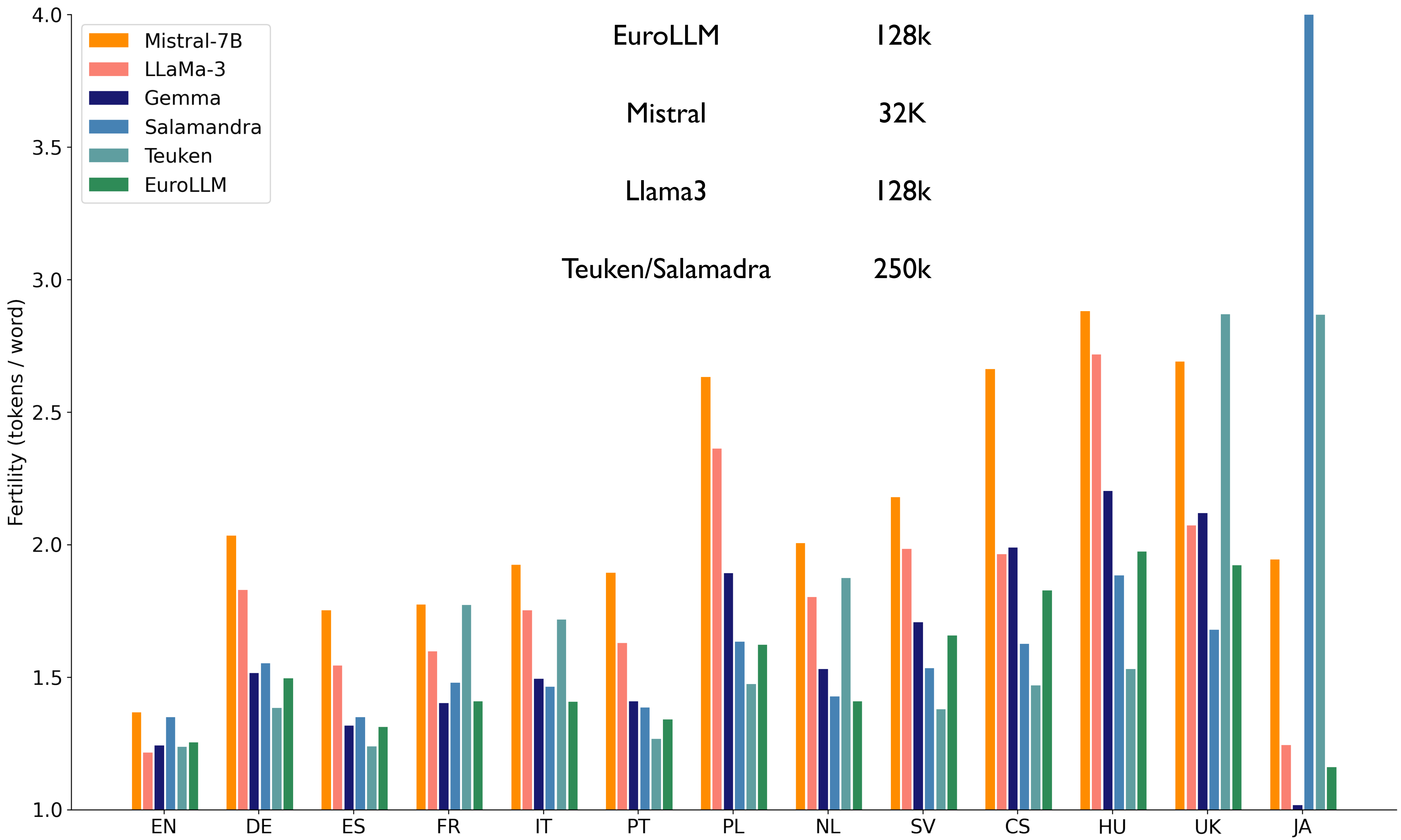
Scaling law: Parallel data experiment from 1.7B

Data Mix

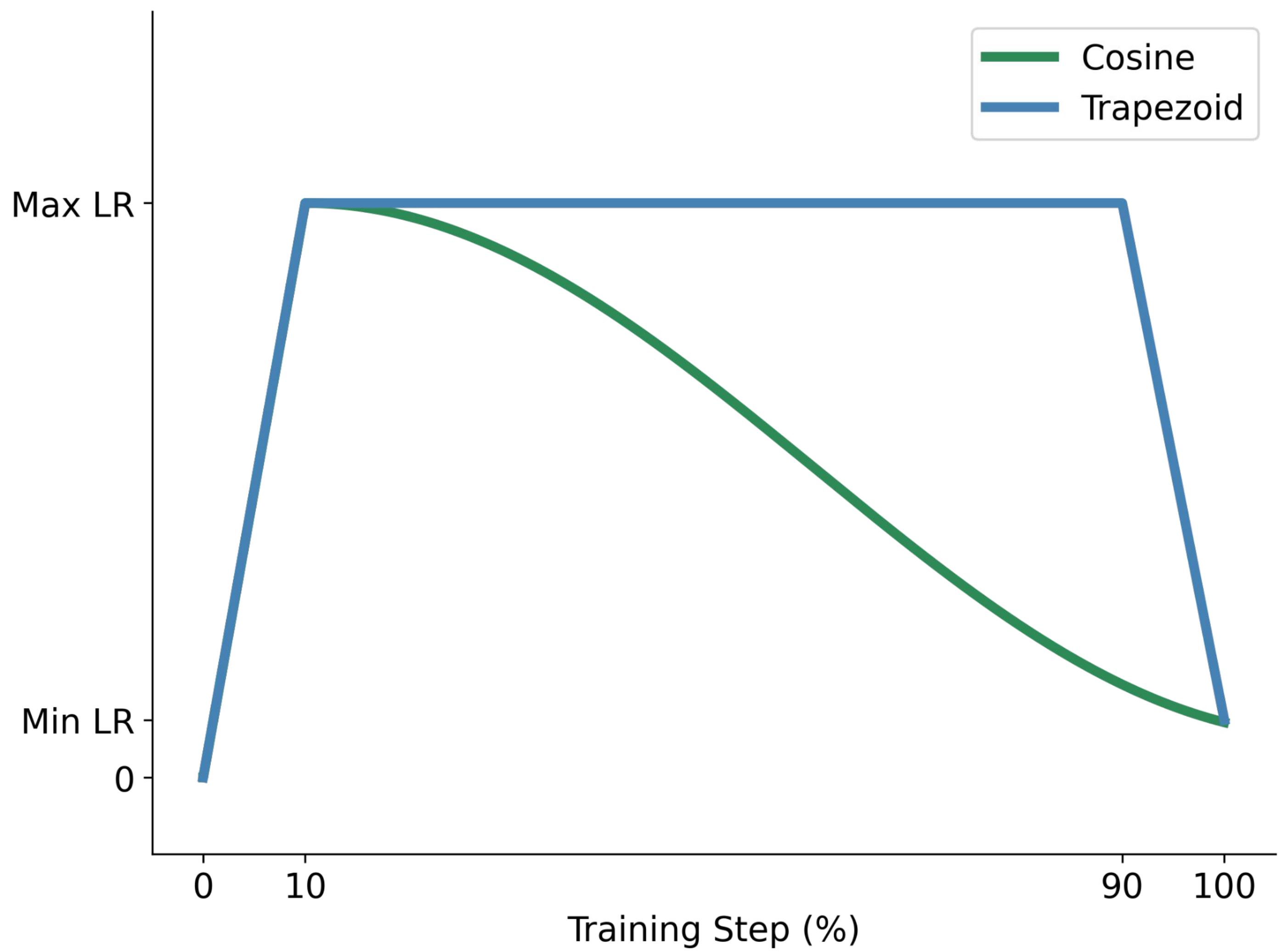


Repeating vs not repeating Wikipedia from 1.9B paper

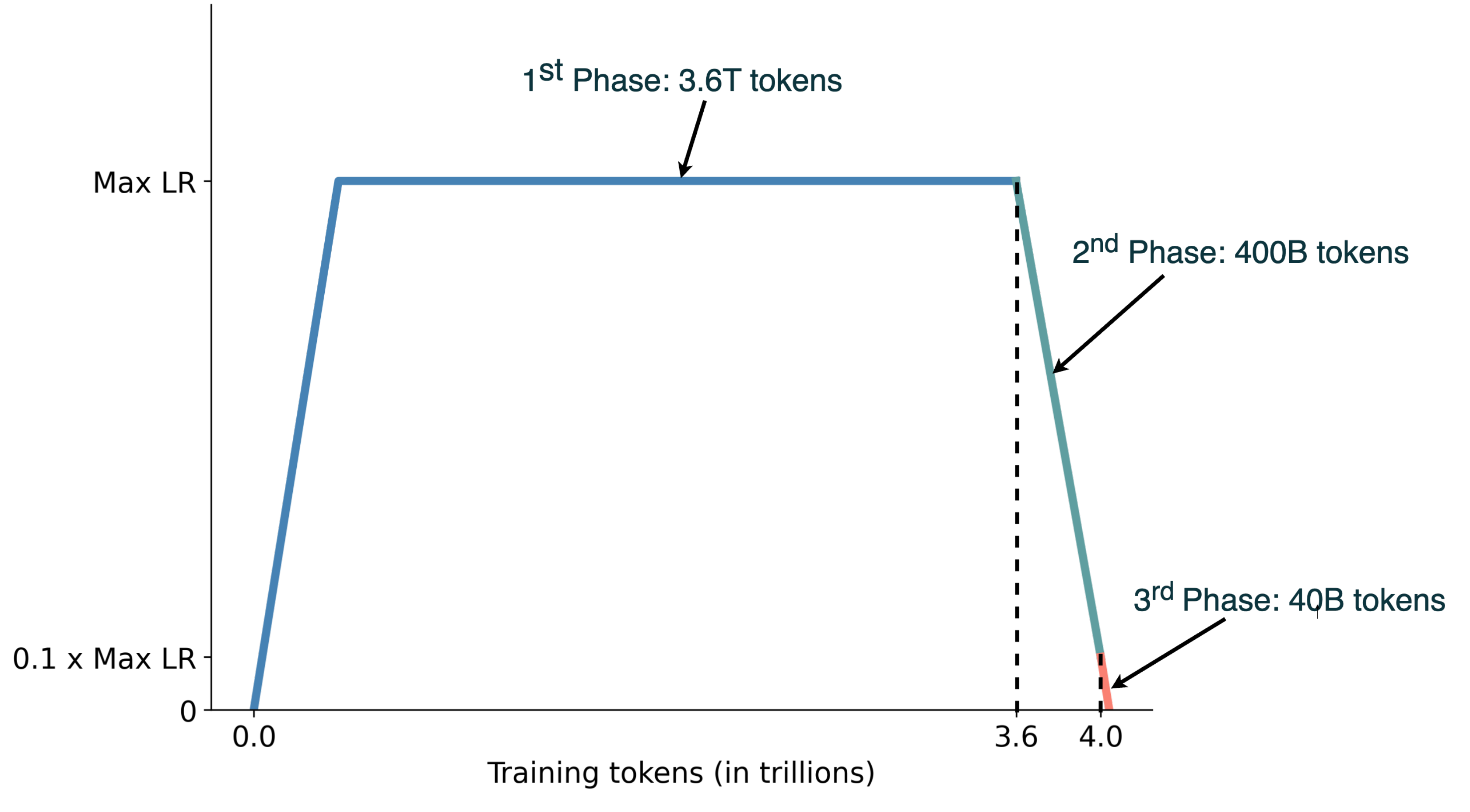
Tokenisation



Pretraining



Pretraining



Data Sources: Web

- Phase 1: English - FinWeb-edu (scores > 2) De,Es,Fr,It - RedPyjama-v2, Remaining languages: HPLT, MADLAD400, Cultural and mC4

Cleaning: deduplicate, heuristic filters (<200char, lorem ipsum, javascript, %symbols), perplexity filtering with KenLM.

- Phase 2 & Phase 3: FinWeb-edu (scores > 3) and filter other languages with model based classifier trained on FinWeb-edu like labelled data

Data Sources: Parallel

- Phase 1: Large collection of corpora: Europarl, Paracrawl, CCMatrix, Opus etc.

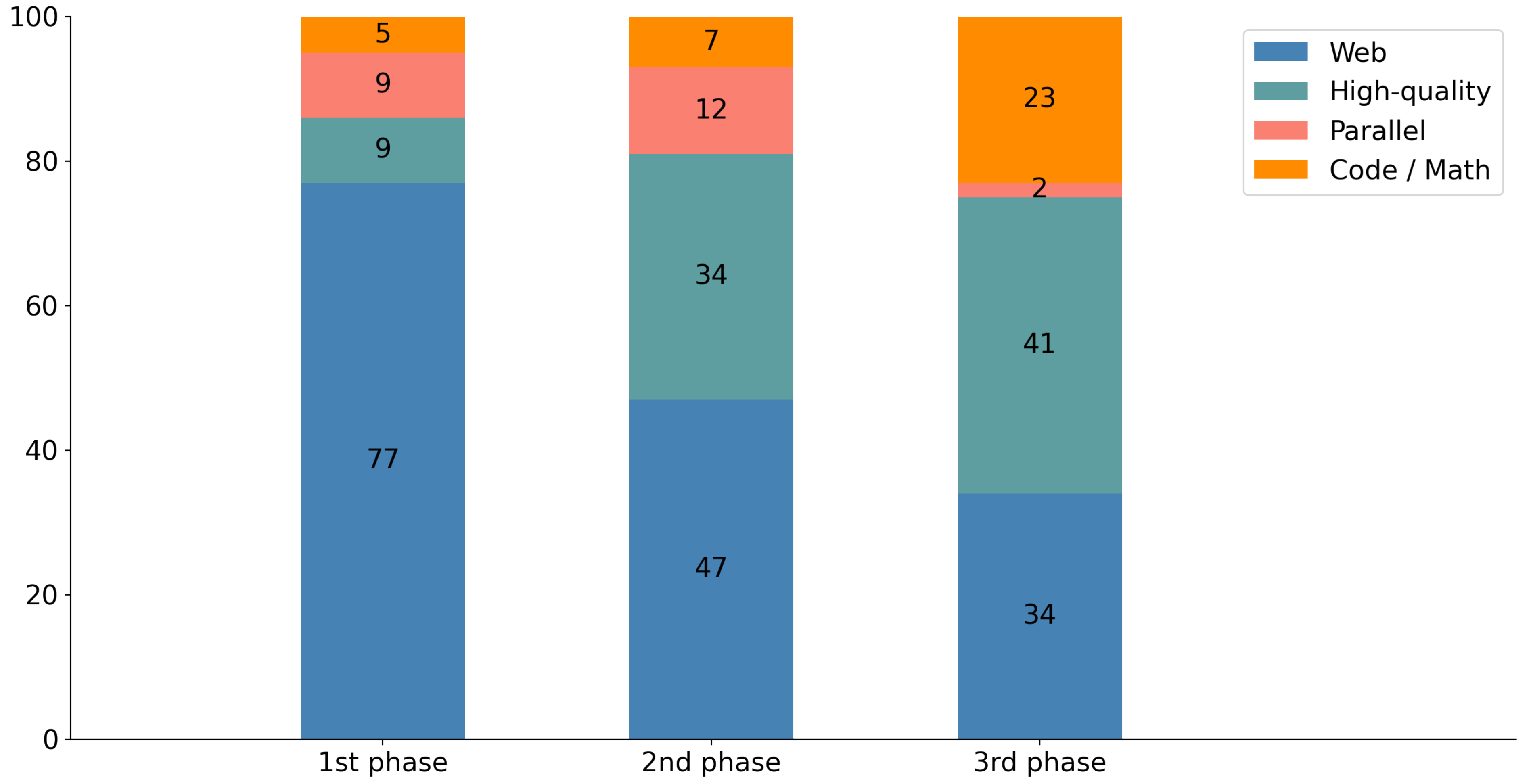
Cleaning: Bifixer to remove duplicates, Bicleaner and CometKIWI-22 to remove low quality.

- Phase 2 & Phase 3: Added document level parallel corpora from Europarl

Data Sources

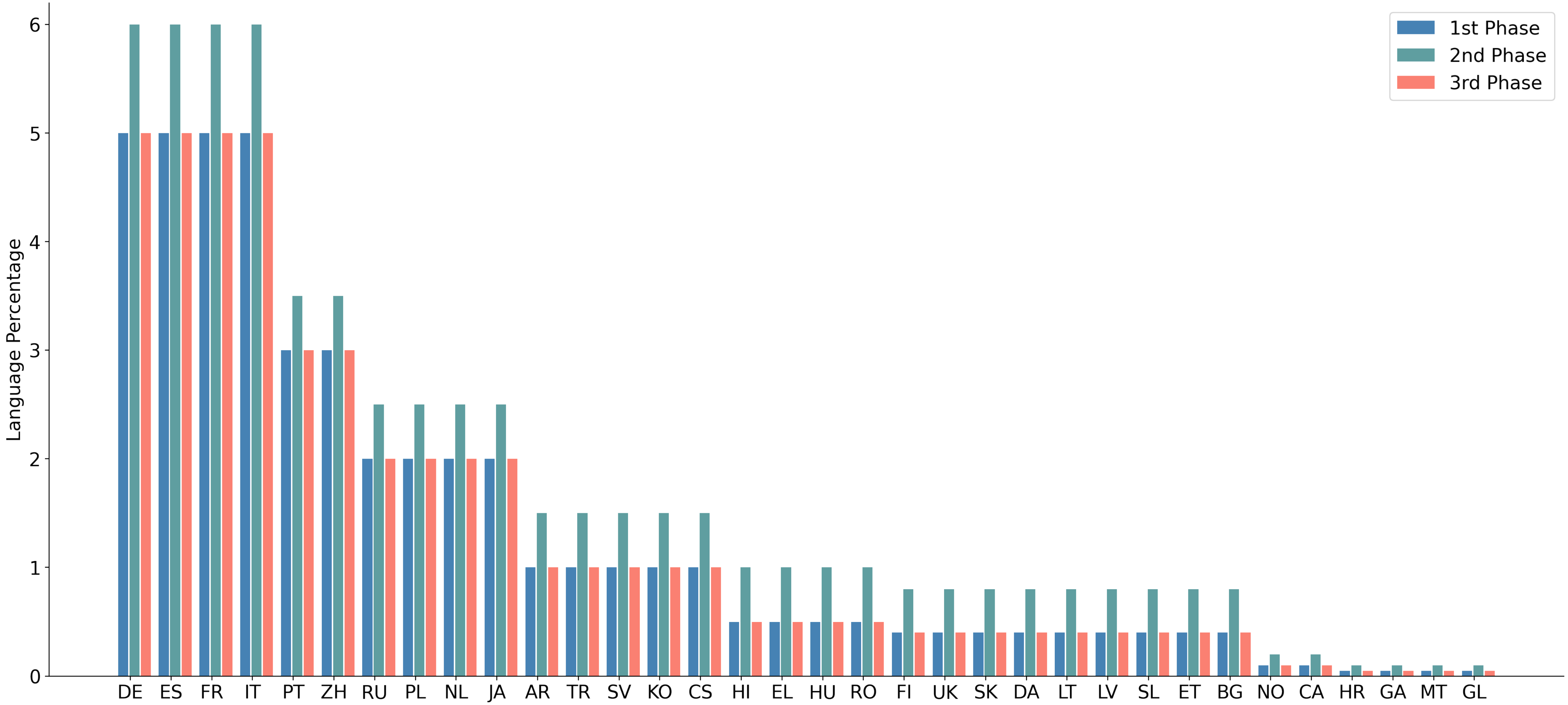
- Code and Math Data: The Stack and Open Web Math. Phase 2&3 added GSM8K training and synthetic Qwen Maths.
- High Quality Data: Wikipedia, Arrive, Books, Apollo and Synthetic Cosmopedia in English. For Phase 3 added Cosmopedia translated using Tower (Alves et al., 2024) to German, Spanish, French, Italian, Portuguese, Dutch, Chinese, and Russian.

Data Mix



% data categories

Data Mixes



1st Phase
2nd Phase
3rd Phase
Language Percentage
0
1
2
3
4
5
6
DE ES FR IT PT ZH RU PL NL JA AR TR SV KO CS HI EL HU RO FI UK SK DA LT LV SL ET BG NO CA HR GA MT GL

Post Training

EuroBlocks

- Consists of TowerBlocks, Aya, OpenHermes, OpenMath2 and others
- Filtered for complexity (remove low complexity) and readability
- Score response using ArmoRM (honesty, verbosity, safety) and remove conversations where the scores fall below a threshold
- Less-represented languages:
    - Synthetic instructions - super-annealing set as seed data and either Llama 3.1 70B or earlier checkpoints of EuroLLM
    - Translation data, and incorporating high-quality translation examples

Baselines

Pre-trained	Post-trained	Technical Report	European	EU Lang. Supp.
Gemma-2-9B	Gemma-2-9B-IT	Gemma Team et al. (2024)	No	—–
LLaMa-3.1-8B	LLaMa-3.1-8B-IT	Llama Team et al. (2024)	No	—–
Granite-3-8B	Granite-3-8B-IT	Granite Team (2024)	No	No
Qwen-2.5-7B	Qwen-2.5-7B-IT	Qwen Team et al. (2024)	No	No
OLMo-2-7B	OLMo-2-7B-IT	OLMo et al. (2024)	No	No
Aya-23-8B	Aya-Expanse-8B	Singh et al. (2024); Dang et al. (2024)	No	No
Mistral-7B	Mistral-7B-IT	Jiang et al. (2023)	Yes	No
Not available	Ministral-8B-IT	—-	Yes	No
Occiglot-7B-eu5	Occiglot-7B-eu5-IT	—-	Yes	No
Salamandra-7B	Salamandra-7B-IT	—-	Yes	Yes
Not available	Pharia-1-LLM-7B-C	—-	Yes	No
Not available	Teuken-7B-IT-R-v0.4	Ali et al. (2024)	Yes	Yes
Not available	Teuken-7B-IT-C-v0.4	Ali et al. (2024)	Yes	No

Evaluation

- Arc-Challenge (Clark et al., 2018): challenging MCQ science exams from grade 3 to grade 9.
- Hellaswag (Zellers et al., 2019): multiple-choice commonsense inference
- MMLU (Hendrycks et al., 2021a) and MMLUPro: MCQ humanities, social sciences, hard sciences
- MUSR (Sprague et al.): MCQ complex problems with around 1,000 words in length generated algorithmically eg. murder mysteries - reason with long-range context.
- GSM8k (Cobbe et al., 2021): multiple-choice grade school math
- IFEval (Kovalevskyi, 2024): set of prompts that test a model's ability to follow explicit instructions

Evaluation

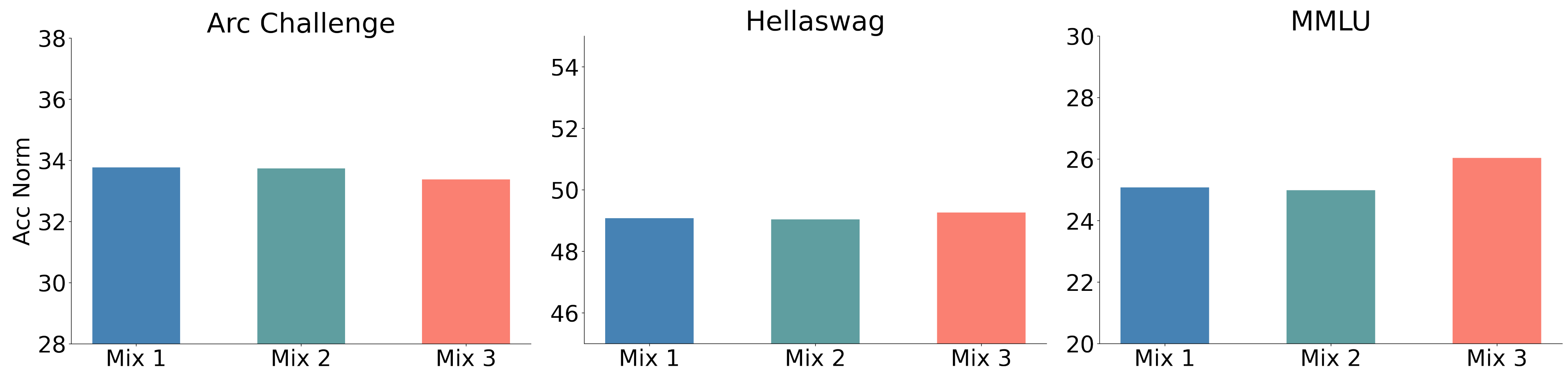
- Translations of Arc-Challenge, Hellaswag, and MMLU from Okapi Lai et al. (2023) in 11 languages, and translate MMLU-PRO and MUSR using Tower into 6 languages
- Translation results 3 WMT translation tasks, and 46 Flores translations tasks - all evaluated with COMET-22
- Reporting a fairer average for ranking: using normalized scores https://huggingface.co/spaces/open-llm-leaderboard/blog - between the random baseline (0 points) and the maximal possible score (100 points)
- Borda count - average rank - not overly influenced by one test set

Phase 2 Data Mix

Mix 1: English 48%, code/math data increased to 7%.

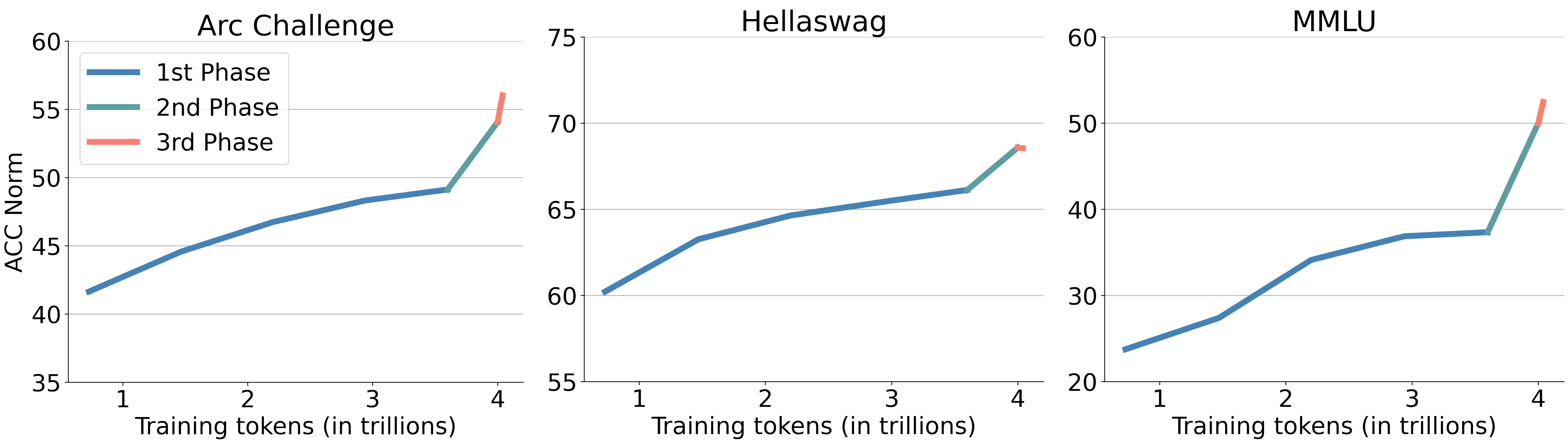
Mix 2: English 40%, code/math data increased to 15%.

Mix 3: English 32.5%, code/math data at 7%, redistributing the remaining percentage across the other languages.



Mix 3 overall best - similar experiment Phase 3 increased code/math

Pretraining progress



Across 11 languages

Evaluation

Averaged EU languages: 11 (Arc,Hella,MMLU) and 6 Non-English languages

Pre-trained	**Arc-C** (25-shot)	**Hellaswag** (10-shot)	**MMLU** (5-shot)	**MMLU-pro** (5-shot)	**MUSR** (0-shot)	**Borda C ↓**
Non-European						
Gemma-2-9B	**59.79**	**70.83**	64.93	29.75	**9.70**	**1.4**
LLaMa-3.1-8B	48.54	65.10	56.01	19.64	5.44	3.2
Granite-3-8B	46.47	61.77	52.35	20.38	9.36	3.2
Qwen-2.5-7B	48.98	60.37	**65.34**	**31.63**	8.04	2.4
OLMo-2-7B	37.35	49.65	45.77	13.91	4.53	5.8
Aya-23-8B	44.15	61.15	47.89	14.04	3.64	5.0
European						
Mistral-7B	48.65	62.10	51.68	17.36	8.69	2.4
Occiglot-7B-eu5	44.99	61.22	45.28	11.98	3.83	3.4
Salamandra-7B	48.89	63.60	40.23	5.25	2.63	3.2
EuroLLM-9B	**56.03**	**68.54**	**52.45**	**17.60**	**10.97**	**1.0**

Evaluation

Averaged EU languages: same as before, adding 3 and 46 translation directions

Post-trained	**Arc-C** (25-shot)	**Hellaswag** (10-shot)	**MMLU** (5-shot)	**MMLU-pro** (5-shot)	**MUSR** (0-shot)	**WMT-24** (0-shot)	**FLORES** (0-shot)	**Borda C ↓**
Non-European								
Gemma-2-9B-IT	**57.98**	**66.95**	**63.07**	27.42	**8.38**	79.82	**86.82**	**1.3**
LLaMa-3.1-8B-IT	52.75	62.40	57.53	24.22	4.01	78.94	84.85	3.0
Granite-3-8B-IT	42.44	55.85	50.15	20.10	7.90	72.18	72.25	4.4
Qwen-2.5-7B-IT	47.09	57.73	62.86	**29.68**	7.62	75.96	76.97	3.1
OLMo-2-7B-IT	40.81	52.02	45.65	12.38	4.02	69.24	71.47	5.9
Aya-Expanse-8B	47.40	61.84	53.58	19.77	5.52	**83.01**	77.73	3.3
European								
Mistral-7B-IT	50.39	61.46	50.75	**18.19**	6.94	75.11	77.98	4.0
Ministral-8B-IT	48.67	61.62	51.55	17.41	6.17	77.13	81.34	3.9
Occiglot-7B-eu5-IT	42.13	59.49	42.08	11.77	4.17	75.10	74.40	6.1
Salamandra-7B-IT	44.69	63.60	44.60	7.01	7.17	80.87	87.35	3.9
Pharia-1-LLM-7B-C	40.55	55.22	39.91	10.10	**9.83**	63.80	58.91	6.4
Teuken-7B-IT-R-v0.4	46.84	62.75	39.81	9.29	2.25	77.91	82.63	5.3
Teuken-7B-IT-C-v0.4	46.28	62.73	41.74	9.79	2.94	77.68	84.41	5.0
EuroLLM-9B-IT	**56.55**	**67.53**	**52.97**	17.04	9.02	**83.61**	**88.87**	**1.4**

Evaluation

English Results

Pre-trained	**Arc-C** (25-shot)	**Hellaswag** (10-shot)	**MMLU** (5-shot)	**MMLU-pro** (5-shot)	**MUSR** (0-shot)	**GSM8k** (5-shot)	**Borda C** ↓
Non-European							
Gemma-2-9B	**68.34**	82.73	70.75	34.87	**14.48**	67.78	**1.8**
LLaMa-3.1-8B	57.68	81.90	65.25	25.17	9.13	49.43	4.5
Granite-3-8B	63.65	**83.29**	64.41	25.82	9.36	64.22	3.5
Qwen-2.5-7B	63.91	80.18	**74.23**	**37.33**	13.06	**83.24**	2.6
OLMo-2-7B	64.68	81.93	68.85	22.74	10.12	68.31	3.0
Aya-23-8B	52.99	78.05	55.18	16.68	5.85	41.85	6.0
European							
Mistral-7B	**60.58**	**83.14**	**62.35**	**21.78**	8.50	37.38	**1.3**
Occiglot-7B-eu5	52.90	78.95	52.78	13.87	2.68	25.70	3.0
Salamandra-7B	55.63	77.17	39.76	5.52	2.58	2.43	3.8
EuroLLM-9B	59.73	78.83	57.32	17.68	**12.47**	**47.69**	1.8

Plans for the future

- Safety and bias evaluation and mitigation
- Speech and images - multimodal model
- Reasoning and scaling inference

Do you speak Finance?

FinLLM

Model Development
- Training of series of FinLLM models: 1.5B, 7B, 30B
- AveniBench: benchmark of financial test sets
- AveniPile: dataset of high quality financial data

Prove Value
- Achieve top performance on key financial NLP tasks
- Deliver GenAI applications in partnership with Lloyds and NW

Regulatory Compliance
- Survey existing regulations and best practices
- Industry leading approach to ethical and safe GenAI with FCA

Integration and Deployment
- APIs for model access
- Successfully integrate with partner environment

AveniPile: Web

Level 1 Category	Level 2 Category
Finance Literacy	Academic and theoretical contents
	Common financial language
	Curriculum in professional qualifications
	Professional associations
Regulatory and accounting	EU Regulations
	UK Regulations
	Taxation and Accounting
Financial news and market data	Financial News and Media
	Financial Markets
Investment insights and analysis	Investment Research
	Sector Analysis
	Market Behaviour and Sentiment
Company information	Financial performance and analysis
	Press releases
Finance products and services	Retail banking
	Corporate banking
	Investment banking
	Private banking

Taxonomy

AveniPile: Web

Level 1	Level 2	Website
financial press	financial news, editorial contents and expert opinions	CNBC
		Forbes
		Yahoo Finance
		Dow Jones
consumer	comparison sites and product reviews	Finder
	Personal finance, goal setting, budgeting, expense tracking, and bill management	Mint
		Personal Capital
market	real-time market data including stock prices, trading volumes, and economic indicators.	Bloomberg Terminal
		Google Finance
		Investing.com - Stock Market Quotes & Financial News
		Alpha Vantage
		IEX Cloud
	developer friendly APIs for market data	Interactive Brokers
		Investors Exchange (IEX)
educational	knowledge sharing sites	wikipedia
		Seeking ALpha

Seed URLs for crawling

AveniPile: Legal

Credit score information
Level 0

Financial planning tools
Level 0

EU regulators
Level 0

UK regulators
Level 0

Taxation and accounting
Level 0

Common financial language
Level 0

Curriculum and professional qualifications
Level 0

Academic and theoretical contents
Level 0

Market data
Level 0

Operational Risk
Level 1

Financial planning
Level 1

Investment management
Level 1

Insurance
Level 1

Corporate banking
Level 1

Investment banking
Level 1

Retail banking
Level 1

Financial performance and analysis
Level 1

Sector analysis
Level 1

Investment research
Level 1

Professional associations
Level 1

Financial news and media
Level 1

Personal finance blogs
Level 2

Press releases
Level 2

Market behaviour and sentiment
Level 2

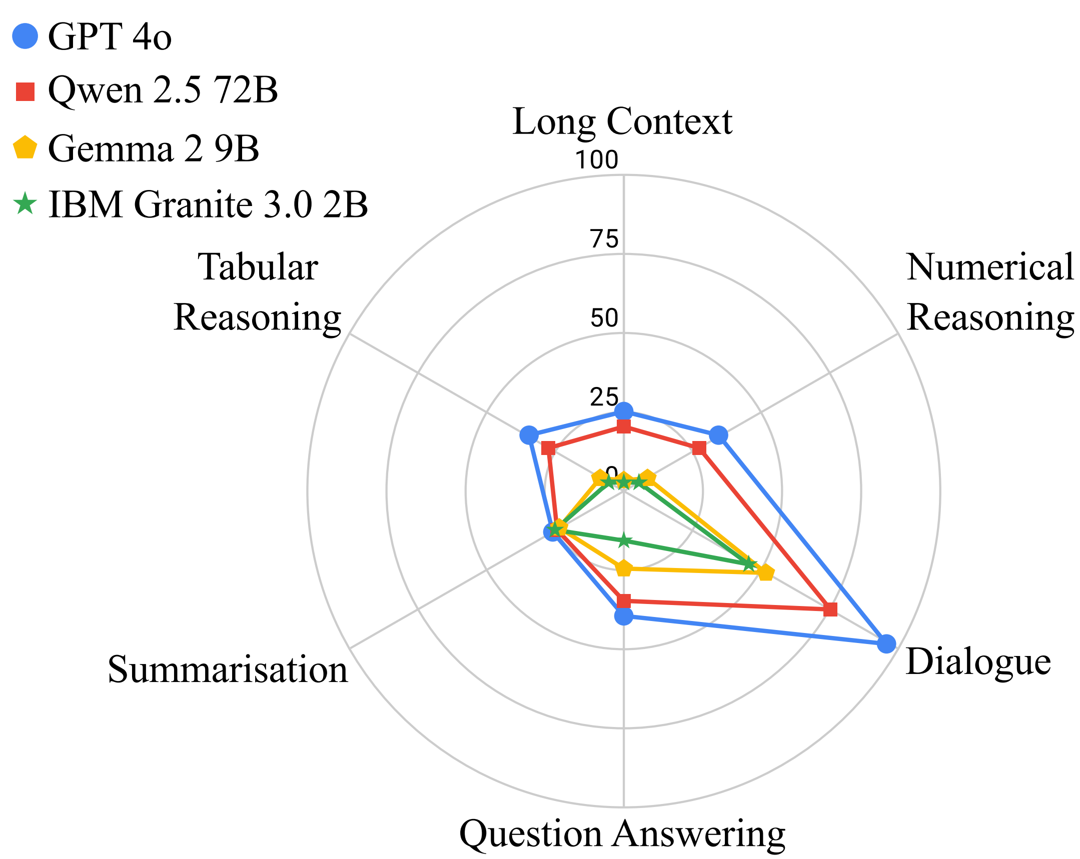
level 0 = no anonymisation
level 1 = iban code/credit card/sort code/account number/ni number/passport number/nhs number
level 2 = level 1 + person (name)/location/nationality, religion, political affiliation/ethnicity/title

AveniPile: Web

Filtering

- Taking HPLT and finWeb and filtering for finance
- Use classifiers trained on LLM labelled data
- 18B Tokens

AveniBench



"AveniBench: Accessible and Versatile Evaluation of Finance Intelligence"
Mateusz Klimaszewski, Pinzhen Chen, Liane Guillou,
Ioannis Papaioannou, Barry Haddow, Alexandra Birch, 2025

https://huggingface.co/spaces/aveni-ai/aveni-bench

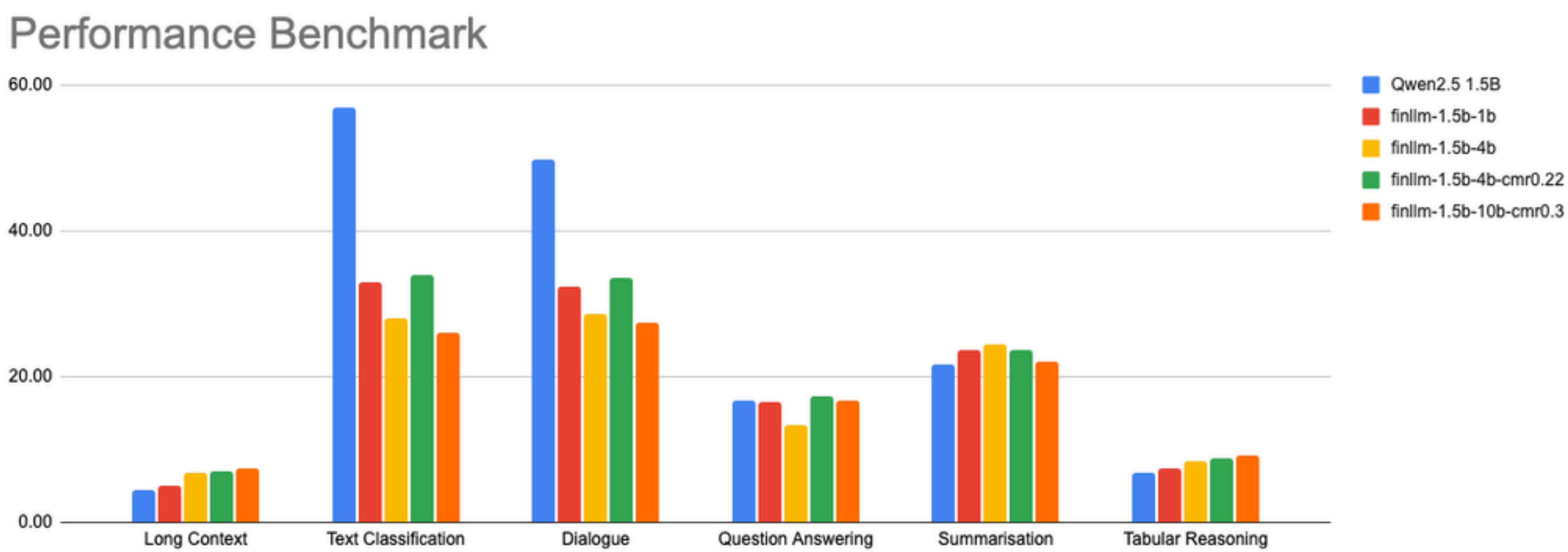
AveniBench

Full Evaluation Set List	NLP Capability	NLP Capability	NLP Capability
Banking 77	Text Classification	Dialogue	
NLU++ EASY	Text Classification	Dialogue	
NLU++ HARD	Text Classification	Dialogue	
FinQA	Question Answer…	Information R…	Tabular Data
ConvFinQA	Question Answer…	Dialogue	
ECTSum	Text Summarisat…	Text Generation	
MultiHiertt EASY	Tabular Data	Long Context …	
MultiHiertt HARD	Tabular Data	Long Context …	
TATQA	Question Answer…	Tabular Data	
TATHQA	Question Answer…	Tabular Data	
Financial Planning Single	Question Answer…	Tabular Data	
Financial Planning Multi	Question Answer…	Tabular Data	

AveniBench

Model	Param.	Banking77 (0-shot)	NLU++ EASY (0-shot)	NLU++ HARD (0-shot)	FinQA (0-shot)	ConvFinQA (0-shot)	ECTSum (0-shot)	MultiHiertt EASY (2-shot)	MultiHiertt HARD (0-shot)	TAT-QA (4-shot)	TAT-HQA (4-shot)	AVG	**Borda Count** Score	**Borda Count** Rank
Proprietary LLMs														
GPT-4o	-	96.43	97.59	94.18	16.98	61.43	25.75	27.33	22.84	41.37	48.06	53.20	189	1
GPT-4o-mini	-	94.94	97.04	91.04	10.57	55.83	22.41	15.33	9.43	31.45	22.57	45.06	153	4
Open-weight LLMs														
Qwen 2.5	72B	95.27	97.85	33.02	13.58	55.43	24.61	24.00	16.48	39.63	30.95	43.08	177	2
Qwen 2.5	32B	94.81	96.51	22.04	10.94	55.36	25.16	23.33	13.21	33.19	23.67	39.82	164	3
Llama 3.1	70B	82.11	94.35	17.79	5.47	48.42	20.99	24.00	10.63	35.84	25.85	36.54	148	5
Gemma 2	27B	76.91	94.89	17.34	4.15	47.40	21.84	9.33	7.65	33.01	10.44	32.30	127	6
Qwen 2.5	7B	88.74	89.52	14.87	2.83	43.02	24.44	13.33	7.75	18.82	8.86	31.22	119	7
Mistral Nemo	12B	41.59	82.26	9.95	3.40	41.27	22.86	20.00	7.75	26.70	11.04	26.68	114	8
Mixtral v0.1	8x7B	52.89	88.98	17.11	3.77	43.83	18.32	18.00	5.06	28.80	9.22	28.60	109	9
Gemma 2	9B	57.36	87.36	11.97	5.09	44.37	23.30	0.00	6.45	25.86	9.34	27.11	107	10
Llama 3.1	8B	45.63	63.71	7.03	2.08	39.65	19.67	14.00	5.36	23.93	6.31	22.74	87	11
IBM Granite 3.0	8B	74.46	58.33	4.34	1.51	29.74	25.04	4.00	1.29	20.02	4.13	22.29	74	12
Qwen 2.5	1.5B	82.07	76.88	11.51	0.19	29.00	21.71	6.67	2.38	13.41	4.73	24.86	72	13
Mistral v0.3	7B	27.52	41.40	0.00	0.94	37.09	22.69	1.33	4.17	18.52	5.70	15.94	63	14
IBM Granite 3.0	2B	32.03	63.97	6.37	0.19	21.51	23.27	2.67	0.99	14.97	4.25	17.02	55	15
SmolLM2	1.7B	29.80	28.23	0.00	0.00	25.76	15.99	9.33	4.57	13.95	4.37	13.20	48	16
Gemma 2	2B	27.74	12.90	0.00	0.57	31.56	20.93	0.67	3.97	12.87	3.64	11.49	42	17
Llama 3.2	1B	22.11	9.14	0.00	0.00	23.40	15.08	7.33	3.48	10.22	2.43	9.32	29	18
OLMo	7B	21.14	5.11	0.00	0.00	18.81	16.07	4.00	1.79	8.90	4.49	8.03	26	19
OLMo	1.5B	20.02	16.67	0.00	0.19	3.10	17.19	4.00	0.40	9.68	1.09	7.23	23	20

Initial Results



"CMR Scaling Law: Predicting Critical Mixture Ratios for Continual Pre-training of Language Models" Gu et al.

Conclusion

- Large amount of skilled engineering involved in training LLMs at scale
- Blizzard of research and competing models - hard to find the best path
- EuroHPC is a key resource!
- Main challenges to make it successful:
  - High quality training data
  - Getting the right evaluation

Plans for the future

- Building application for demonstrating FinLLM with Lloyds/NW
- More and better data, synthetic data, partner data, industry body data
- Collect human preference judgements
- Explore multimodal vision-text models mainly for documents