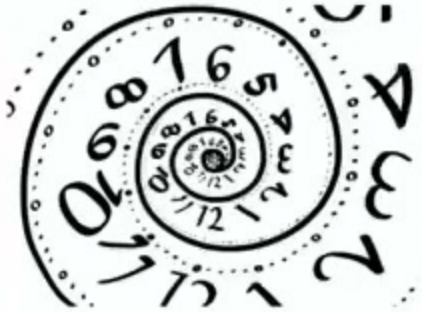
Open Source Artificial Intelligence

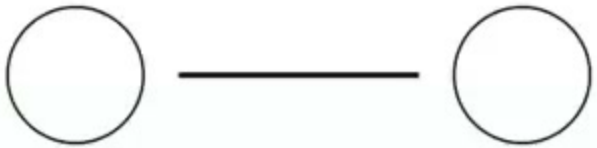
Pontus Stenetorp
p.stenetorp@cs.ucl.ac.uk
https://pontus.stenetorp.se

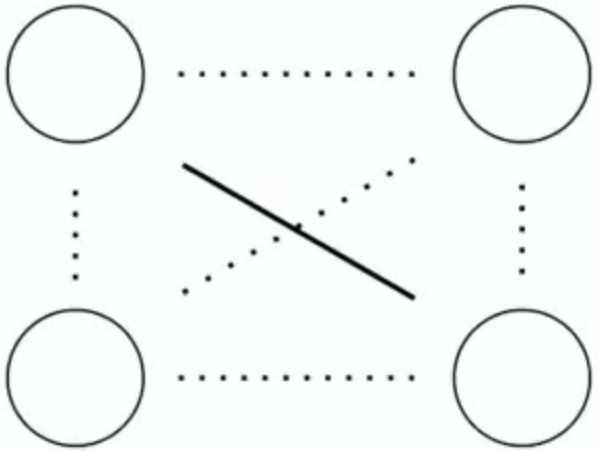
Professor of natural language processing
University College London

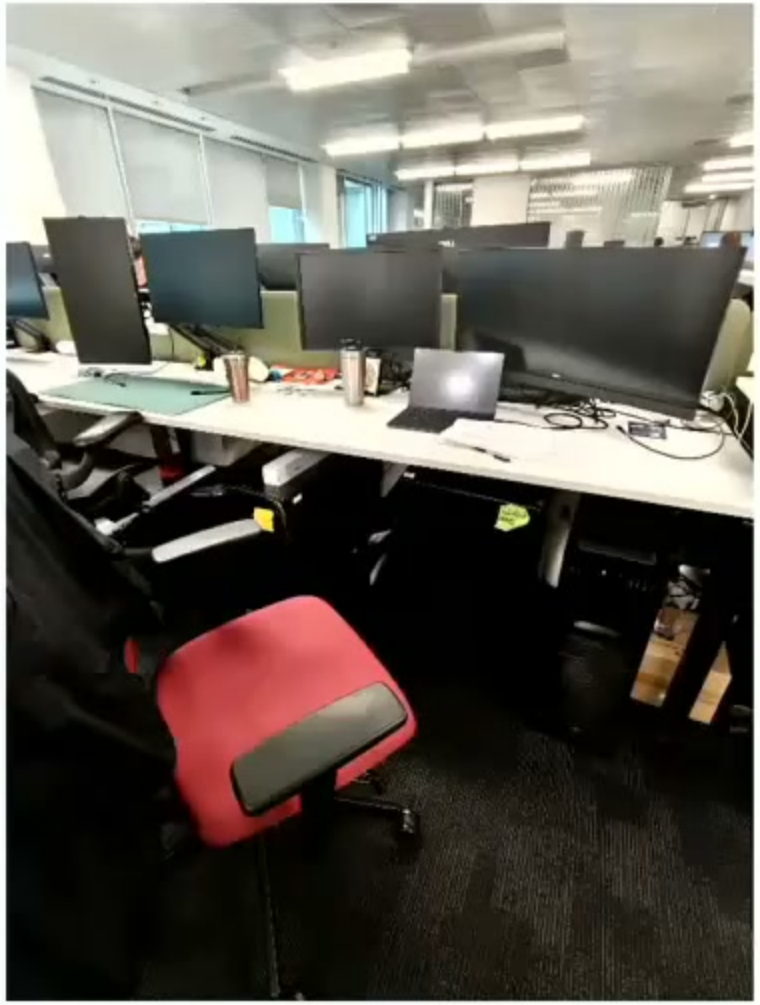


1. Unit of innovation
2. UK-LLM
3. Open Source Artificial Intelligence
4. What it will take

Unit of innovation







"When the bubble bursts again..."

Large Language Models

Training a large language model

Pre-training

Fine-tuning

Why pre-training?

UK-LLM

January 2023

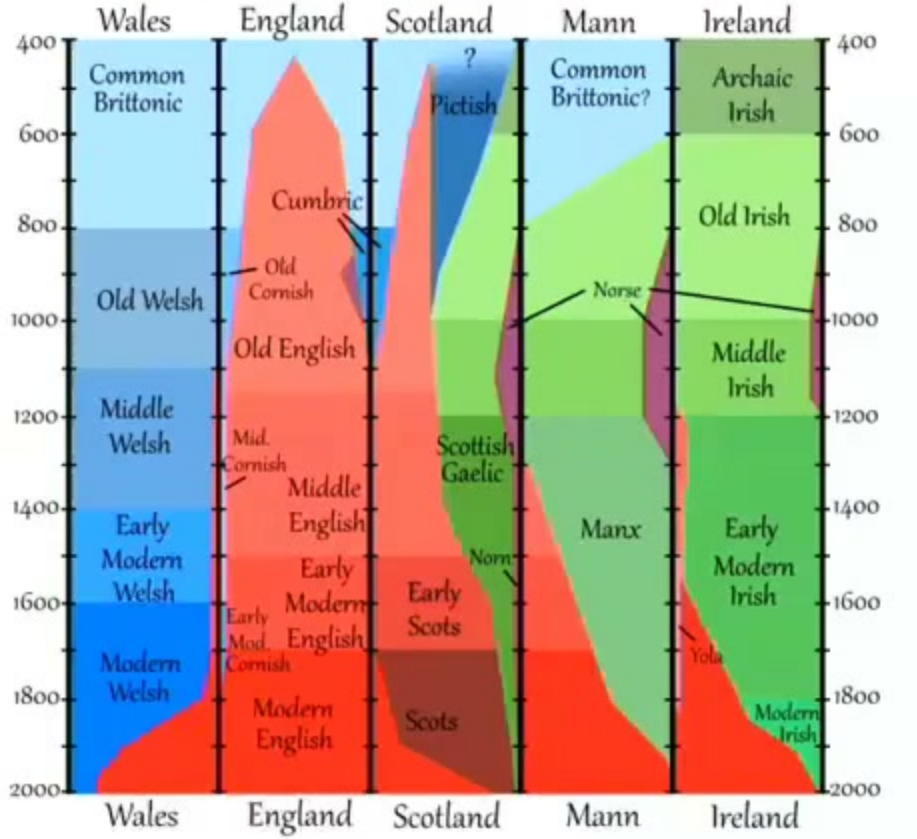
Why UK-LLM?

Know-how

Cultural/societal differences

Languages of the British-Irish Isles

Native Languages of Britain & Ireland Over Time



Caution

UK-Eval

Caernarfon 3B

Spring 2023

DiRAC

BritEval (Welsh)

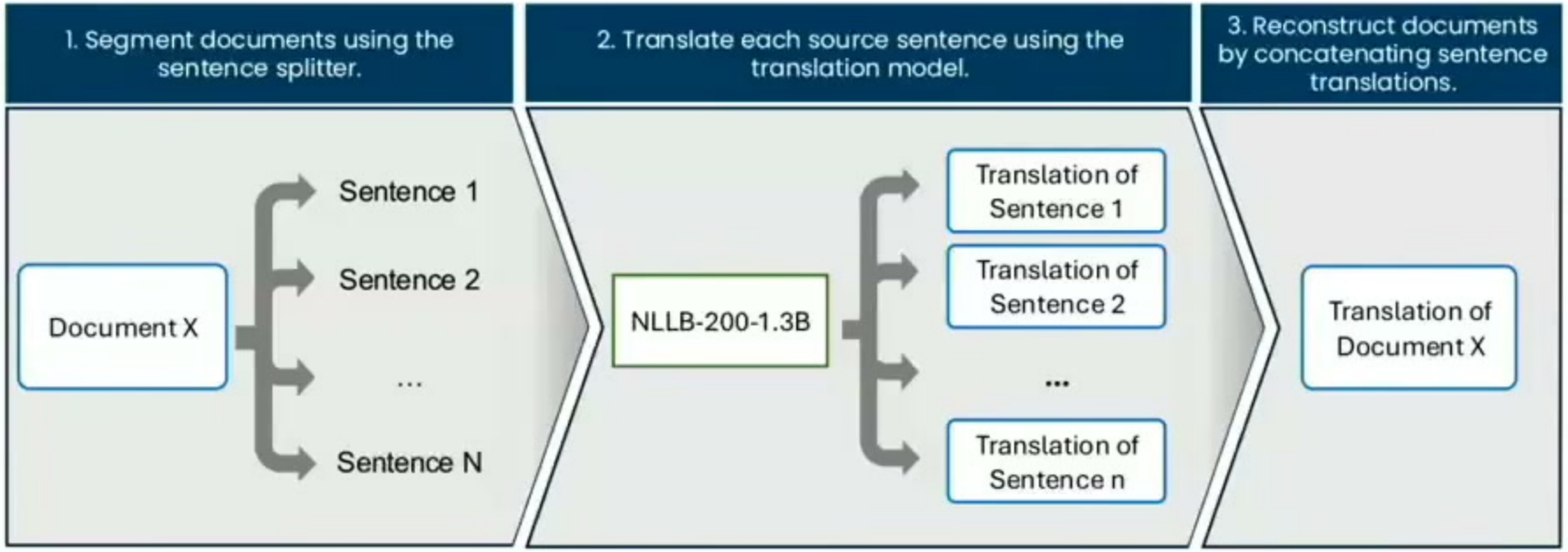
Model	Mean	ARC-e	PIQA	XNLI
Caernarfon 3B	46.2	40.8	60.1	37.7
LLaMA 3 8B	43.6	40.2	55.0	35.5
mGPT 13B	38.4	27.0	53.6	34.6
Phi 2	38.3	29.1	52.6	33.2
Mistral 7B	38.2	28.4	52.3	33.9
Bloom 7B	38.1	26.4	53.2	34.7

TransWebLLM 1B

Summer 2024

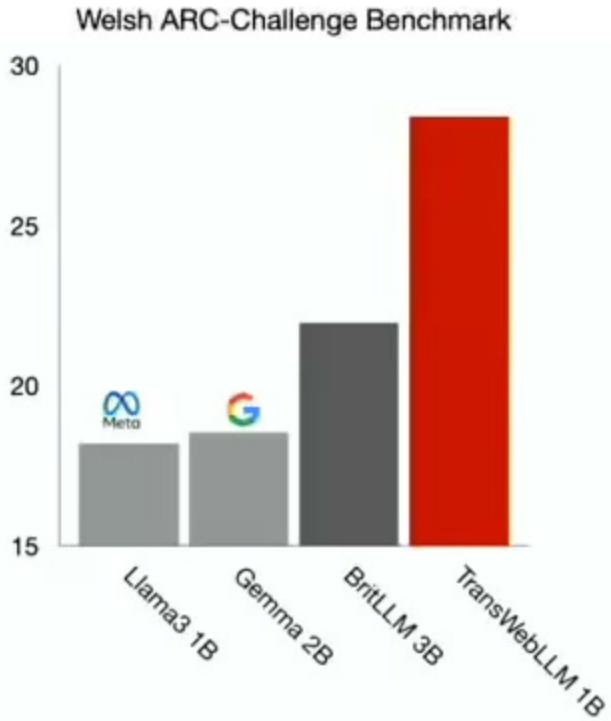
Curse of multi-linguality

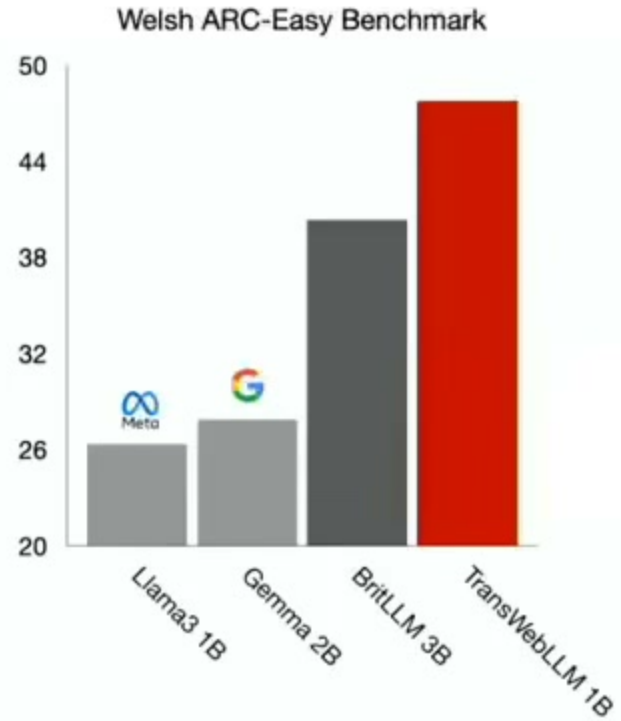
English superiority

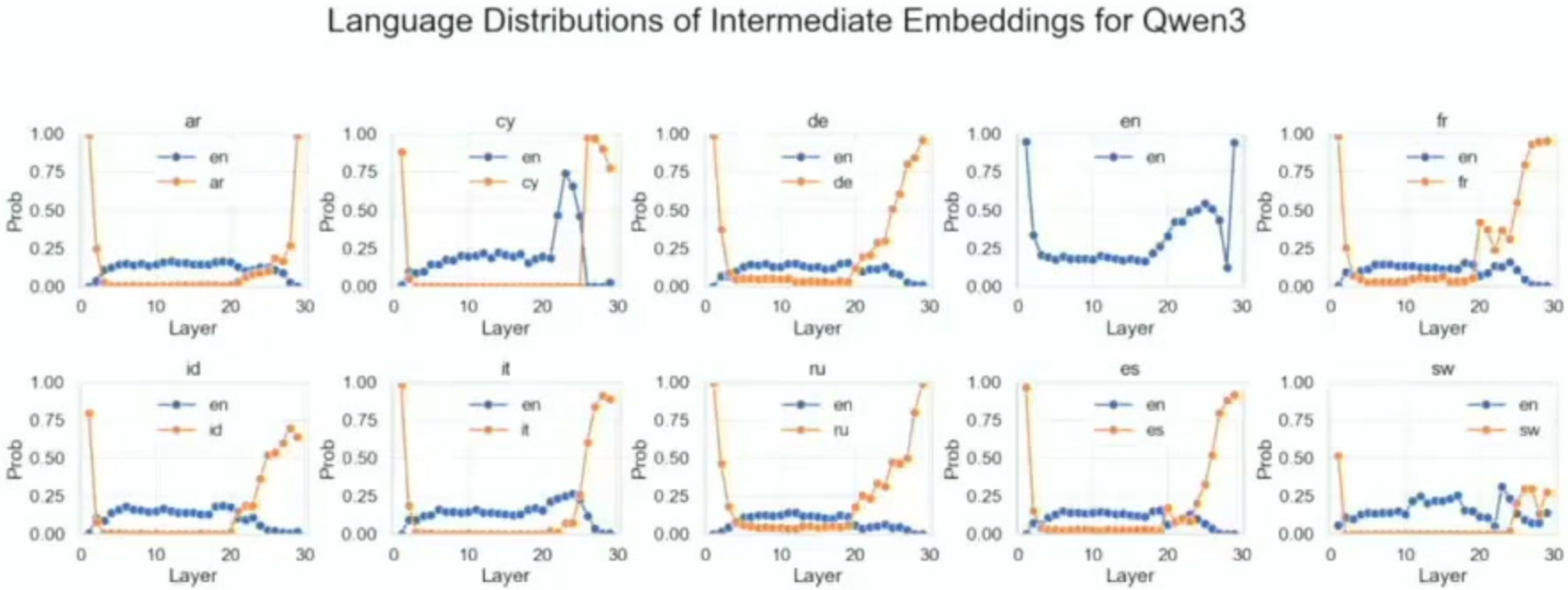

1. Segment documents using the sentence splitter.
2. Translate each source sentence using the translation model.
3. Reconstruct documents by concatenating sentence translations.
Document X
Sentence 1
Sentence 2
...
Sentence N
NLLB-200-1.3B
Translation of Sentence 1
Translation of Sentence 2
...
Translation of Sentence n
Translation of Document X

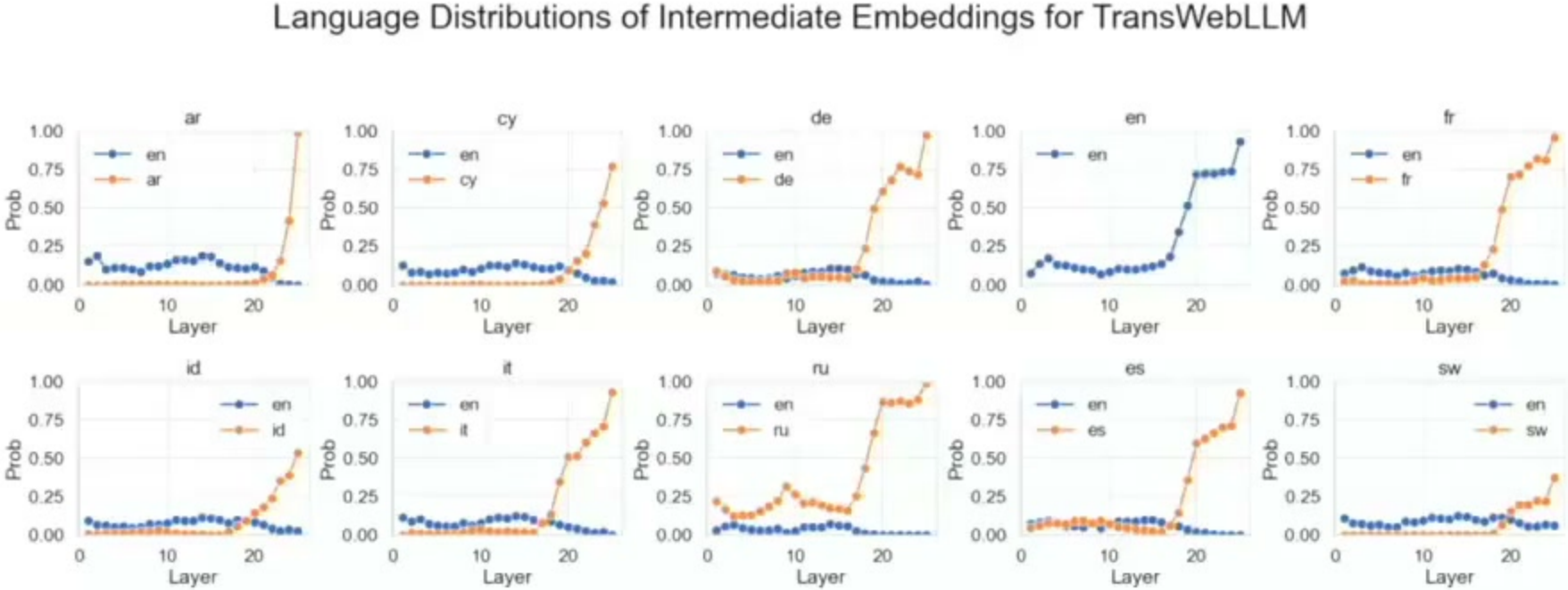
Language	Tokens (B)	Avg. Doc Length (tokens)
Arabic	311.35	3,201
English	114.95	1,182
French	143.71	1,479
German	140.70	1,447
Indonesian	174.12	1,792
Italian	140.32	1,447
Russian	157.40	1,618
Spanish	140.99	1,449
Swahili	183.55	1,887
Welsh	201.49	2,071
Total	1,708.58	1,757

Isambard-AI Phase 1


Welsh ARC-Challenge Benchmark
30
25
20
15
Meta
Llama3 1B
Gemma 2B
BritLLM 3B
TransWebLLM 1B


Welsh ARC-Easy Benchmark
50
44
38
32
26
20
Meta
Llama3 1B
Gemma 2B
BritLLM 3B
TransWebLLM 1B

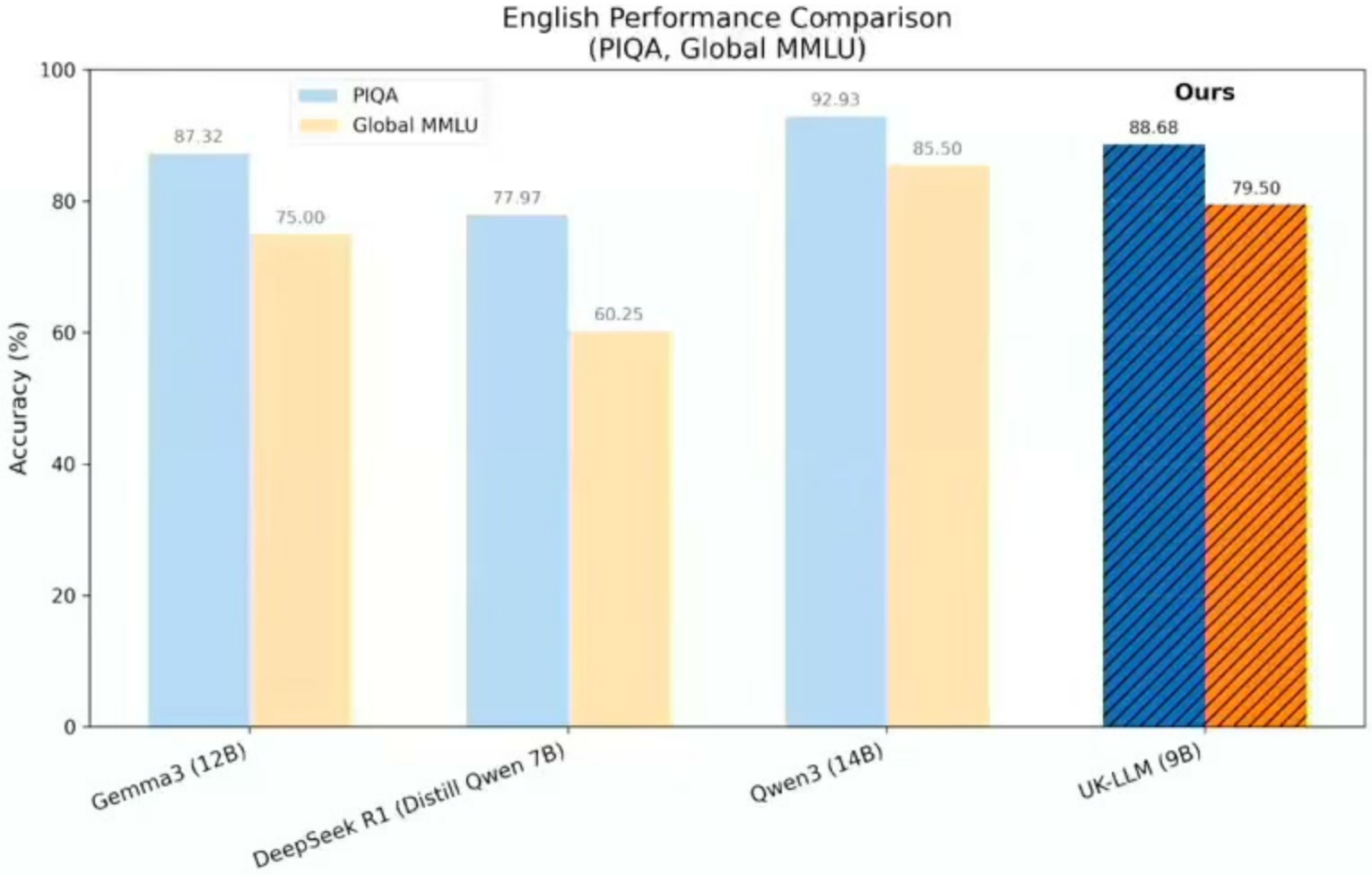

Language Distributions of Intermediate Embeddings for Qwen3
ar
cy
de
en
fr
id
it
ru
es
sw
en
ar
cy
de
fr
id
it
ru
es
sw
Prob
Layer
0.00
0.25
0.50
0.75
1.00
0
10
20
30

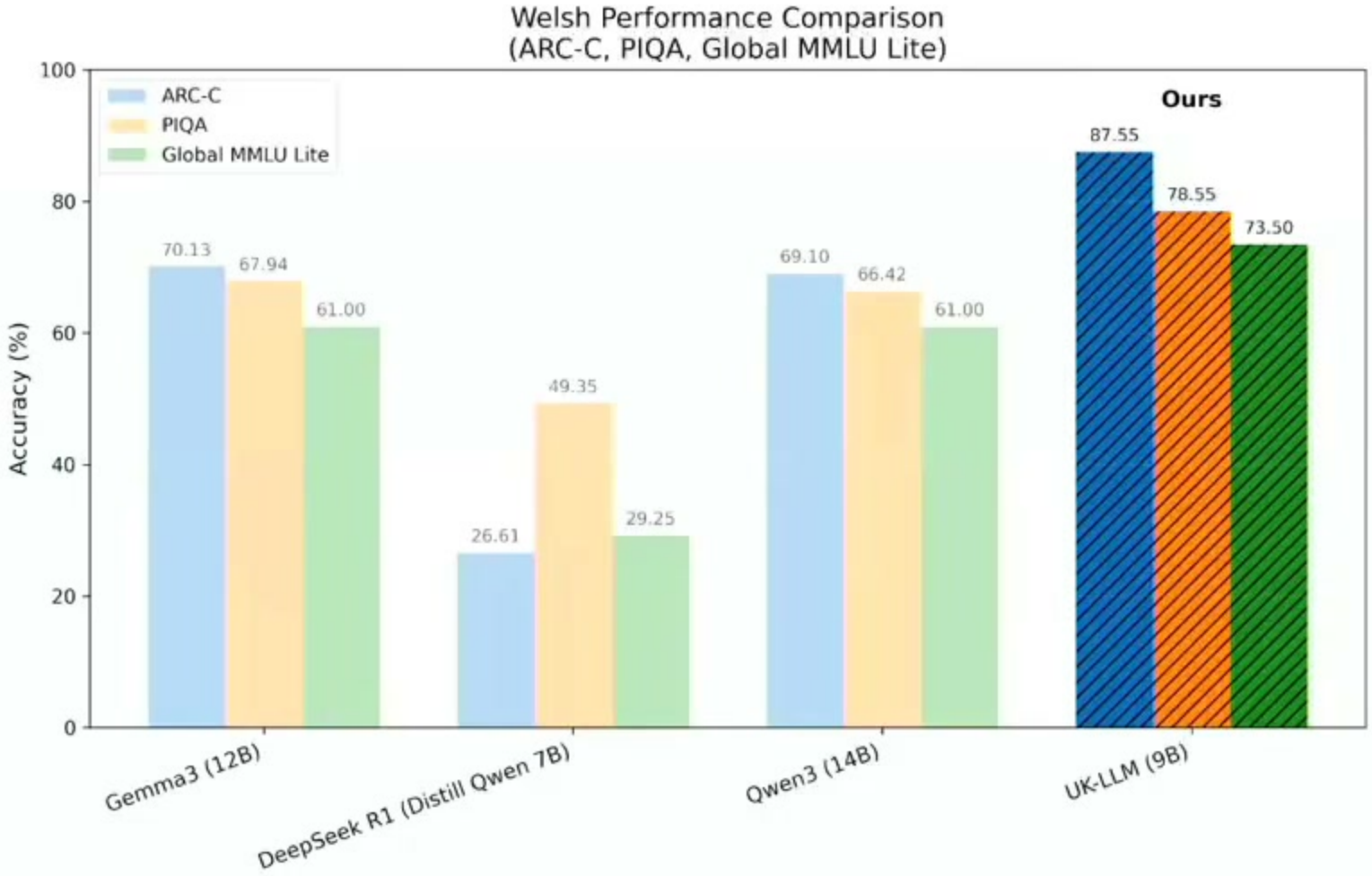

Language Distributions of Intermediate Embeddings for TransWebLLM
ar
cy
de
en
fr
id
it
ru
es
sw
en
ar
cy
de
fr
id
it
ru
es
sw
Prob
Layer
0.00
0.25
0.50
0.75
1.00
0
10
20

Model	English	French	German
English Model	54.56	-	-
English and French Model	54.86	39.80	-
English, French, and German Model	54.39	38.59	37.09
CuatroLLM 1.3B	54.58	38.53	37.58

UK-LLM Nemotron 9B

August 2025

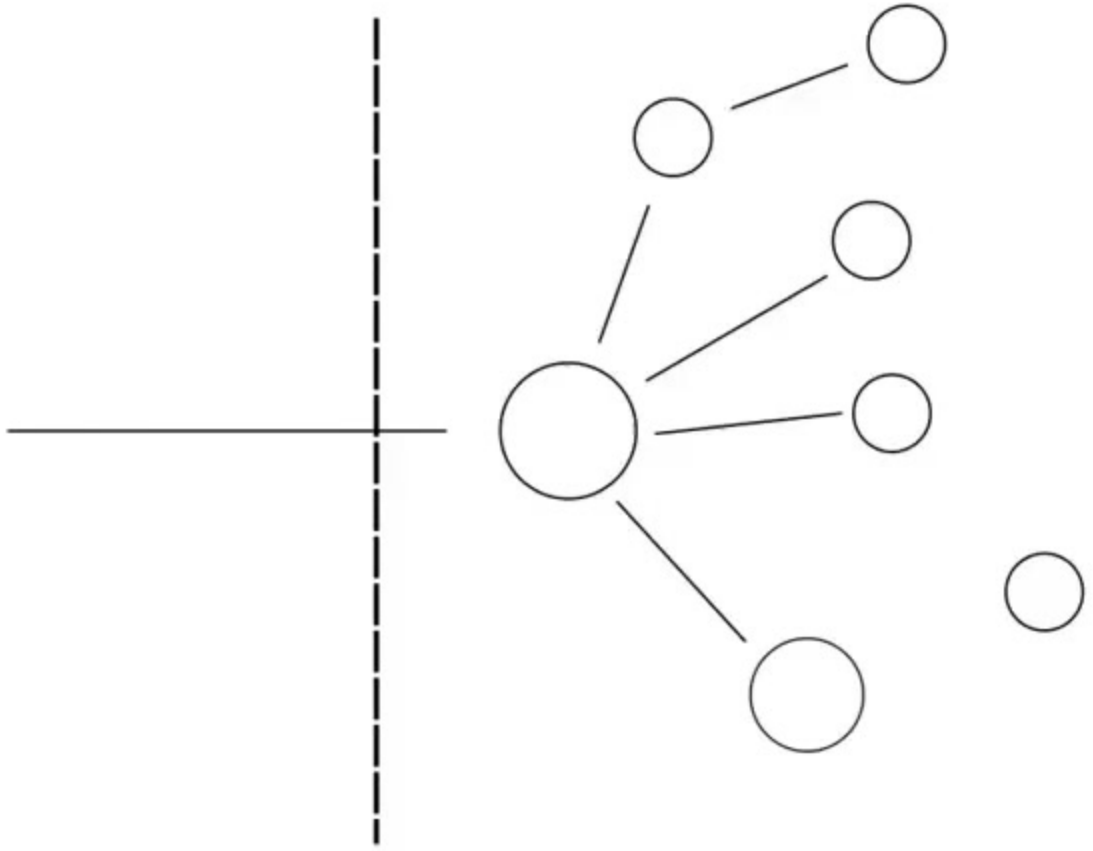

English Performance Comparison
(PIQA, Global MMLU)
PIQA
Global MMLU
Ours
Accuracy (%)
100
80
60
40
20
0
87.32
75.00
77.97
60.25
92.93
85.50
88.68
79.50
Gemma3 (12B)
DeepSeek R1 (Distill Qwen 7B)
Qwen3 (14B)
UK-LLM (9B)


Welsh Performance Comparison
(ARC-C, PIQA, Global MMLU Lite)
ARC-C
PIQA
Global MMLU Lite
Ours
Accuracy (%)
100
80
60
40
20
0
70.13
67.94
61.00
26.61
49.35
29.25
69.10
66.42
61.00
87.55
78.55
73.50
Gemma3 (12B)
DeepSeek R1 (Distill Qwen 7B)
Qwen3 (14B)
UK-LLM (9B)

Press release "story"

"This is a powerful example of how the latest AI technology, trained on the UK's most advanced AI supercomputer... can serve the public good"

Sofair



OSAI Working Group

UK-MLM: Multi-modal, joint pre-training

Geospatial: Highest detail, open source model

Large-scale, distributed agentic coordination

Open Source Artificial Intelligence

Open Source AI ≠ Open Source Software

Degrees of openness

Why does it matter?

Science

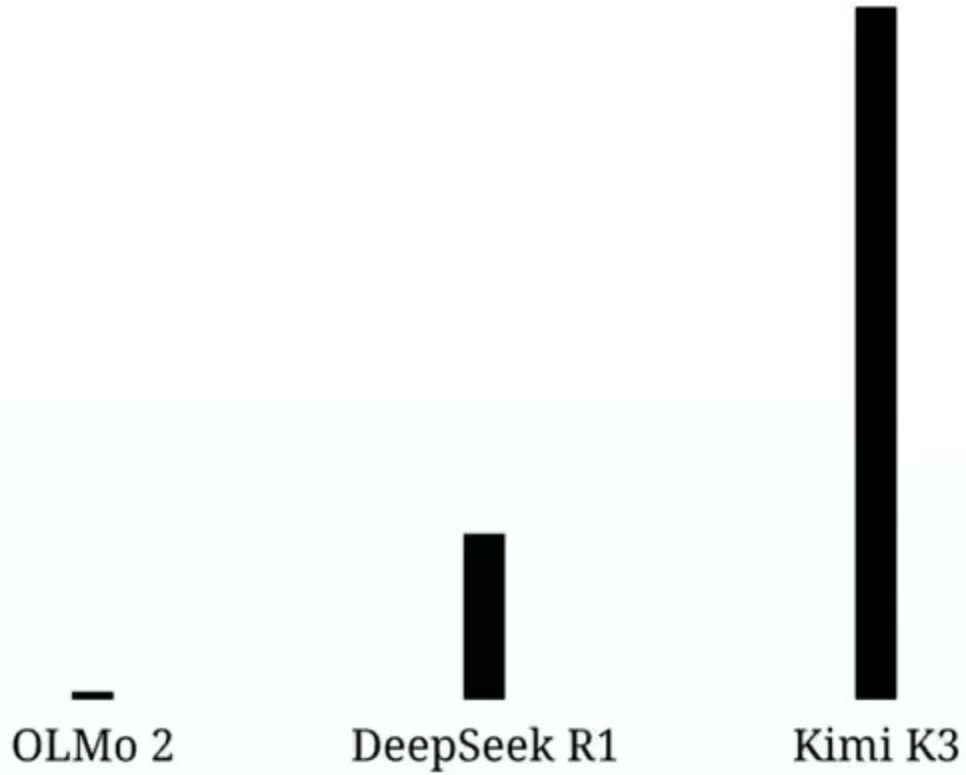
Economy

Society

There is a community waiting

What it will take

Know-how and manpower


OLMo 2
DeepSeek R1
Kimi K3

* Sofair: 2,000,000 GPUh/year
* ◆◆◆◆◆◆◆: ~10,000,000
* ◆◆◆◆◆: ~20,000,000
* Isambard-AI Phase 2: ~46,000,000

Largest viable OSAI model

40,000,000 GPUh/year

Industry donations or collaborations?

Sovereign ⊭ Isolated

Global Open Source AI Summit

Sofair: A *starting* point

1. Reconsider our "unit of innovation"
2. Expectations on publishing (for careers too)
3. "Refresh" strategy for compute
4. Long-term copyright reform

Thank you for your attention, questions or comments?