📈 ARCHITECTURE & SCALING SUB-WORKING GROUP

YOU ONLY TRAIN ONCE: MAKING ARCHITECTURAL DECISIONS FOR A >100B MODEL

Big Science Episode #2 — INLG, 2021/09/20

Iz Beltagy
Ai2

Big Science

Julien Launay
LightOn

📈 The Scaling and Architecture Sub-Working Group

What?

Draft and validate an architecture & training setup to get the best out of our GPU budget.

How?

By establishing principled baselines, carefully evaluating novel modelling choices, and studying the scaling of candidate architectures.

Constraints.

💯 **proven**

no unnecessary risks

📈 **scalable** 🔥 **efficient**

final run: >200B param., 4MGPUh

🌍 **multilingual**

🧠 **emergent**

few-shot, prompt tuning, etc.

Main unknowns in 🌸 Big Science

Scale

Very few models have been trained in the 100-200B range.

GPT-3 (English, OpenAI), Jurassic-1 (English, A21),
HyperClova (Korean, Naver), PanGu-Alpha (Chinese, Huawei).

🤗 with engineering working group.

Multilinguality

Limited knowledge on extreme-scale generative multilingual models.

Closest comparison: mT5, 100 languages, 11B parameters. No large generative-only model.

Can we avoid the *curse of multilinguality*?

Severely underperforming monolingual counterparts.

🤗 with multilingual working group.

Architecture

Bridge the LM and encoder-decoder performance gap with prefix LM.

How to validate prefix LM at scale?

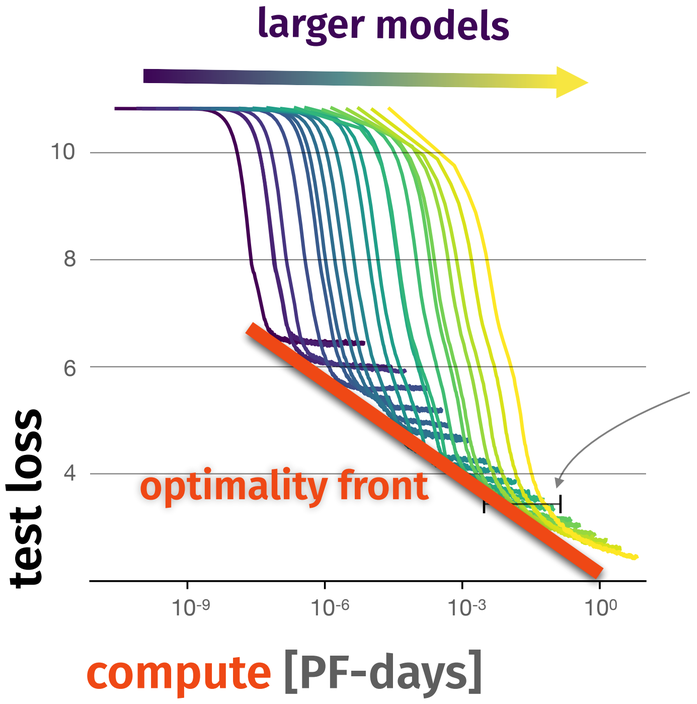
Evaluations and metrics to benchmark architectures

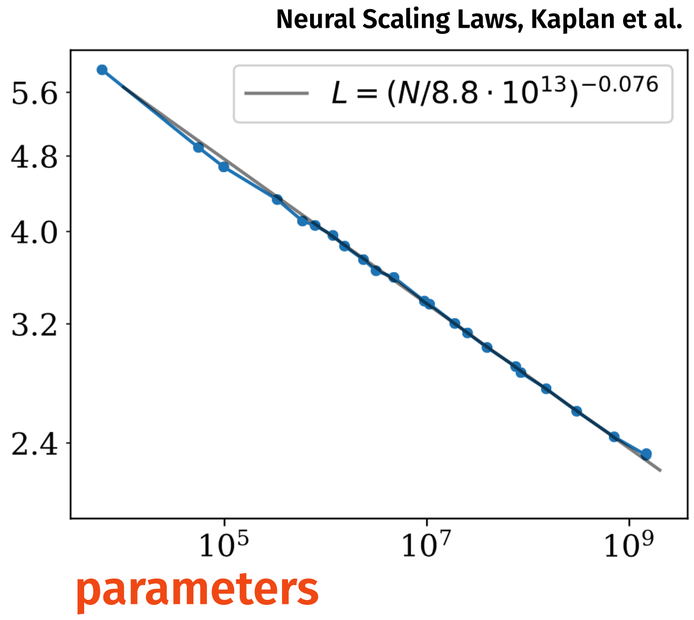
👀 **Usual and simple metrics: validation loss, training time/throughput, etc.**

efficiency & stability are key metrics at the 100B+ scale!

📈 **Empirical backing of scaling laws to evaluate scaling:**

performance is predictable using simple power laws





still, some behaviours are scale-emergent ⟶ train as large as possible, ~1B scale at least

Evaluations and metrics to benchmark architectures

👩‍🏫 Zero/few-shot performance evaluation on a large range of datasets.

currently using Eleuther AI evaluation harness for English baselines.

🤗 with evaluation group ⟶ multilingual evaluation, etc.

🚨 Big unknown: how will final 200B model be used by the community? 🚨

🤩 Weights offloading/streaming make inference "accessible"...

ZeRO-infinity

but still very expensive to run in practice! 💸

🔎 Currently, OpenAI/A21/Cohere ⟶ hosted API with a text/log-prob interface.

fine-tuning only offered for small models.

🧠 Other approaches: efficient fine-tuning, adapter, prompt tuning, etc.

keep emergent possibilities open!

Unknown #1: Scale

📈 **100B+ scale is unforgiving: we need excellent tooling, scalable architecture, etc.**

every FLOP counts!

💥 **"Unstable" behaviour in training at scale, not fully explained.**

numerical instabilities: float16, etc. → can be avoided with bfloat16 on modern hardware (TPUs/A100s)

data-related instabilities? → see work on curriculum learning

diagnostic tools? → gradient noise scale, weightwatcher, etc.

🤖 **Engineering working group: "big" exploratory runs at the >10B scale.**

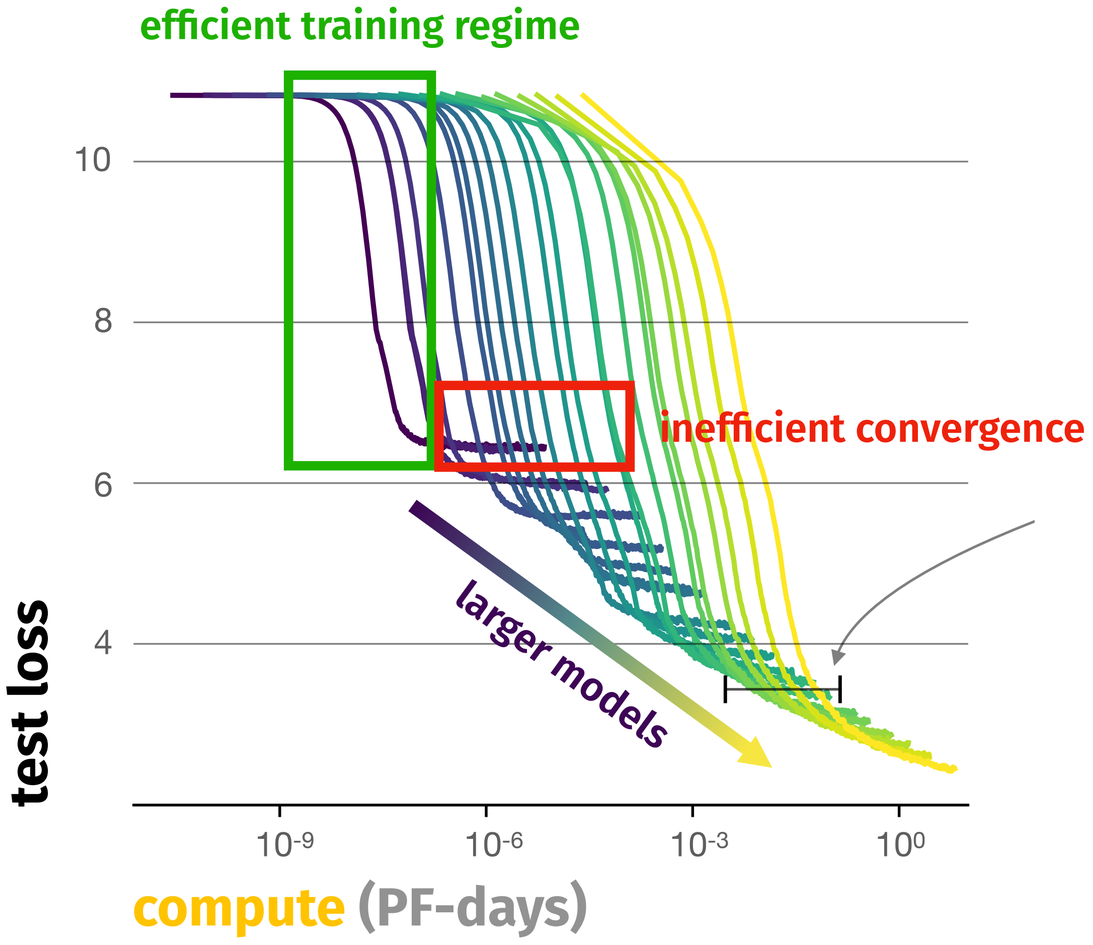
training #1 (13B English-only) complete, now looking at 13B multilingual for training #2.

one lesson already: dataset matters *a lot* for end-task performance!

Training setup: at scale, training to convergence vs optimality

🧠 **Don't train to convergence, but to optimality for efficiency in final run.**

training budget: 200B, 4,400 PF-days (~4 MV100h@25 TFLOPs) to optimality, 30,000 PF-days (~30 MV100h) to conv.

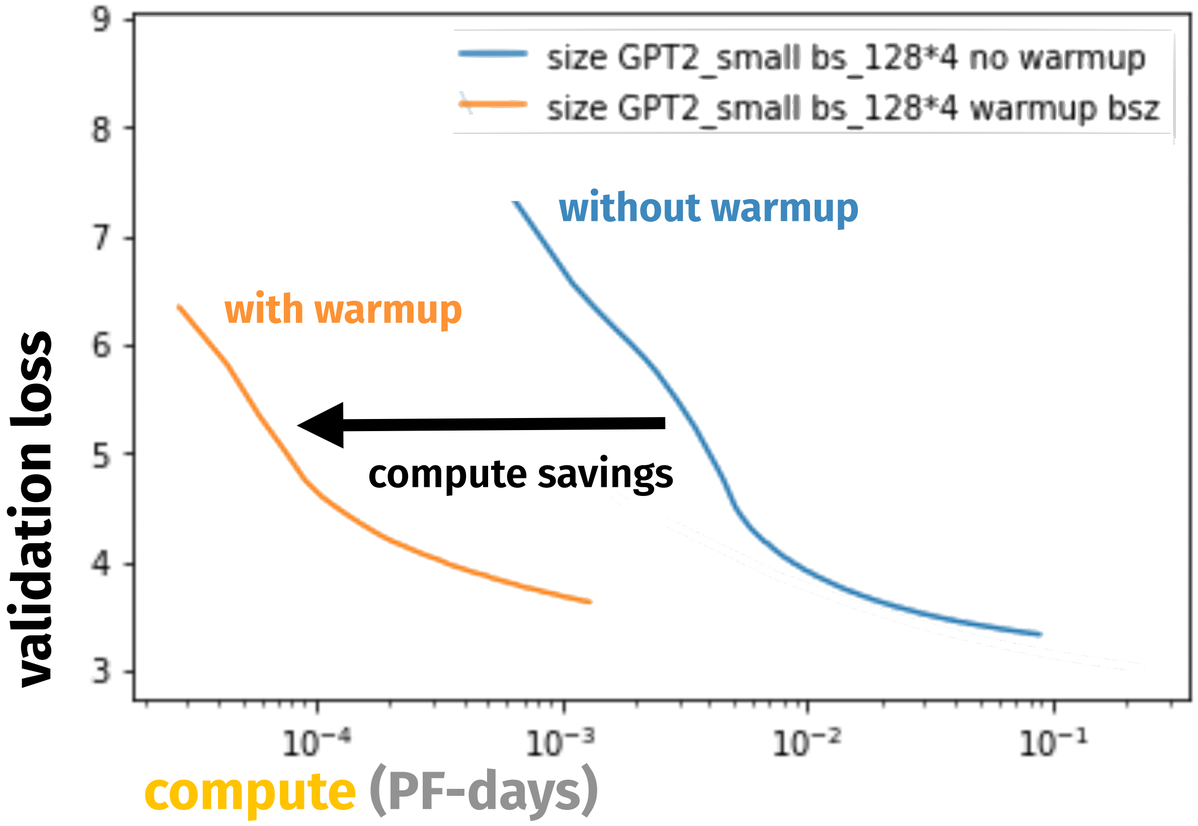


Neural Scaling Laws, Kaplan et al.

Batch size warmup saves compute

🏋️ Batch size warmup: start with a small batch size, then linearly increase to max batch size.

🤓 Intuition: gradient noise is high early in training, so large batch size is wasteful.



Scaling laws as a diagnostic tool

🤔 Big Science training #1 (13B, English-only): disappointing few-shot performance.

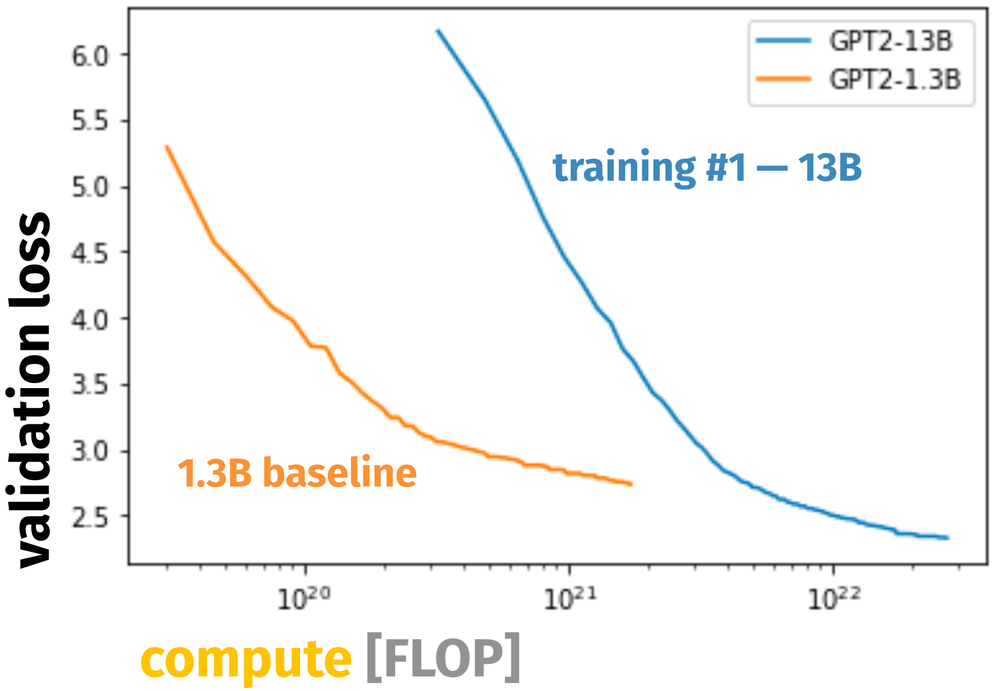
Results from EAI harness obtained by Stella Biderman.

5.02	65%	54%
lambada ppl.	winogrande acc.	hellaswag acc.

more in line with a 2.7-6B model!

is this a data (OSCAR) problem?

or a setup problem?



Unknown #2: Multilinguality

🤗 **Build a model that is valuable to the community at large.**

languages selection, data collection, release licenses, etc. → many other WGs in Big Science!

🤔 **Under-explored at scale, with *curse of multilinguality* problem.**

if multilingual model severely underperforms monolingual counterparts, not that interesting!

no large-scale generative multilingual model exists... → ⚠️ very sensitive to data,
no high-quality multilingual dataset!

100B English tokens vs 100B multilingual tokens, what's the gap?

😓 **Evaluation of multilingual models is more challenging.**

less big and "wide" benchmarks than in English for low-ressources languages.

Tackling multilinguality under the angle of scaling laws

🌍 **Can we establish multilingual scaling laws?**

quantify how languages scale differently...

quantify benefits from one language to another, like has been done for multimodal setups...

connect to fundamental linguistics works and validate findings

💪 **Can we use this law for more principled multilingual training.**

inform sampling strategy/scaling of gradients, etc.

We will be answering this questions soon 😊

Unknown #3: Architecture

GPT-3 as our base architecture, however...

⚠️ From the T5 paper: performance of autoregressive LM is lower than encoder-decoder

T5, Raffel et al.

Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	Denoising	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	Denoising	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	Denoising	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	Denoising	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39

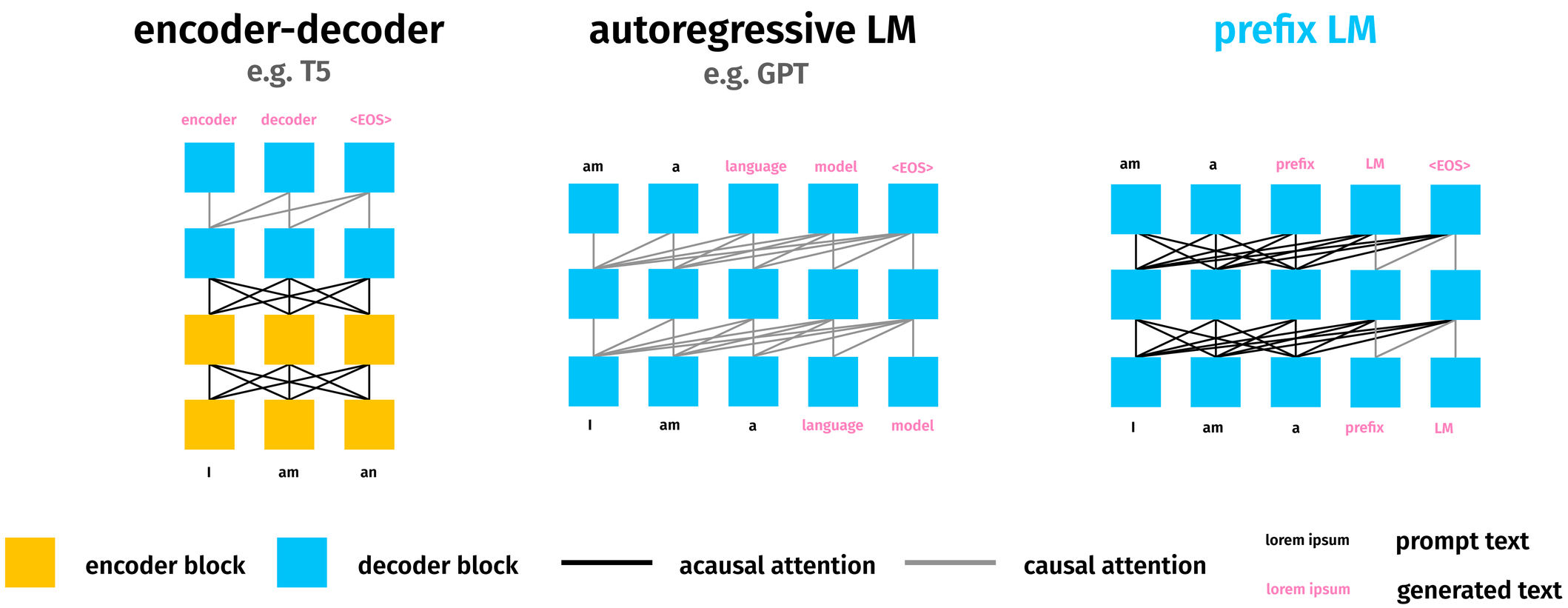
Can we use a prefix LM model to bridge the gap?

❓ Other architectural choices: embeddings, activation functions, etc

rotary, ALiBi

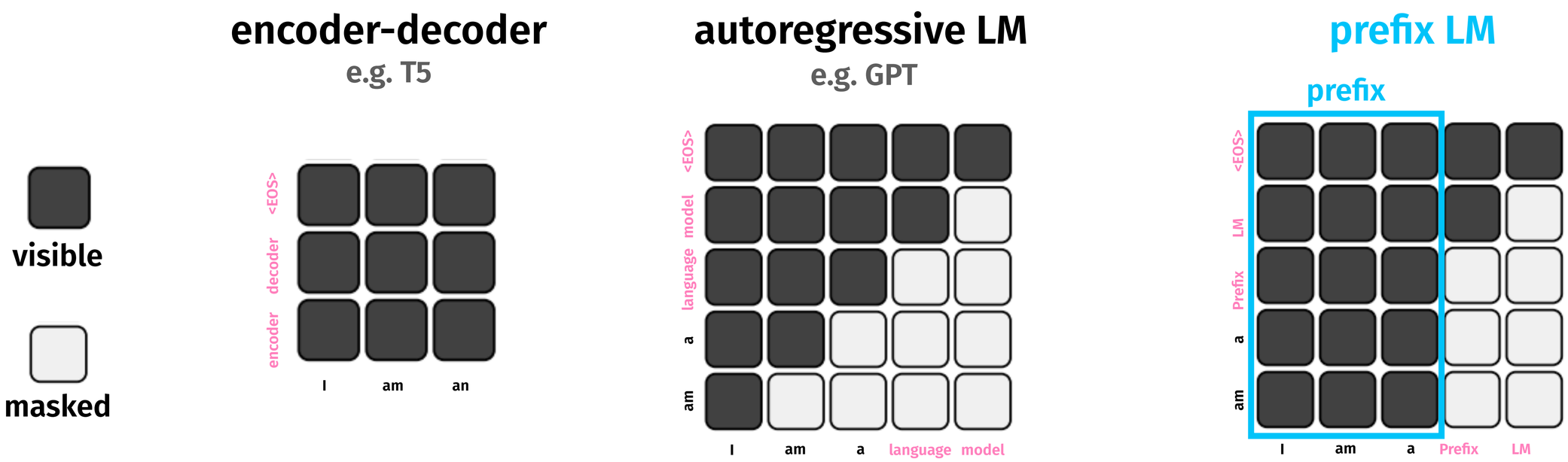
GeLU-GLU, squared ReLU

Bridging the performance gap with prefix language modelling



⚠️ Prefix LM: same architecture as autoregressive LM, but with a different attention pattern.

Bridging the performance gap with prefix language modelling



🧪 Intuition: tokens in the prefix/prompt don't have restricted view, thus better representation.

🦾 As per T5, could bridge encoder-decoder/LM gap, but never demonstrated at scale nor for few-shot!

train with a randomly selected prefix during training, then prefix is prompt at inference time.

Megatron+DeepSpeed implementation ready, 1.3B results soon.

Choosing a positional embedding: state-of-the-art

📅 Better embeddings have been a hot topic: rotary, ALiBi, etc.

different metrics of importance: speed, stability, modeling loss, extrapolation.

🔄 Rotary embeddings

clear performance advantage, very small cost in speed.

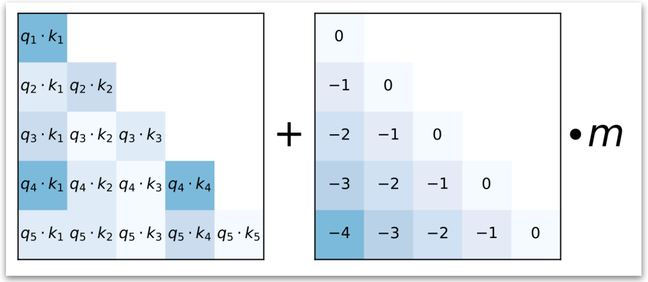
how it works: adds positional information to every layer, at the keys/queries.

✨ ALiBi: newest embedding, with extrapolation capabilities.

Extrapolation: pretrain on short sequences then evaluate on longer ones
potentially opens the door to training with a smaller context size!

very simple and fast, performance on large models to be confirmed.
how it works: simple additive bias to attention scores

ALiBi, Press et al.

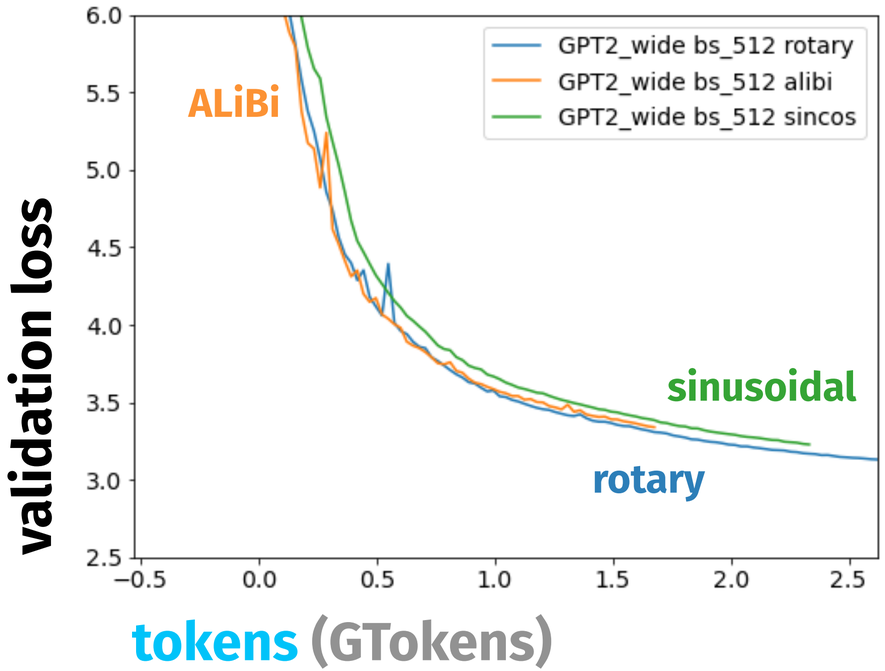


Choosing a positional embedding: first experiments

👍 **Rotary and ALiBi consistently outperforms sinusoidal embeddings**

why? they inject position information in each self-attention layer, not just in input embeddings;

they use relative position information, so the model can't overfit certain locations.



Limitations of evaluation so far:

medium model (350M) only, move to 1.3B

LM loss only, should evaluate few-shot and more

Where we are and where are we going

✅

done

Implementation of different candidate architectures (mostly);

Preprocessing multilingual training data;

English-only baseline 1.3B run;

English-only evaluation benchmark.

⏭️

next steps

Debug training #1 (13B run) and understand few-shot performance;

Evaluate English-only baseline on downstream tasks;

Train and evaluate multilingual 1.3B baseline;

Train and evaluate 1.3B ALiBi, rotary, and prefix LM.

🤗 Joining and contributing!

🌸 Join Big Science: https://bigscience.huggingface.co/ and sign-up for modeling group.

🐙 GitHub: https://github.com/bigscience-workshop/Megatron-DeepSpeed/issues

📅 Weekly meetings: Wednesday 8am PT, 5pm CEST

😍 Contributors 😍

Teven Le Scao

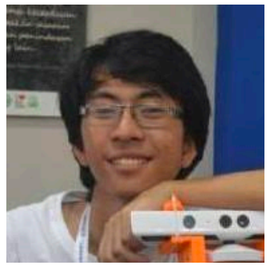
Sheng Shen

Thomas Wang

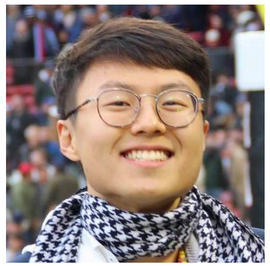
Ofir Press

Stella Biderman

M Saiful Bari (Maruf)



Lintang Sutawika



Jake Tae

Huu Nguyen