ALLaM: Large Language Models for Arabic and English

M Saiful Bari*, Yazeed Alnumay*,
Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya,
Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie,
Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran,
Yousef Almushayqih, Raneem Alnajim,
Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan,
Majed Alrubaian, Ali Alammari, Zaki Alawami,
Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose,
Abdalghani Abujabal*, Nora Al-Twairesh*, Areeb Alowisheq*, Haidar Khan*

National Center for Artificial Intelligence
Saudi Authority for Data and Artificial Intelligence

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

NCAI
المركز الوطني للذكاء الاصطناعي
National Center for AI

ALLaM: Large Language Models for Arabic and English

- **Second Language Acquisition**
  - Language Alignment at Scale
  - Continue Pre-train from LLaMA-2
    - 7b, 13b, 70b llama-2
  - Process of Vocabulary Expansion
  - Vocabulary initialization
- **Training From Scratch**
  - 7B, 34B param
- **Post training Alignment**
  - Scaling alignment data
  - Multi-turn DPO
- **Training Details**
  - data mixture
  - learning rate
  - grad norm
  - eval indicators during training
- **Evaluation**
  - Arabic specific eval
  - Multi-turn eval
- **Opensource Assets**

  https://huggingface.co/ALLaM-AI/

ALLaM: Large Language Models for Arabic and English

NCAI المركز الوطني للذكاء الاصطناعي National Center for AI

SDAIA الهيئة السعودية للبيانات والذكاء الاصطناعي Saudi Data & AI Authority

M Saiful Bari *† Yazeed Alnumay†
Norah A. Alzahrani Nouf M. Alotaibi Hisham A. Alyahya
Sultan AlRashed Faisal A. Mirza Shaykhah Z. Alsubaie
Hassan A. Alahmed Ghadah Alabduljabbar Raghad Alkhathran
Yousef Almushayqih Raneem Alnajim
Salman Alsubaihi Maryam Al Mansour
Majed Alrubaian Ali Alammari Zaki Alawami
Abdulmohsen Al-Thubaity Ahmed Abdelali Jeril Kuriakose
Abdalghani Abujabal† Nora Al-Twairesh† Areeb Alowisheq† Haidar Khan†

National Center for AI (NCAI), Saudi Data and AI Authority (SDAIA)
Riyadh, Saudi Arabia

Cross-lingual Alignment at Scale, WHY?

ogy (Weidinger et al., 2021). Judging from the initial capabilities (Bubeck et al., 2023), the potential of these frontier models are reinventing the way humans interact with machines, impacting social norms, productivity, trends, and culture on a broader scale (Zhou et al., 2024). However, most of these frontier-class models are primarily trained on English and often lack a connection to localized regional cultures and norms (Naous et al., 2024). This gap has the potential to result in slow and irreversible manipulation of regional identities and lead to cultural homogenization.

Scare 1: Scaling manipulation ***campaign*** via alignment hacking

Scare 2: Scaling ***Autonomous*** intelligence without empathy, helpfulness etc.

Scare 3: Discounting mistake as "Humanly", Are we ready to do that with AI systems? "Anthropomorphizing" or not

Sovereign LLM: by the people, for the people

The natural course to reverse this trend is to invest resources in curating data and building models to support the diversity of languages and cultures represented in the modern world. While this is possible, the significant training costs of LLMs and their environmental impact have become major concerns in recent years (Strubell et al., 2019). The vast computational resources required to train LLMs contribute to substantial carbon emissions (Luccioni & Hernandez-Garcia, 2023). Governments [2] and non/for-profit organizations (Dodge et al., 2022; Google, 2021; Amazon, 2021), are increasingly aware of these issues. This awareness has led to discussions about the ethical implications of AI development and the need for sustainable practices concerning "*When and how to scale the training of these models*". In addition, curating data for each language/region at pretraining scale is also a difficult task, since most available data comes from a few high-resource languages.

Responsible Scaling

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

LLM training at Scale

Training an LLM is like *cooking*, you can prepare the same dish in many different ways.
The convergence behavior is mostly theoretically non-interpretable.

At Scale model training is always a **trade-off** between:

(i) the compute resources available
(ii) the algorithms available (both optimization and structural priors)
(iii) the data available
(iv) understanding of scaling laws

Model Trainer(s) take a series of decisions based on

(i) Compute Risk
(ii) Prior research
(iii) Theoretical background
(iv) Ablations
(v) Intuitions

Cross-lingual Alignment at Scale

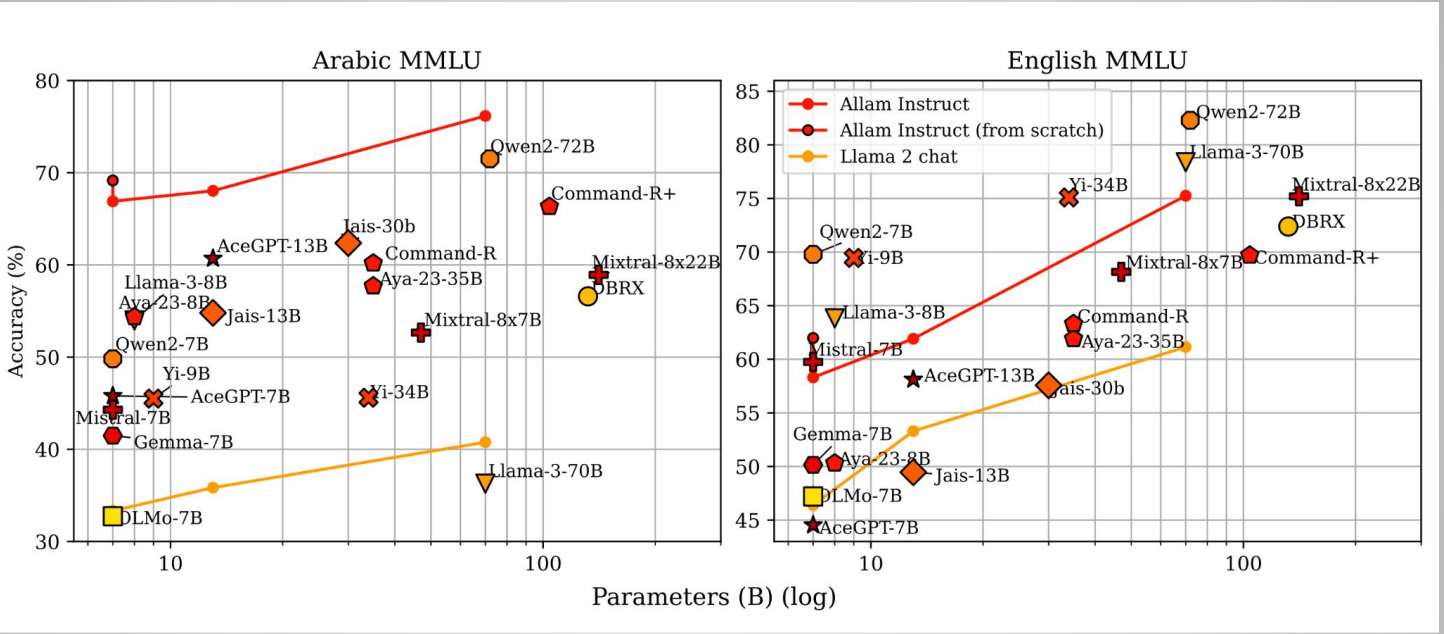


Figure: Performance on Arabic (Koto et al., 2024)[1] and English (Hendrycks et al., 2020) MMLU [2] Benchmarks

[1] Koto, Fajri, et al. "Arabicmmlu: Assessing massive multitask language understanding in arabic." arXiv preprint arXiv:2402.12840 (2024).
[2] Hendrycks, Dan, et al. "Measuring massive multitask language understanding." arXiv preprint arXiv:2009.03300 (2020).

Challenges in a Second Language Acquisition - Tokenization

Tokenization Example

LLaMA-2 – 11 Tokens

'ال', 'س', 'ل', 'ا', 'م', ' ', 'ع', 'ل', 'ي', 'ك', 'م'

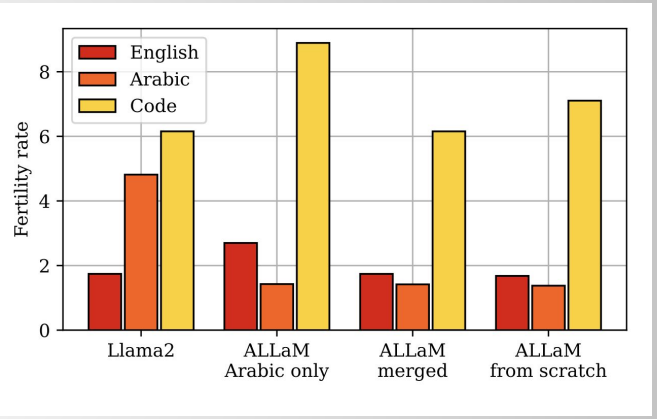
LLaMA-3 – 4 Tokens

'ال', 'سلام', ' عليك', 'م'

ALLaM – 2 Tokens

'السلام', 'عليكم'

Does it only helps inference?



Challenges in a Second Language Acquisition - Tokenization

Tokenization Example

LLaMA-2 – 11 Tokens

'ال', 'س', 'ل', 'ا', 'م', ' ', 'ع', 'ل', 'ي', 'ك', 'م'

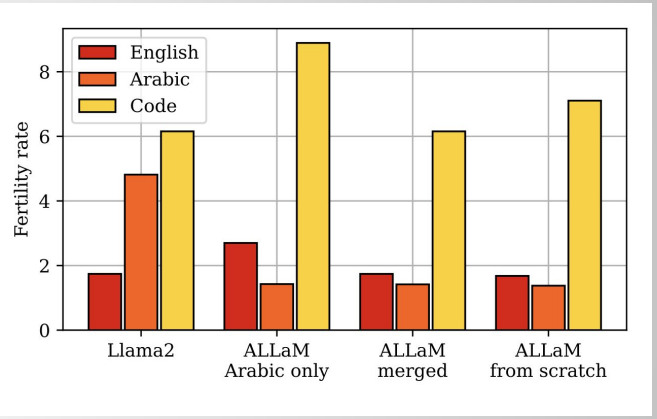
LLaMA-3 – 4 Tokens

'ال', 'سلام', ' عليك', 'م'

ALLaM – 2 Tokens

'السلام', 'عليكم'

Does it only helps inference?

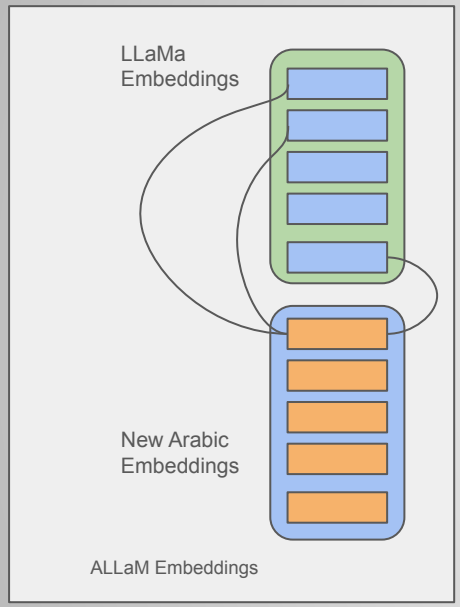


230B Word → 1.2T token

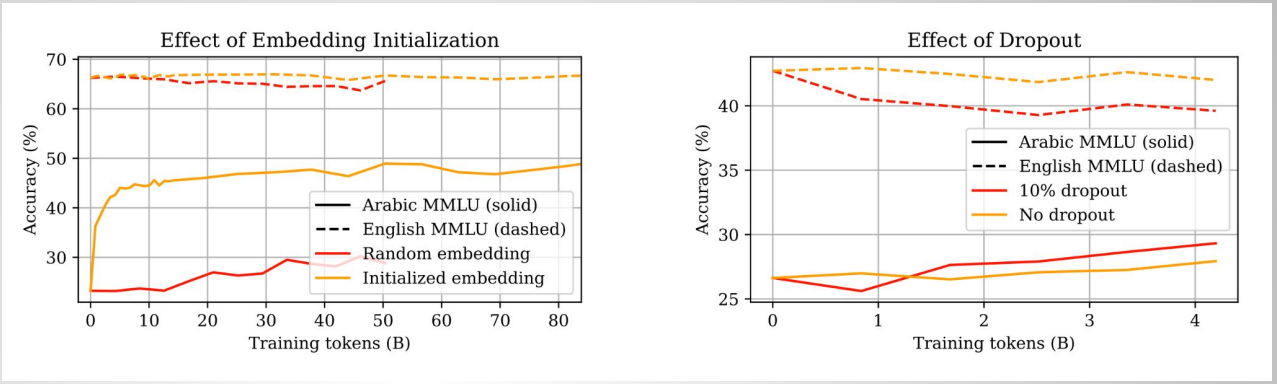
230B Word → 308B token

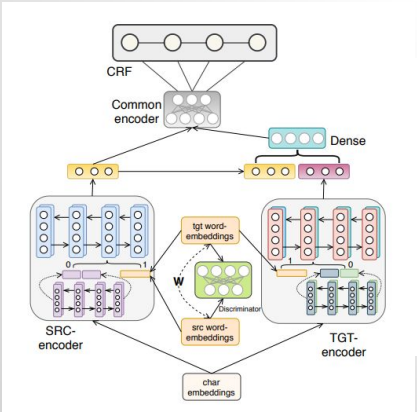
Challenges in a Second Language Acquisition - Embedding Initialization; Things to ignore

[1]

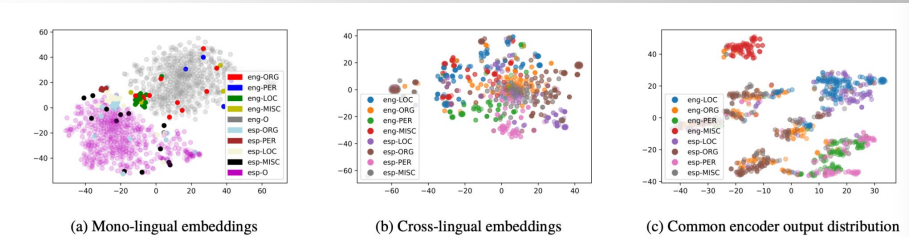


[1]





[2]

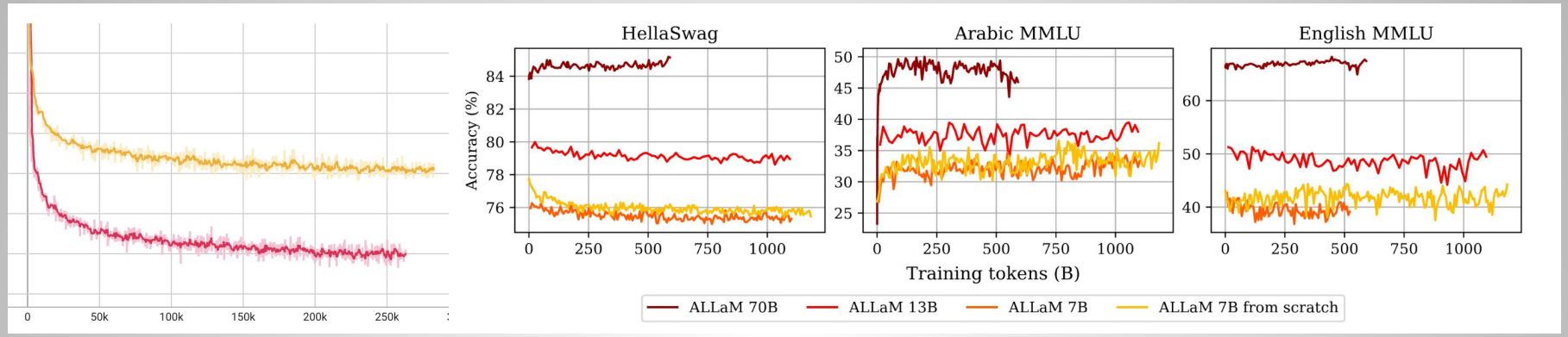


(a) Mono-lingual embeddings (b) Cross-lingual embeddings (c) Common encoder output distribution

Figure 3: t-SNE plot of NER tagged embeddings of two languages with 1000 samples: (a) Mono-lingual embeddings (fasttext), (b) Cross-lingual embeddings after word-level adversarial training, (c) Embeddings from our common encoder.

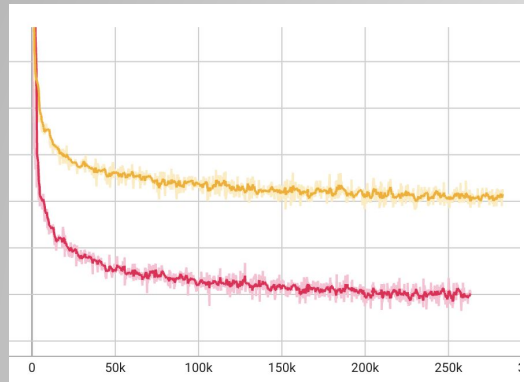
[1] Saiful Bari, M., et al. "ALLaM: Large Language Models for Arabic and English." arXiv e-prints (2024): arXiv-2407.
[2] Bari, M. Saiful, Shafiq Joty, and Prathyusha Jwalapuram. "Zero-resource cross-lingual named entity recognition." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 05. 2020.

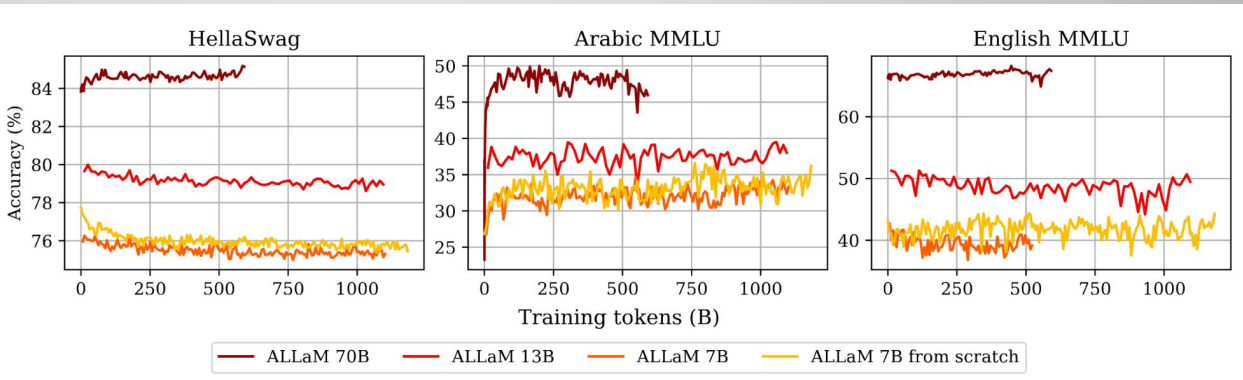
Challenges in a Second Language Acquisition: Evals



What's the most boring part of the figure? No significant benchmark improvement

Challenges in a Second Language Acquisition: Evals

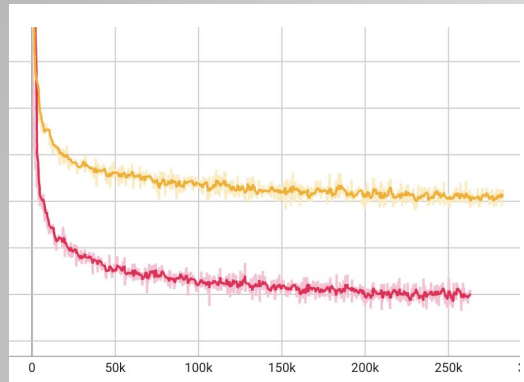


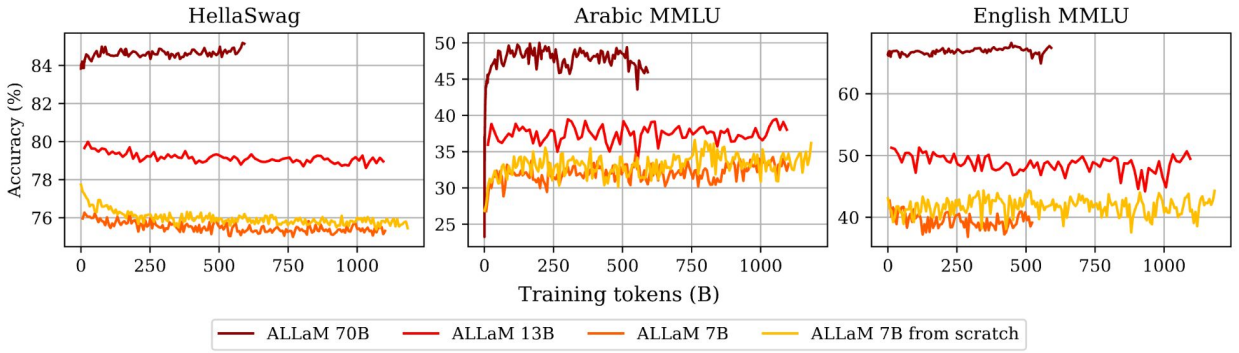


What's the most boring part of the figure? No significant benchmark improvement

What's the confusing part of the figure? Loss is still decreasing

Challenges in a Second Language Acquisition: Evals



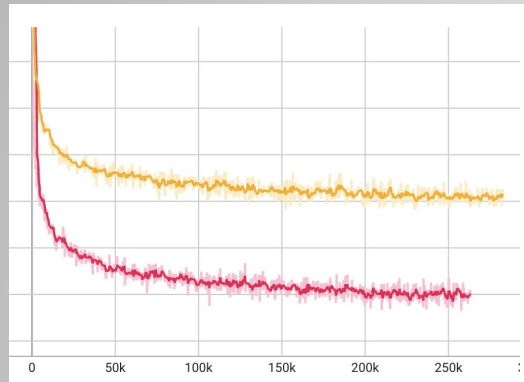


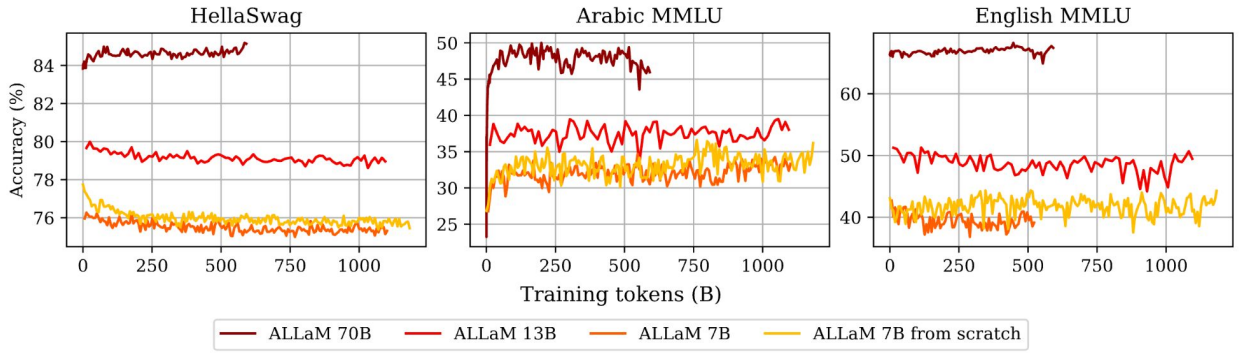
What's the most boring part of the figure? No significant benchmark improvement

What's the confusing part of the figure? Loss is still decreasing

Are we watching the right thing? Probably not

Challenges in a Second Language Acquisition: Evals





Are we watching the right thing?

Probably not

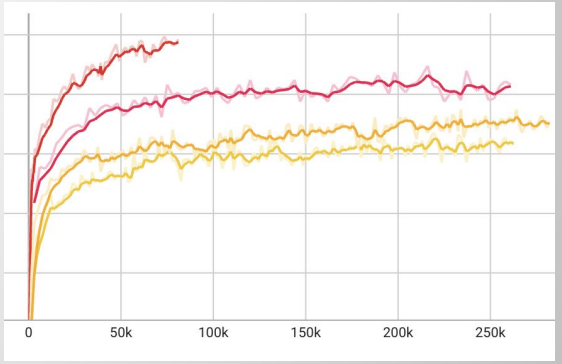
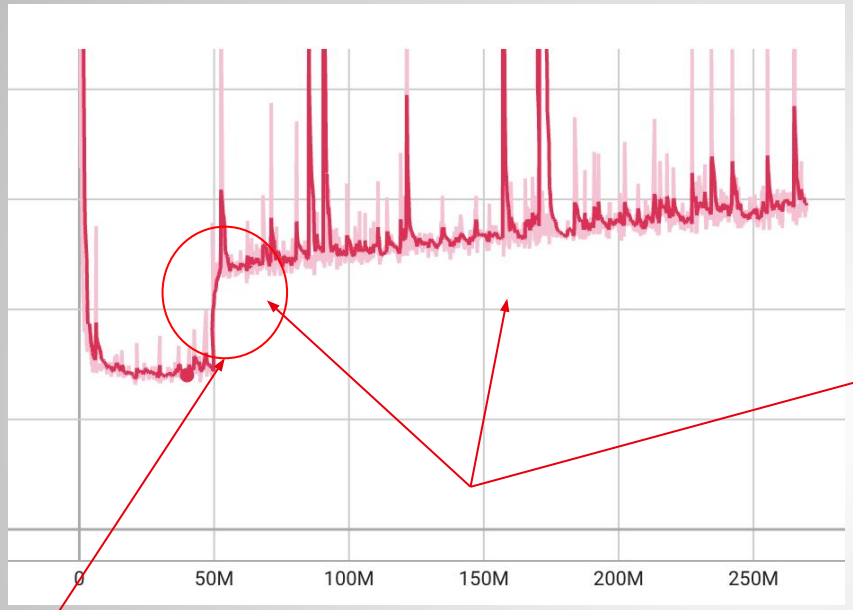
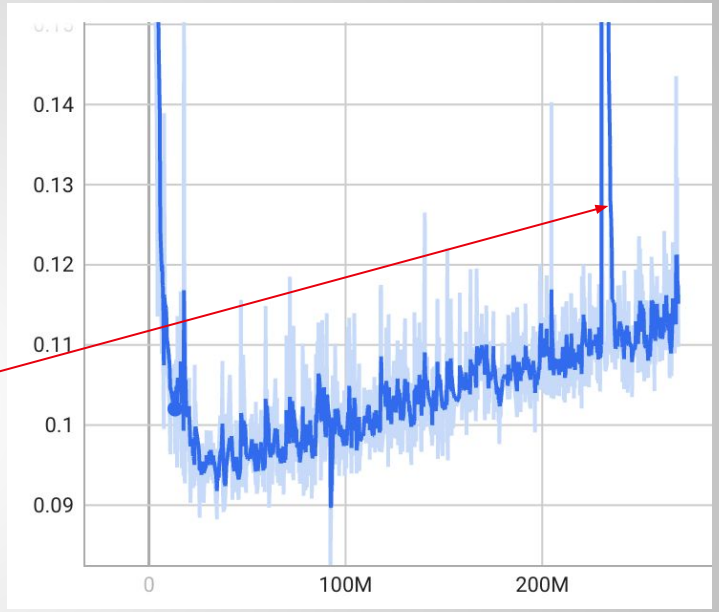


Fig: Araswag 10 shot

What to look during training: Grad-Norm



Resume implemented with warmup. Affected by lr-rate.

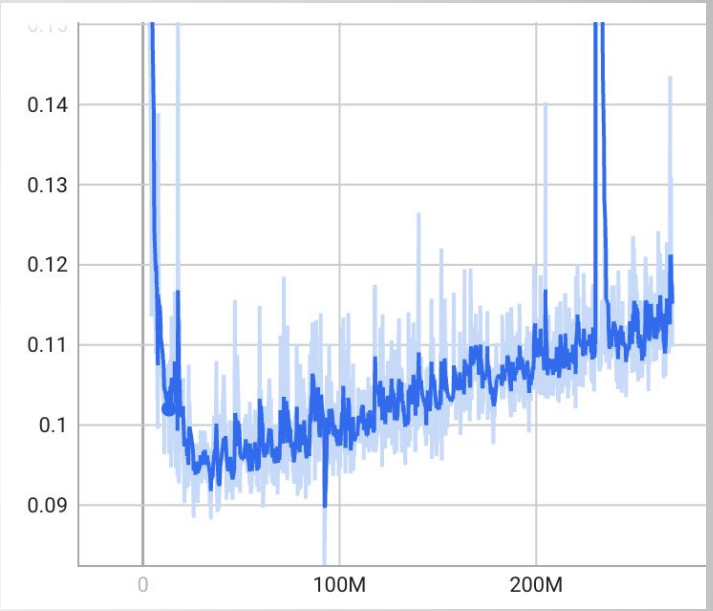


A run without any issue and proper resume.

$$||\nabla\mathcal{L}(\theta)||_2 = \sqrt{\sum_{i=1}^{n}\left(\frac{\partial\mathcal{L}}{\partial\theta_i}\right)^2}$$

What to look during training: Grad-Norm (Scare!)

- Gradient norm starts very high at training initialization, showing maximum model uncertainty.
- Gradient norm sharply drops and reaches its lowest point during early training as basic patterns are learned.
- Gradient norm gradually climbs up again during later training phases as model fine-tunes complex patterns.
- Gradient Norm is affected by
  - Learning rate
  - Data distribution



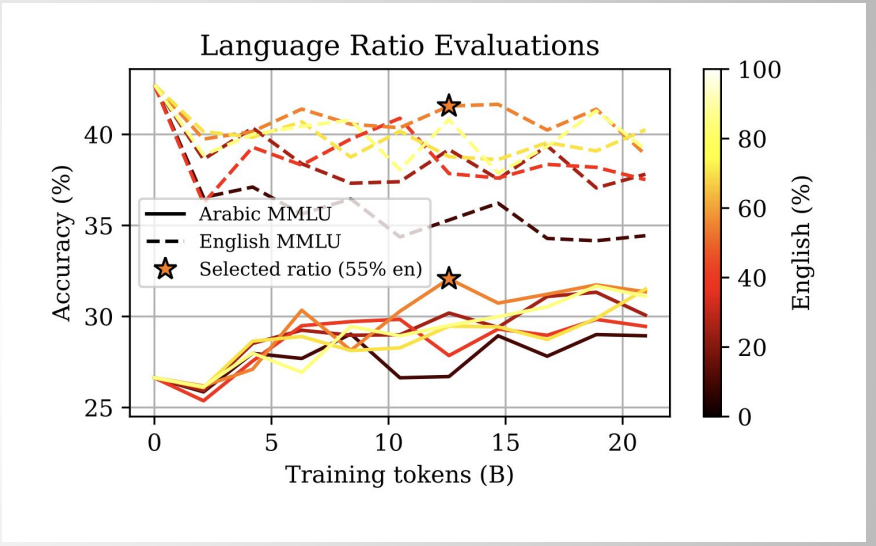
This is a constant learning rate experiments

Cross-lingual Alignment at Scale

What happens when you don't know the Scaling Law?

Issue with early-fusion on multilingual data.

- Curse of Multilinguality.
- Lower bound to the regular scaling law.
- Mu-transfer[3] may break.
- Working on a Scaling law with 500B-1T compute budget is not possible.



Recipe to cheat Scaling Law:

Responsible Scaling: Follow Scaling law for English -> rapidly adapt/transfer knowledge to a target distribution (in this case, it's a language)

[3] Yang, Greg, et al. "Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer." arXiv preprint arXiv:2203.03466 (2022).

Cross-lingual Alignment at Scale: Notion of data sampling, Domain, Epoch

- Epoch vs Percentage of Domain
- Adding Code related data
- Mechanistic Interpretability approach
  - What makes less hallucination

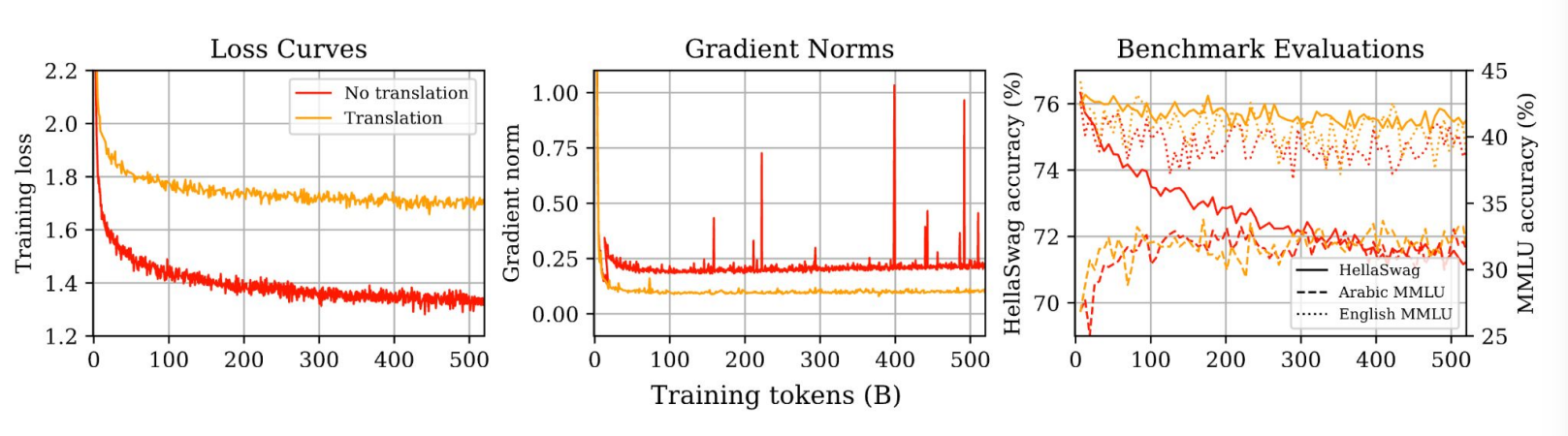
Domain	Mixed Arabic & English				English Only
	English	Arabic		Mixed	
		Natural	Translated		
Web	31%	71%	65%	48%	71%
Books	9%	13%	12%	11%	3%
Wiki	—	0.70%	0.61%	0.3%	0.1%
News	—	14%	—	3%	—
Science	16%	—	22%	14%	6%
Code	39%	—	—	21%	17%
Math	5%	—	—	2.5%	0.9%
Other	—	1.3%	0.39%	0.2%	2%
Lang Mix	55%	22.5%	22.5%	100%	100%
Tokens	660B	270B	270B	1.2T	4T

Cross-lingual Alignment at Scale: Filtering english data

- Red-pajama-v2 snapshot
- Meta data basis filtering
- Only done for english

Annotation Tag	Description	Threshold
ccnet_language_score	Language identification model score.	Keep ≥ 0.6
ccnet_length	Number of characters in the document.	Drop < 150 characters
ccnet_nlines	Number of lines in the document.	Drop < 3 lines
rps_doc_ml_palm_score	FastText classifier prediction for document classification as Wikipedia, OpenWebText, or RedPajama-V1 book (English only).	Sample according to distribution
rps_doc_frac_lines_end_with_ellipsis	Fraction of lines ending with an ellipsis ("..." or "..").	Drop ≥ 0.8
rps_doc_frac_no_alph_words	Fraction of words without any alphabetical characters.	Drop ≥ 0.9
rps_doc_lorem_ipsum	Ratio of occurrences of "lorem ipsum" to total characters in content (after normalization).	Drop ≥ 0.5
rps_doc_stop_word_fraction	Ratio of stop words to total words in the document, using stop words from here.	Drop ≥ 0.9
rps_doc_symbol_to_word_ratio	Ratio of symbols ("#", "..." or "..") to words in content.	Drop ≥ 0.9
rps_doc_ldnoobw_words	Count of sequences from the List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words blocklist (see here).	Drop ≥ 0.9 (ldnoobw_words/total_words)
minhash_signature	Minhash signature for fuzzy deduplication at Jaccard similarity of 0.7, based on 128 hash functions grouped into 14 bands of 9 rows for LSH.	0.7 (ldnoobw_words/total_words)

Cross-lingual Alignment at Scale: Adding parallel translation data



Adding translation data helped up stabilizing training, specifically having smooth grad norms. Non-intuitive loss?

Cross-lingual Alignment at Scale: Learning Rate

Learning rate vs convergence

- LlaMa-2's final learning rate was 3e-5.
- Llama-3.1's final learning rate is 0.

How learning rate interfere convergence

- Once the learning rate went down, it's very difficult to ramp-up learning rate without some forgetting
- hyperparameter tuning
  - 1e-3 … 3e-4
  - each for 10B tokens
- We had to train on the constant learning rate 3.0e-5

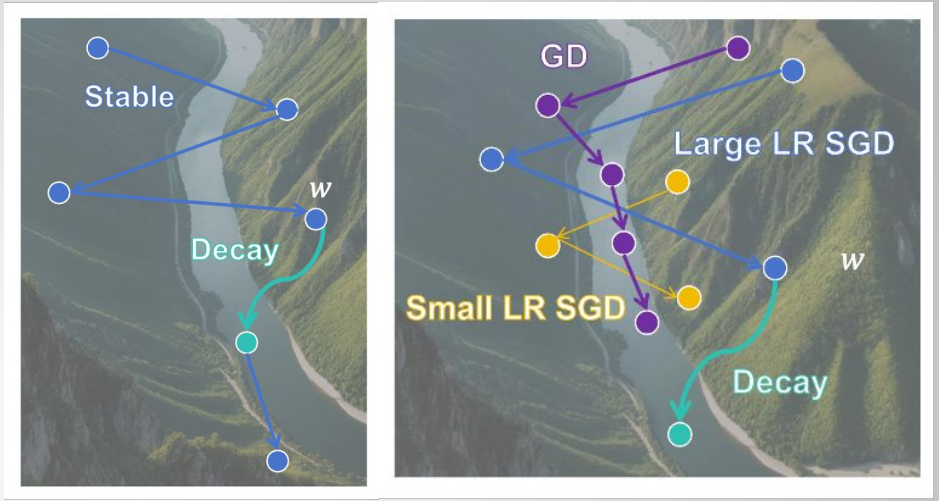


Figure 2 : River valley loss decay analogy [1]

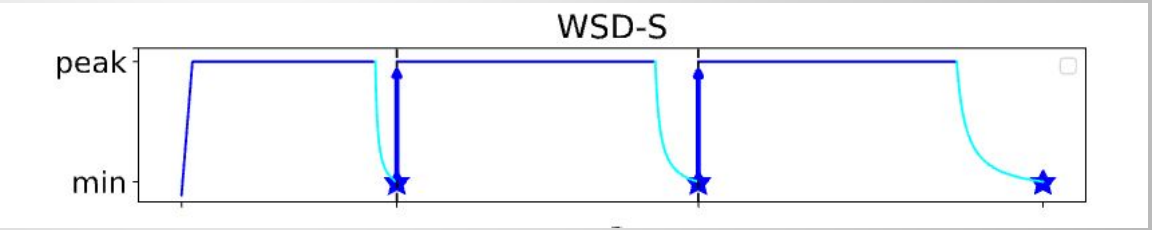


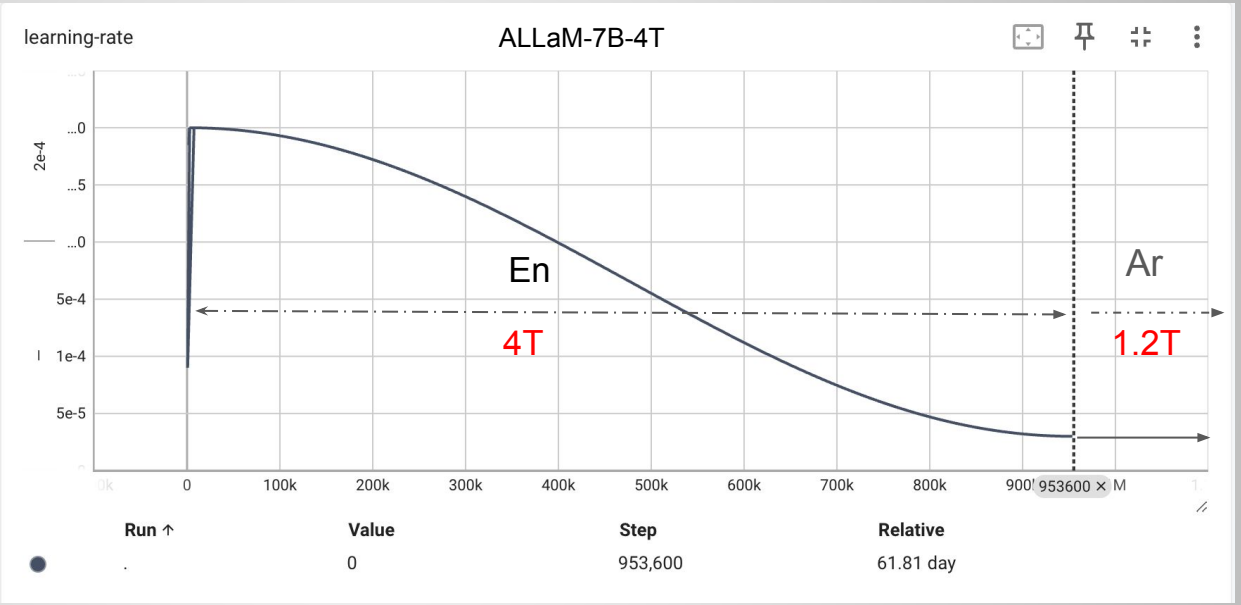
Figure 3: Learning rate scheduling of Warmup-Stable-Decay

[4] Wen, Kaiyue, et al. "Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape View." The Thirteenth International Conference on Learning Representations.
[5] Hu, Shengding, et al. "Minicpm: Unveiling the potential of small language models with scalable training strategies." arXiv preprint arXiv:2404.06395 (2024).

Cross-lingual Alignment at Scale: Effect of learning rate

Issue:

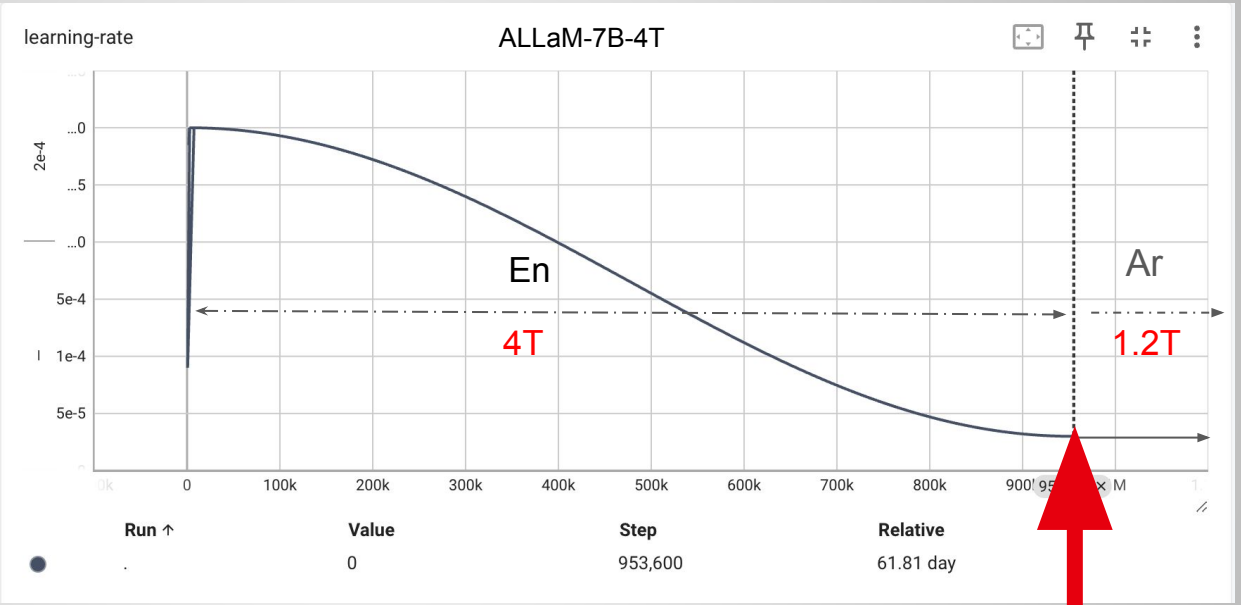
- Once the learning rate went down, it's very difficult to ramp-up learning rate without some forgetting
- hyperparameter tuning
  - 1e-3 … 3e-4
  - each for 10B tokens
- We had to train on the constant learning rate 3.0e-5



Cross-lingual Alignment at Scale: Effect of learning rate

Issue:

- Once the learning rate went down, it's very difficult to ramp-up learning rate without some forgetting
- hyperparameter tuning
  - 1e-3 … 3e-4
  - each for 10B tokens
- We had to train on the constant learning rate 3.0e-5



Data team: Comes up with new data that will increase the GDP by 10x. (:D)

Cross-lingual Alignment at Scale

How 34b is being trained:

- In previous literature, T5 [5] was trained on comparatively large learning rate (but with adafactor optimizer)
- For 34b, we train the LLM with large and constant learning rate.
- When we need a product release, we just do a pass of Warmup State Decay with GDP breaking dataset. (:D)
-

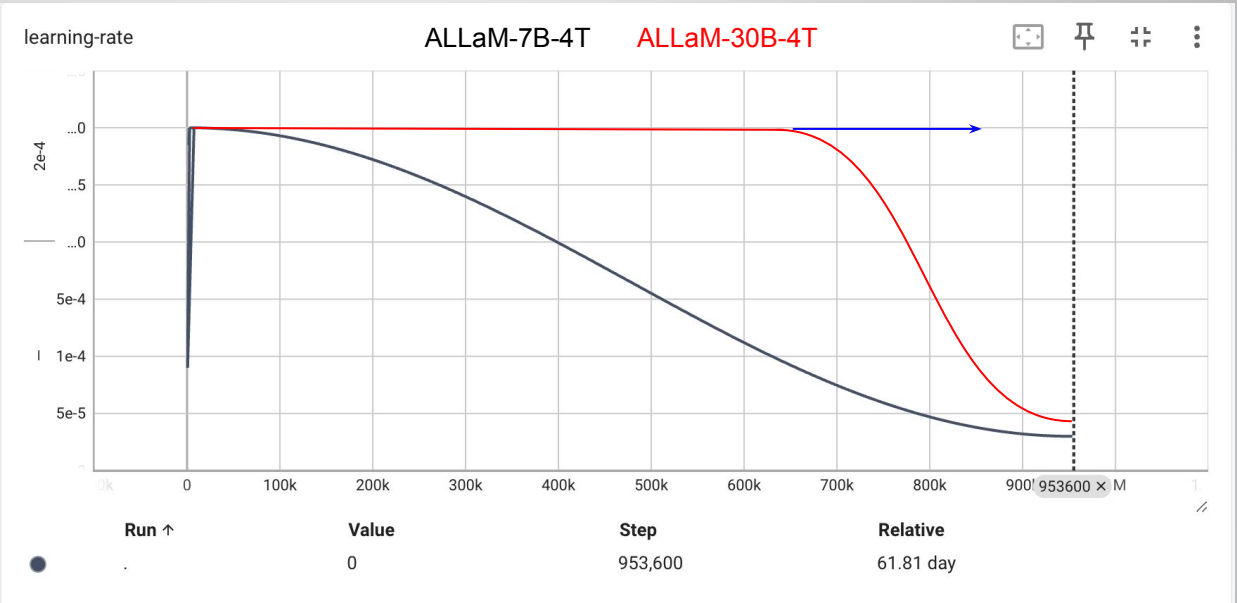
[6] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." Journal of machine learning research 21.140 (2020): 1-67.

Cross-lingual Alignment at Scale

Issue:

- If we train constant learning rate, we can go back to pretraining without any issue
- Do a WSD step training with good amount of data
- Once the release is done, we can go back to the previous checkpoint with large learning rate.

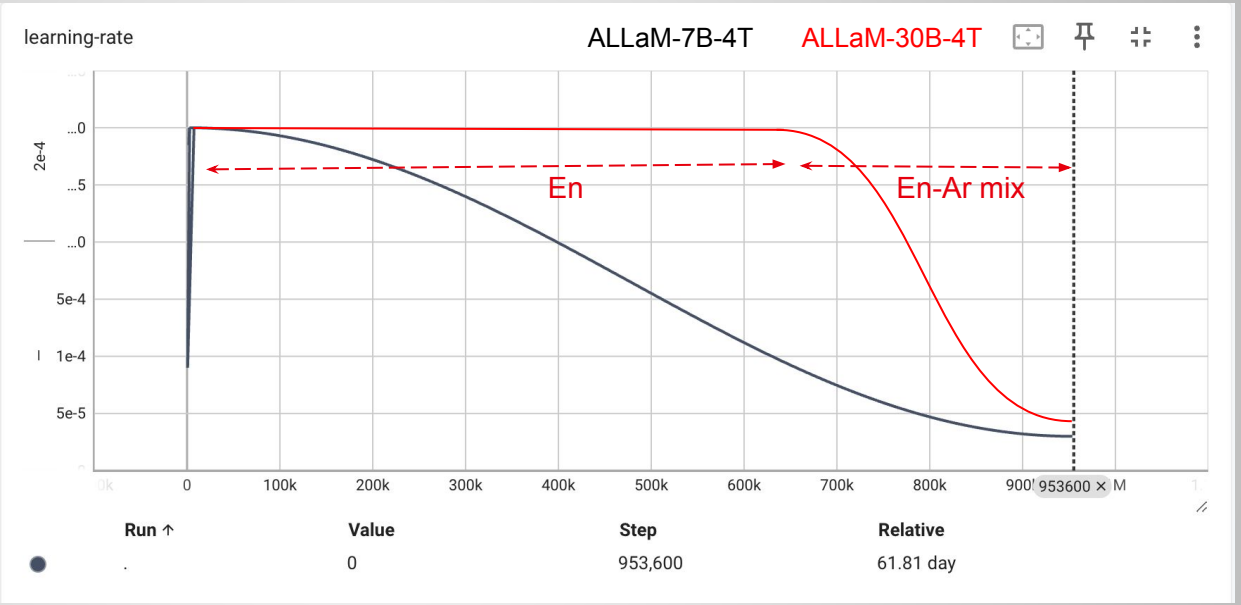
LIMITATION: We don't know how this affects the scaling laws.



Cross-lingual Alignment at Scale

Issue:

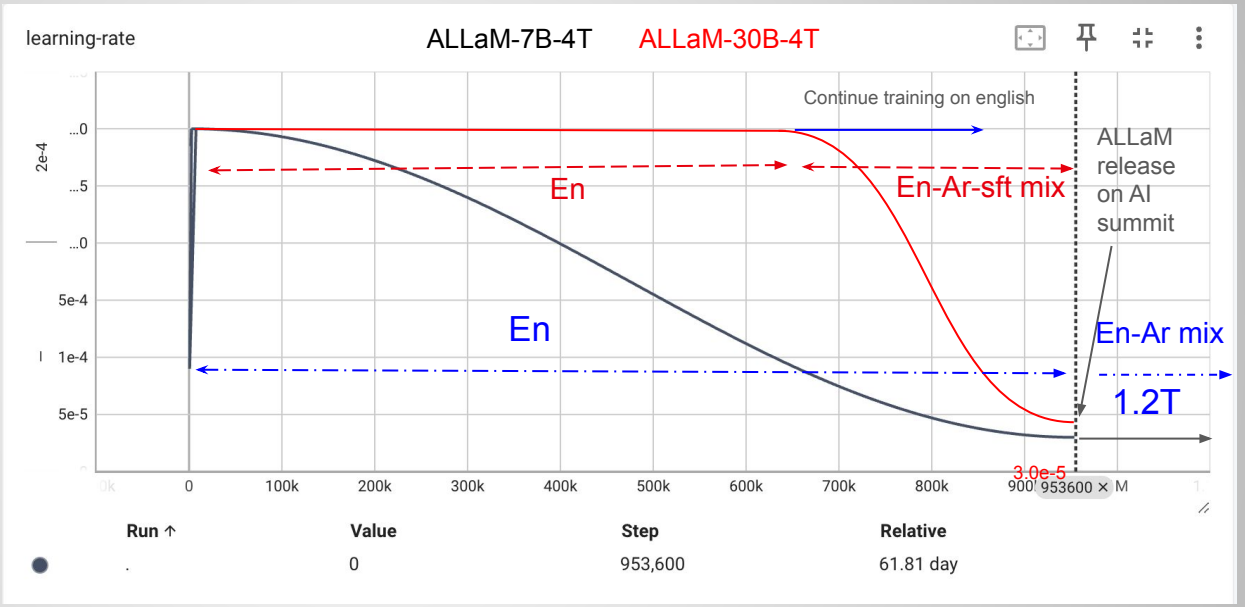
- We decided to use English Arabic dataset to use as WDS step.



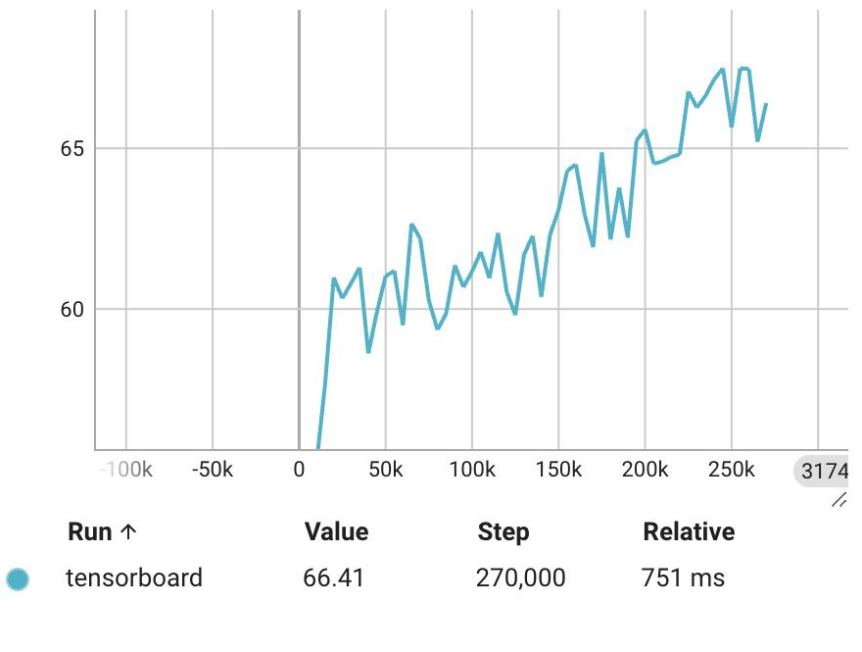
Cross-lingual Alignment at Scale

Issue:

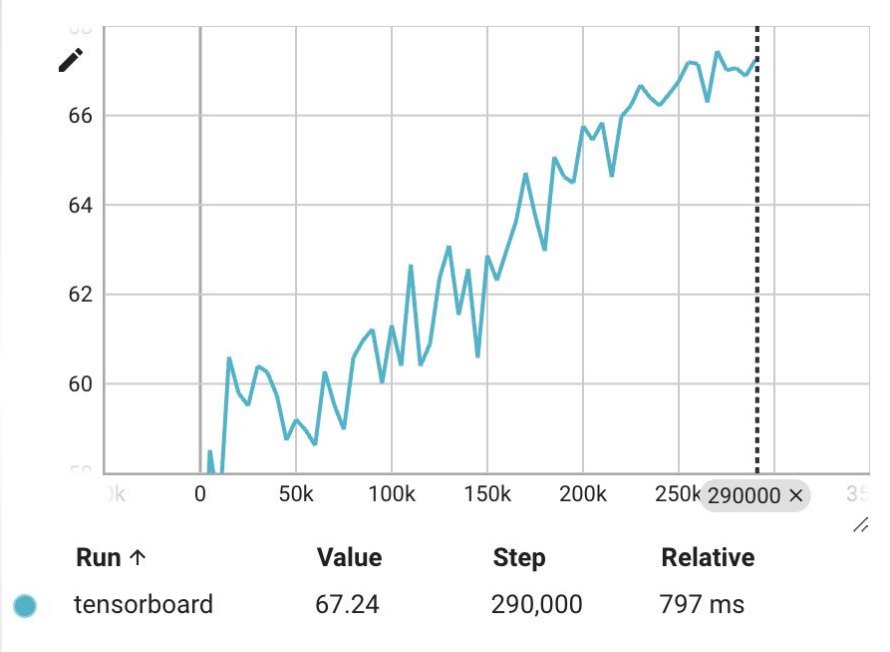
- We decided to use English Arabic dataset to use as WDS step.
- Why not add the SFT dataset.



The last stage of Pre-Training- **Midtraining**: **The model just wants to learn**



MMLU-ar



MMLU-en

Figure: 34B model pre-training.

Model Merging works, … BUT !!!

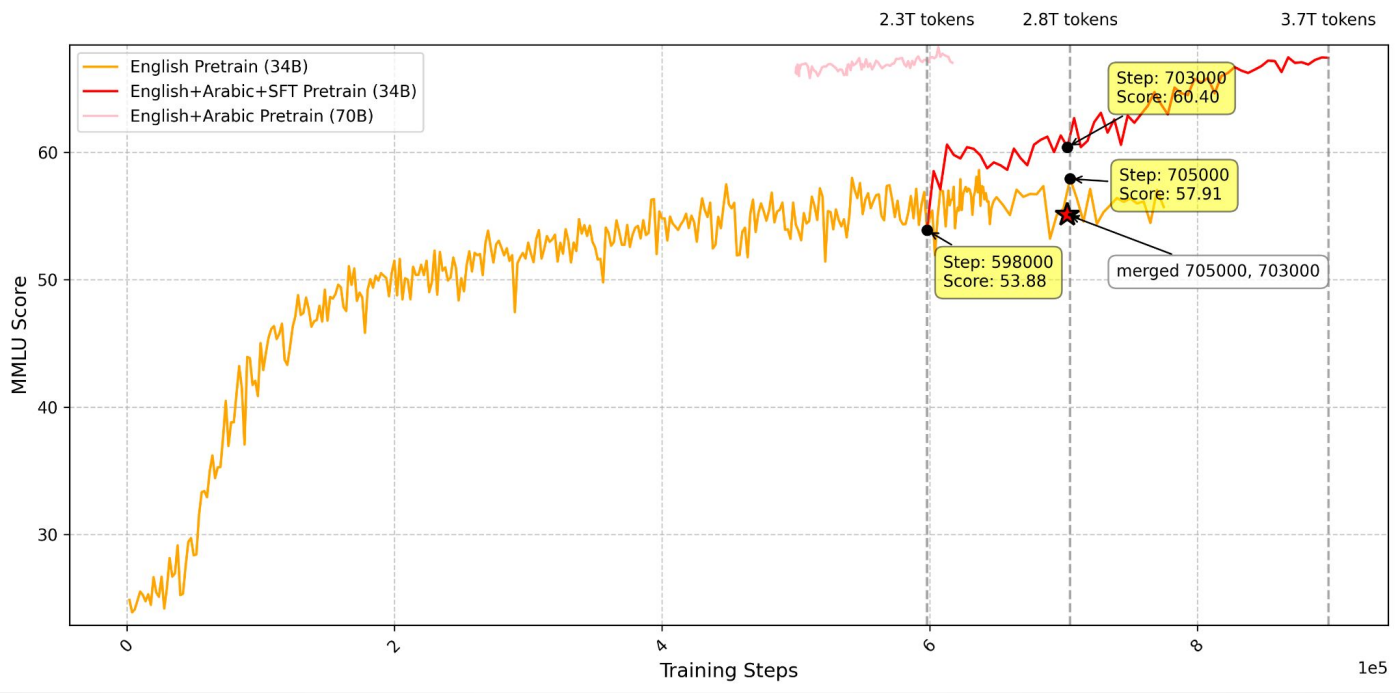
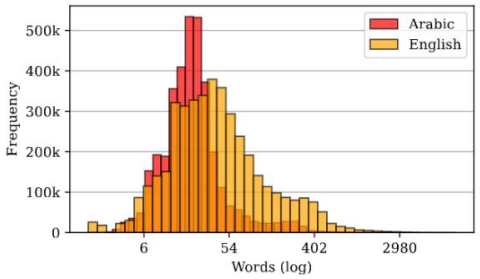
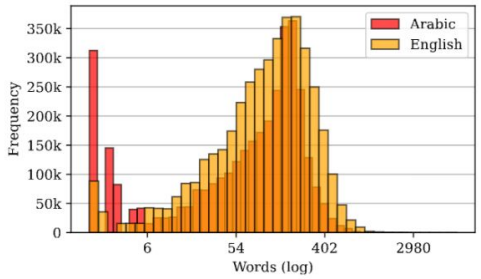


Figure: Effect of adding en+ar+sft data and model merging at scale

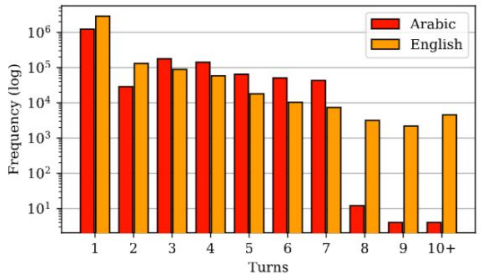
Scaling SFT data: ***Less is not More and certainly not Enough***!



(a) Prompt word count.



(b) Response word count.



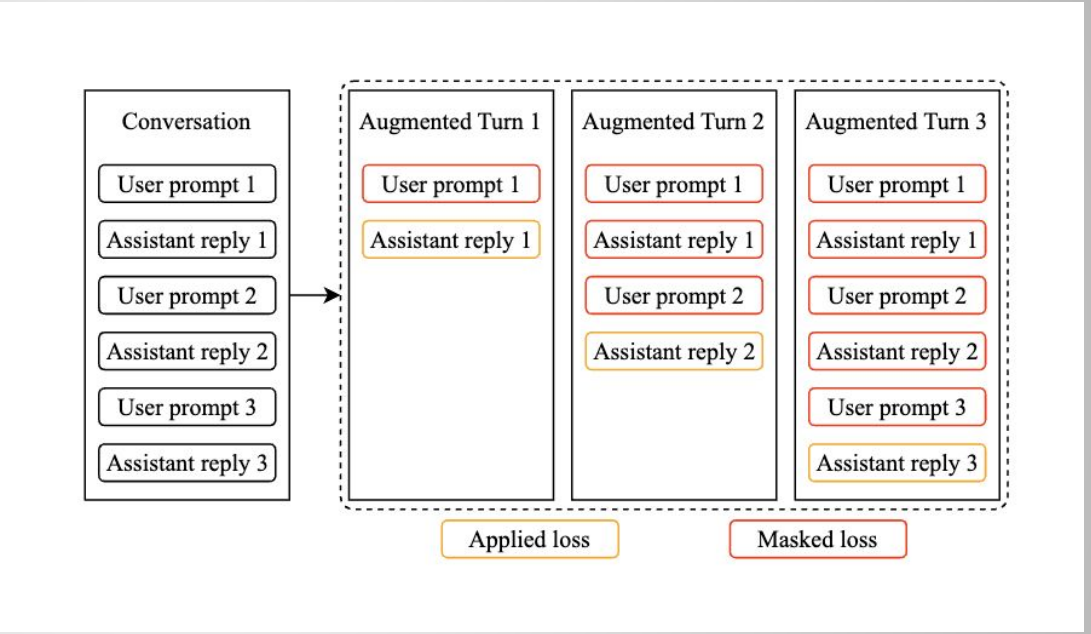
(c) Number of conversation turns.

Figure 7: Word count and turn distributions of SFT data. There are two main differences in our Arabic and English SFT datasets: shorter responses are more frequent in our Arabic SFT dataset, while our English SFT dataset contains more dialogues with more than 8 turns.

- As we increase more data, we see consistent improvement in generalization.
- Larger models are easy to generalize, align.
- Quality matters a lot vs Quantity.
- Human driven alignment.

Alignment: Multi-Turn Training

- N turn training
- Scaling SFT Data
- Direct Preference Optimization
  - Multi-Epoch Sampling from the same data.
  - Emphasize on KL penalty



Results

		araSwag	ACVA	MMLU (ar)		Exams (ar)	ETEC	araTruthfulQA	araMath
				Koto et al. (2024)	Huang et al. (2023)				
		10-shot	5-shot	0-shot	0-shot	5-shot	0-shot	0-shot	5-shot
ALLaM-Instruct	7B	**49.28**	**80.33**	**66.9**	**49.6**	**52.7**	**62.95**	**36.4**	36.5
AceGPT-Chat	7B	43.4	59.35	45.8	33.58	35.57	36.05	**37.9**	22.5
Llama 2-Chat	7B	24.44	52.46	33.33	26.45	25.33	26.69	29.9	21.5
Mistral-Instruct-v0.3	7B	30.59	60.7	44.3	34.06	31.1	34.41	30.3	26.0
Llama 3-Instruct	8B	33.99	75.21	53.98	41.49	44.32	49.42	34.0	**38.3**
ALLaM-Instruct	13B	54.77	**78.59**	**68.11**	**51.03**	**54.93**	**65.59**	**37.5**	**46.8**
Llama 2-Chat	13B	25.75	60.14	35.84	28.73	22.91	30.44	31.4	22.3
Jais-Chat	13B	**77.12**	70.68	54.8	41.43	46.93	48.68	31.6	25.3
ALLaM-Instruct	34B	59.34	**93.41**	**76.17**	**46.36**	**80.74**	**80.22**	**38.17**	**74.6**
Jais-Chat-v3	30B	**88.37**	70.05	62.37	30.15	51.21	38.53	37.3	32.5
ALLaM-Instruct	70B	57.91	79.01	**75.92**	**62.23**	58.47	**78.38**	38.4	56.8
Llama 2-Chat	70B	30.72	59.49	40.77	32.86	28.68	30.6	32.3	25.5
Llama 3-Instruct	70B	45.75	**80.26**	36.27	60.11	**58.47**	71.41	37.7	**59.70**

Figure: Arabic benchmark results for instruction tuned models. Follow Table 10 for detailed results.

Results

		AGIEval	MMLU Average	MMLU-Pro	Ethics	TruthfulQA	ARC Challenge	HellaSwag	MixEval Hard	MixEval Standard
		0-shot	0-shot	CoT 5-shot	0-shot	0-shot	0-shot	0-shot	5/0-shot (base/ft)	5/0-shot (base/ft)
ALLaM-Instruct	7B	**47.09**	58.31	27.78	69.8	42.11	51.45	75.2	28.9	67.6
AceGPT-Chat	7B	26.33	44.53	—	53.38	49.34	42.32	70.92	—	—
Llama 2-Chat	7B	35.55	46.4	22.87	58.88	45.32	44.28	75.52	30.8	61.7
Mistral-Instruct-v0.3	7B	42.22	59.75	36.33	73.59	59.65	58.7	82.88	36.2	70.0
Llama 3-Instruct	8B	44.35	**63.82**	**41.32**	**68.07**	**51.72**	**56.83**	**75.81**	**45.6**	**75.0**
ALLaM-Instruct	13B	**48.42**	**61.8**	**34.05**	**76.47**	**57.69**	**55.89**	**81.14**	**37.2**	**72.8**
Llama 2-Chat	13B	37.73	53.3	27.19	70.52	43.95	50.17	79.66	—	—
Jais-Chat	13B	31.45	49.46	—	64.92	39.66	46.84	77.6	—	—
ALLaM-Instruct	**34B**	**52.47**	**71.24**	**43.61**	**72.84**	**56.27**	**60.15**	**81.25**	—	—
Jais-Chat-v3	30B	36.78	57.57	26.45	68.03	42.34	51.02	78.91	—	—
ALLaM-Instruct	70B	**65.67**	75.43	48.61	76.16	58.78	59.56	**84.97**	51.60	83.5
Llama 2-Chat	70B	46.0	61.15	35.16	68.5	52.77	54.27	82.14	38.0	74.6
Llama 3-Instruct	70B	63.78	**78.38**	**59.52**	**77.09**	**61.79**	**64.33**	82.49	**55.90**	**84.00**

Figure: English benchmark results for instruction tuned models. Follow Table 11 for detailed results.

Results

Model	English			Arabic		
	Avg.	Turn 1	Turn 2	Avg.	Turn 1	Turn 2
AceGPT 13B-chat	5.44	6.76	4.12	6.33	7.01	5.64
ALLaM 13B Instruct	7.34	7.67	7.01	7.57	7.9	7.23
ALLaM 70B Instruct	7.44	**7.91**	6.96	**8.19**	**8.4**	**7.97**
Jais 13B Chat	4.18	4.39	3.96	4.72	5.07	4.36
Jais 30B Chat v1	3.89	4.13	3.64	3.54	4.13	2.95
Jais 30B Chat v3	5.86	6.25	5.47	6.28	6.78	5.78
Cohere Command R+	7.41	7.63	7.18	7.97	8.28	7.65
Cohere Command R	6.99	7.19	6.79	7.47	7.82	7.12
DBRX Instruct	7.16	7.33	6.98	7.83	8.19	7.46
GPT 3.5 Turbo	**7.55**	7.79	**7.31**	8.12	8.39	7.84

Figure: MT bench score

Results

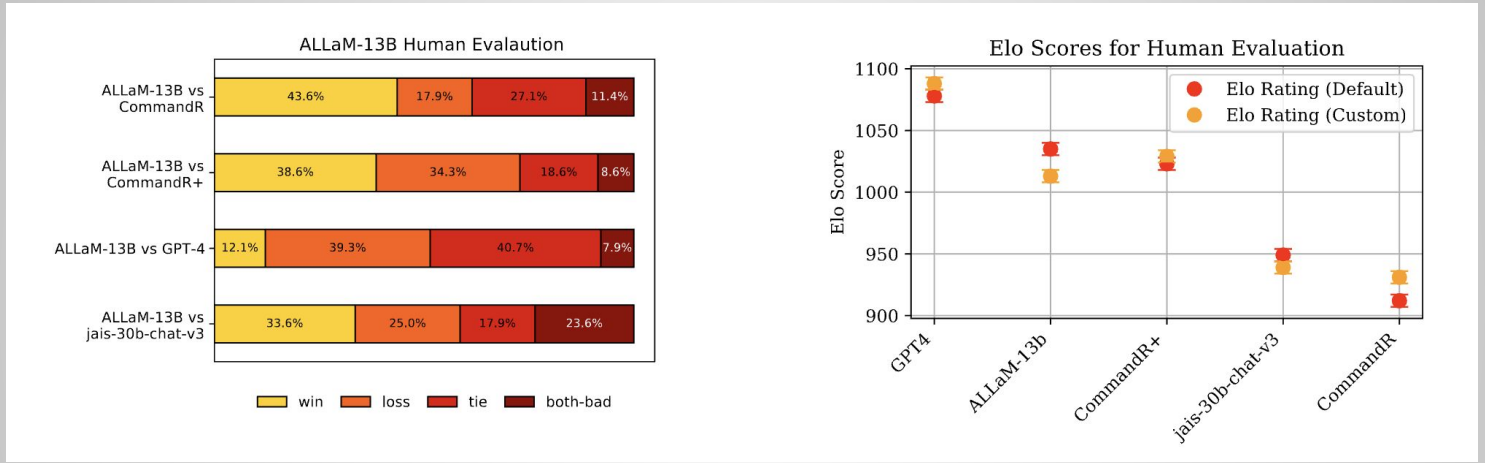


Figure: (Left) Pairwise win rates as judged by human evaluators. ALLaM-13B wins against many much larger models.

Figure: ELO scores from human evaluator preferences. ALLaM is tied with Command-R+ and lags only behind GPT-4.

Pathways towards **Super-Intelligence** (KNOWLEDGE EXPLOSION)