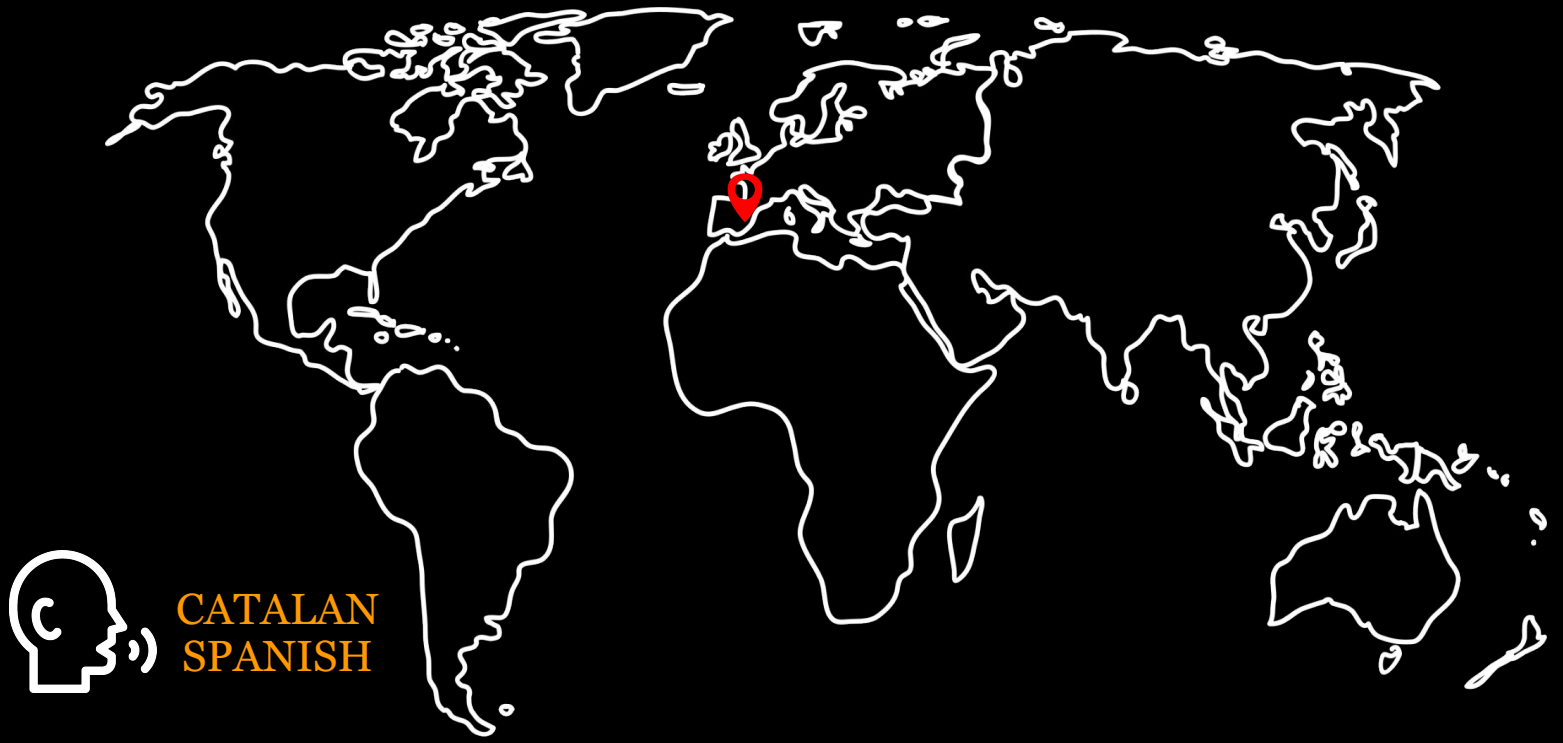
FLOR

Severino Da Dalt, Joan Llop, Irene Baucells, Marc Pàmies, Yishi Xu,
Aitor Gonzalez-Agirre, Marta Villegas

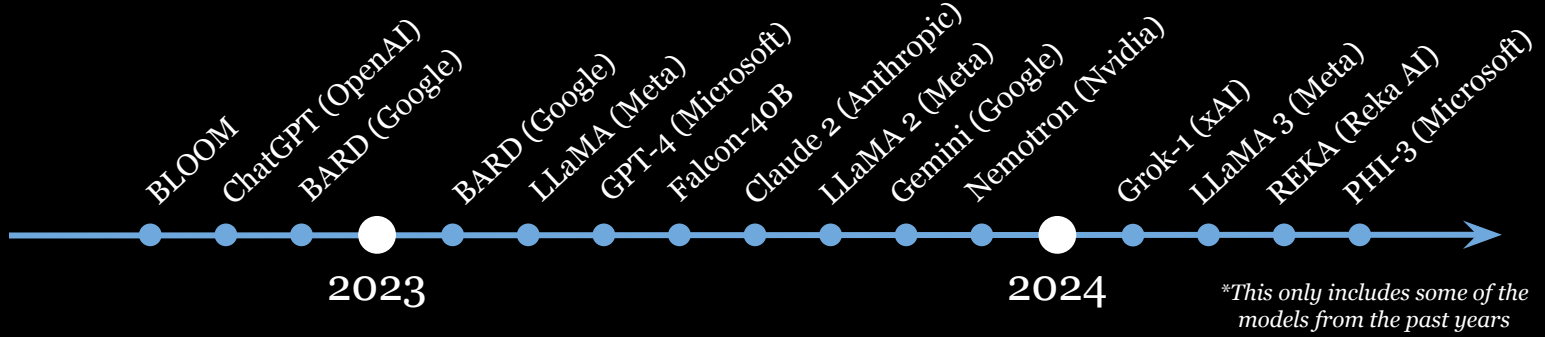
On the Effectiveness of Language Adaptation

Context



CATALAN
SPANISH

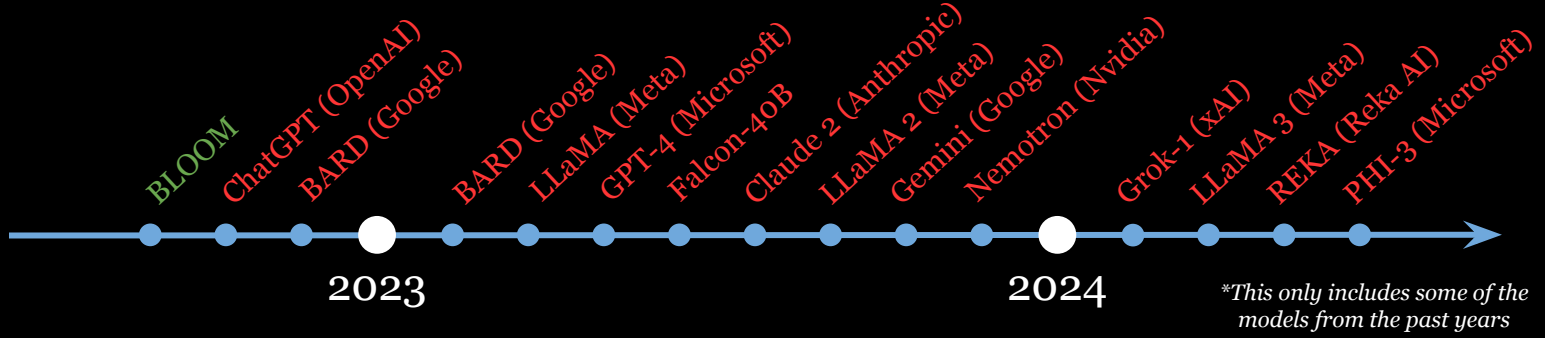
Context


BLOOM
ChatGPT (OpenAI)
BARD (Google)
BARD (Google)
LLaMA (Meta)
GPT-4 (Microsoft)
Falcon-40B
Claude 2 (Anthropic)
LLaMA 2 (Meta)
Gemini (Google)
Nemotron (Nvidia)
Grok-1 (xAI)
LLaMA 3 (Meta)
REKA (Reka AI)
PHI-3 (Microsoft)
2023
2024
*This only includes some of the models from the past years

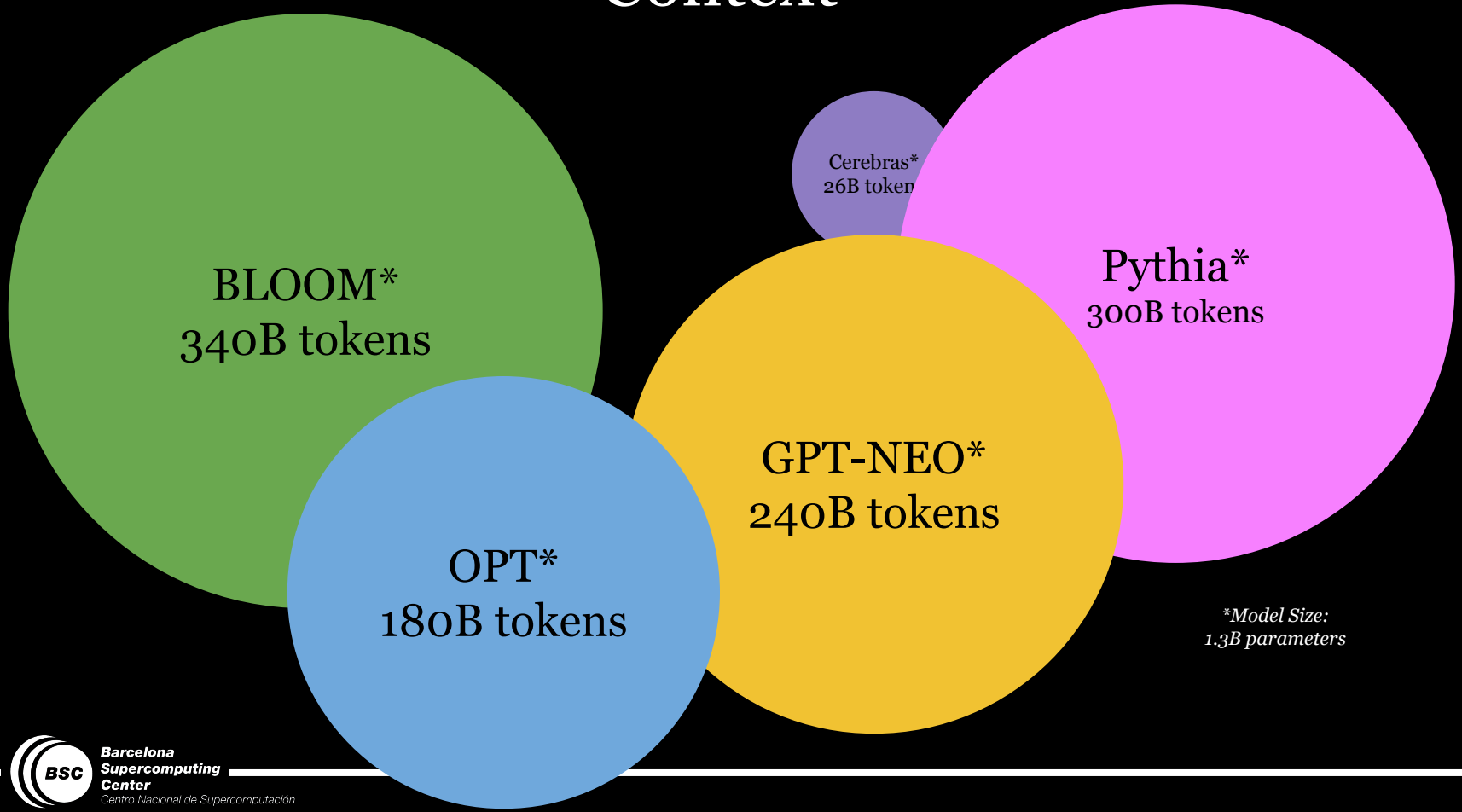
Context

Open and equal access

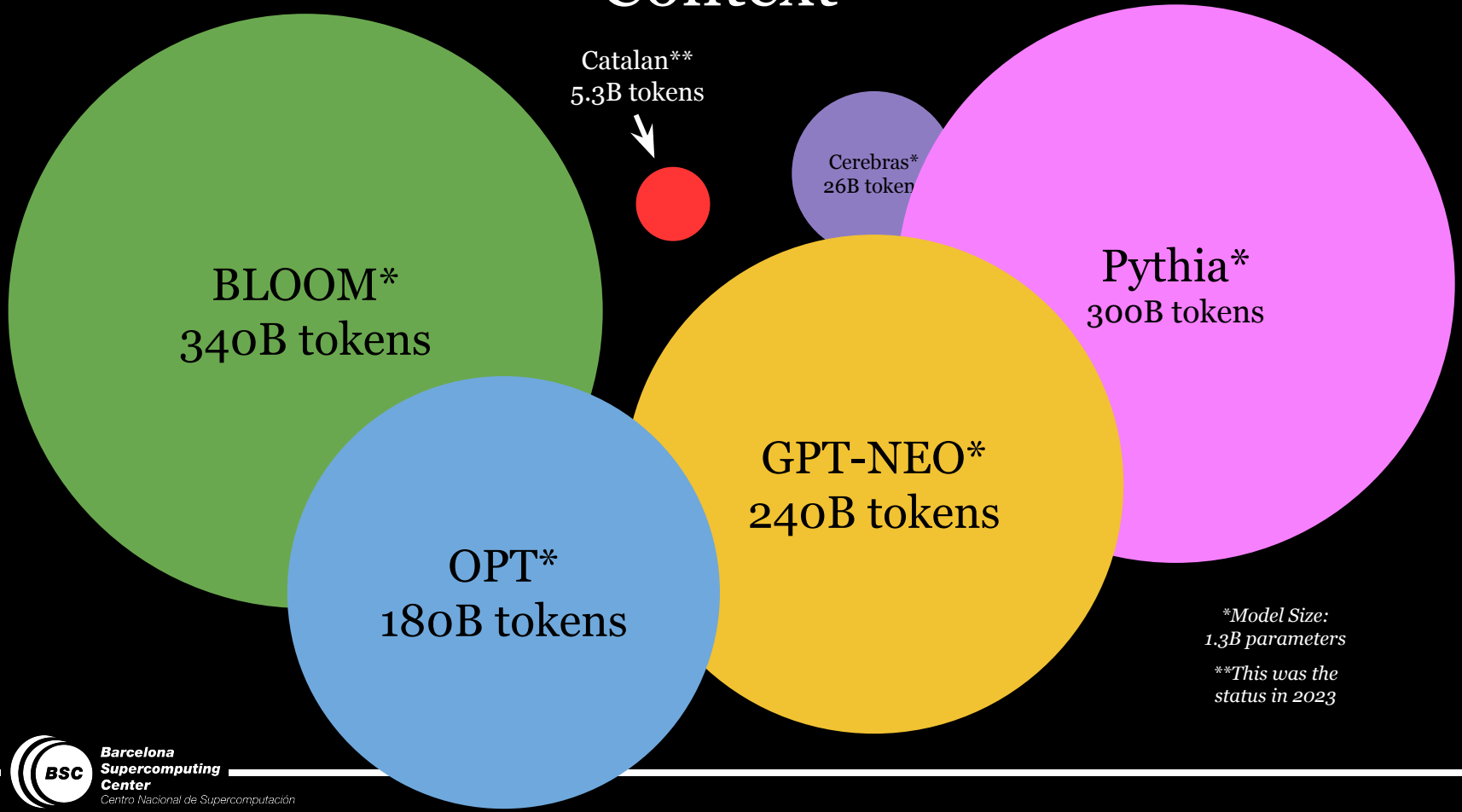
Target Language: Catalan



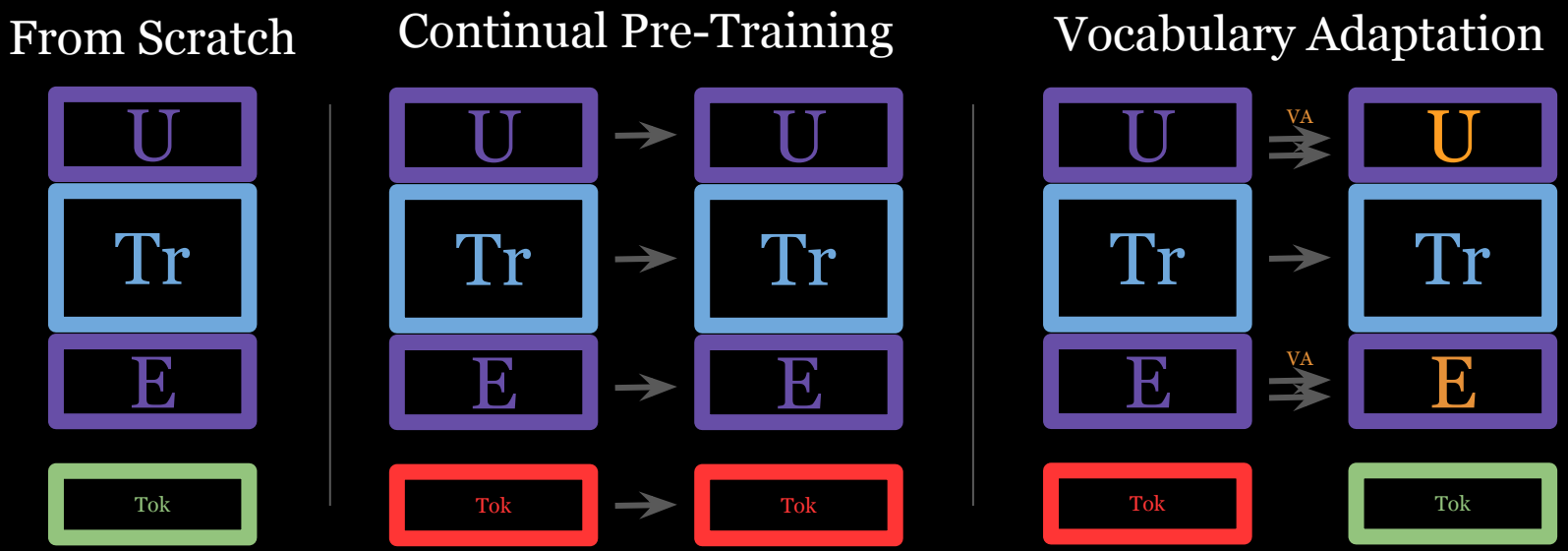
Context


BLOOM*
340B tokens
Cerebras*
26B token
Pythia*
300B tokens
GPT-NEO*
240B tokens
OPT*
180B tokens
*Model Size:
1.3B parameters

Context

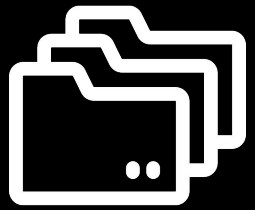

Catalan**
5.3B tokens
Cerebras*
26B token
BLOOM*
340B tokens
Pythia*
300B tokens
GPT-NEO*
240B tokens
OPT*
180B tokens
*Model Size:
1.3B parameters
**This was the
status in 2023

Best Strategy?



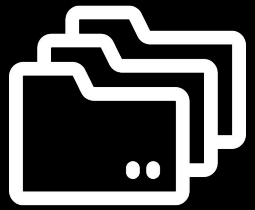
Contributions

Contributions

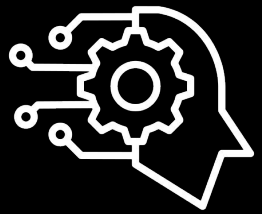


Training dataset (ca, es, en)

Contributions

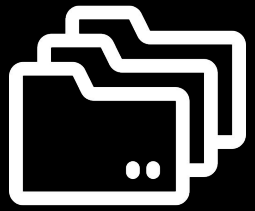


Training dataset (ca, es, en)

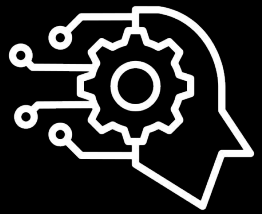


FLOR-760M and FLOR-1.3B, two autoregressive language models that achieve state-of-the-art results in several Catalan and Spanish downstream tasks.

Contributions



Training dataset (ca, es, en)



FLOR-760M and FLOR-1.3B, two autoregressive language models that achieve state-of-the-art results in several Catalan and Spanish downstream tasks.

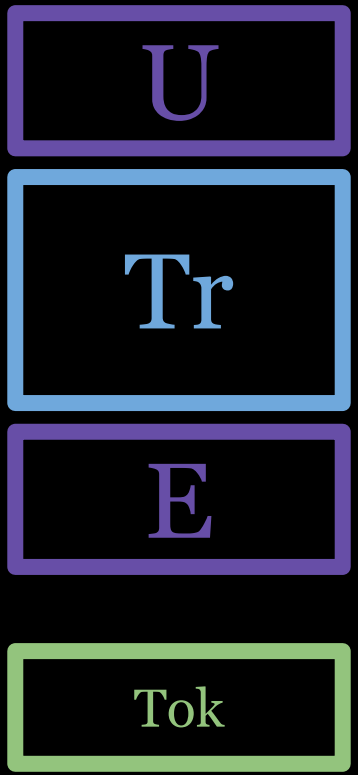
A novel evaluation benchmark for Catalan and Spanish decoder-only models.

Training Data

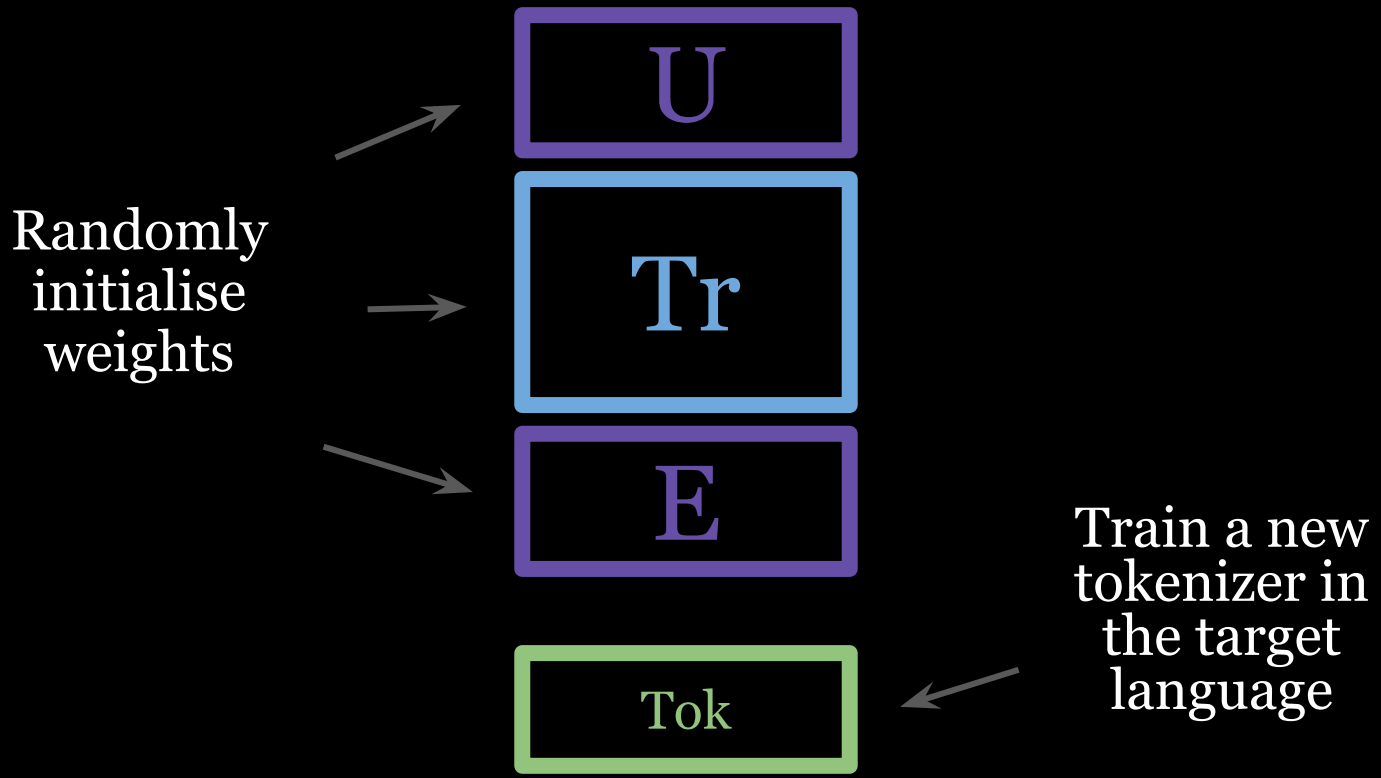
Dataset	Lang.	Epochs	Tokens (M)
Wikipedia	CA	3.5	1127.08
C4_ca	CA	2.1	8381.54
Biomedical	CA	1.4	23.33
VilaWeb	CA	2.1	149.29
CaWaC	CA	2.1	171.41
Racó - Notícies	CA	2.1	50.89
Racó - Fòrums	CA	2.1	989.81
Wikipedia	ES	1.4	1371.43
C4_es	ES	0.1	7805.53
Biomedical	ES	0.7	449.85
Legal	ES	0.7	984.37
Gutenberg	ES	0.7	52.57
Wikipedia	EN	1.4	4290.56
Total			25847.66

Training Strategy

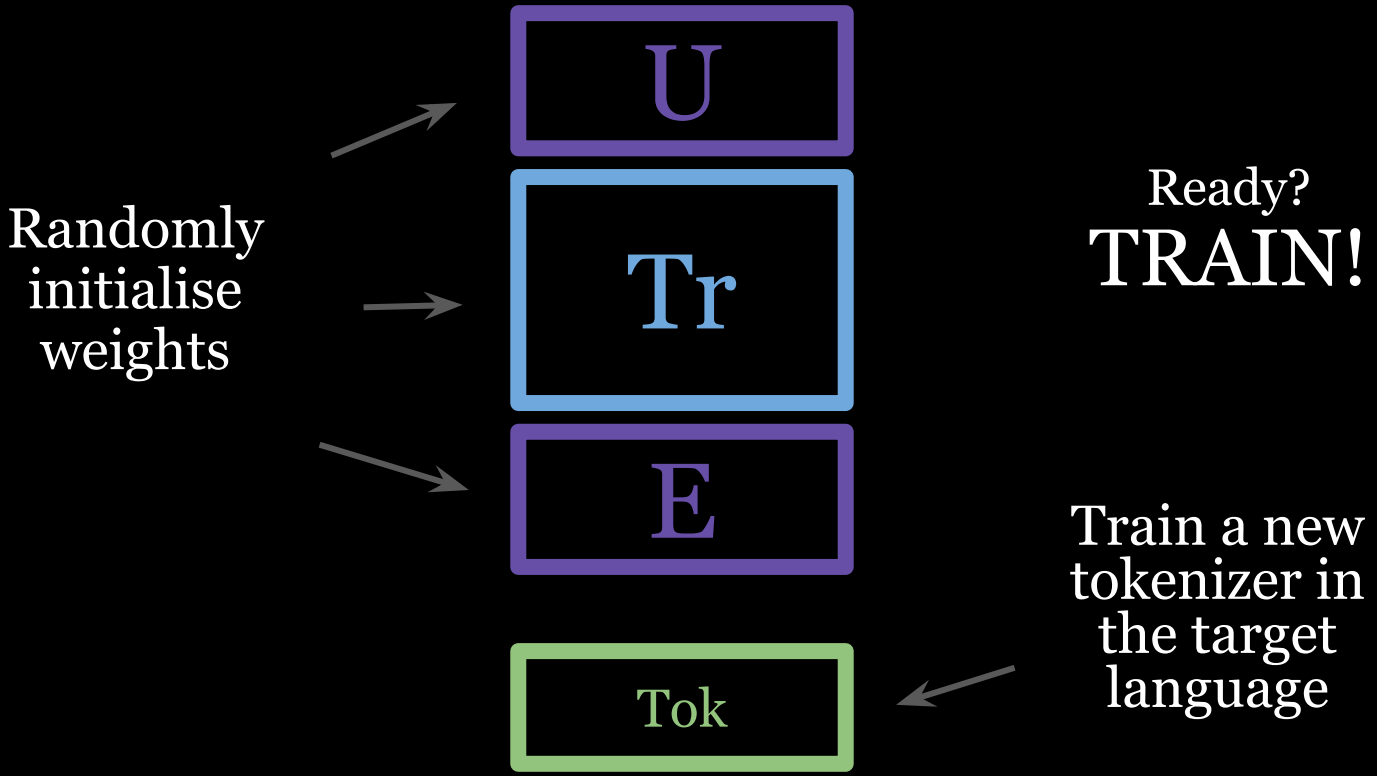
From Scratch



From Scratch



From Scratch



From Scratch

GitHub: Russian GPT-3 models
By Anton Emelyanov et al.

PAGnol: an extra-large French generative model
By Julien Launay et al.

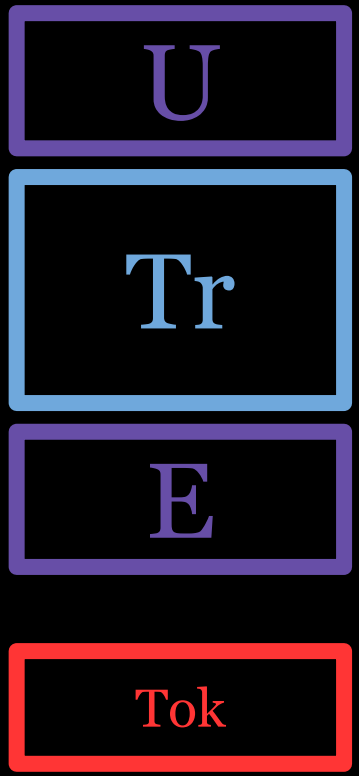
AraGPT2: Pre-trained transformer for Arabic language generation
By Wissam Antoun et al.

Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation
By Wei Zeng et al.

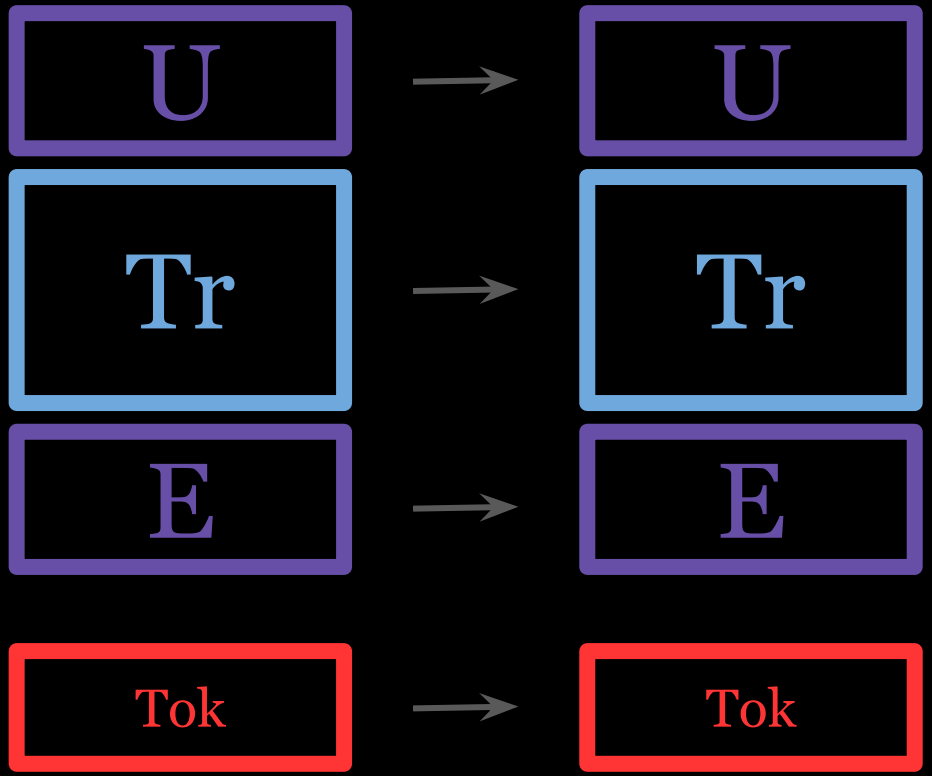
German GPT2
By Stefan Schweter

Language Adaptation

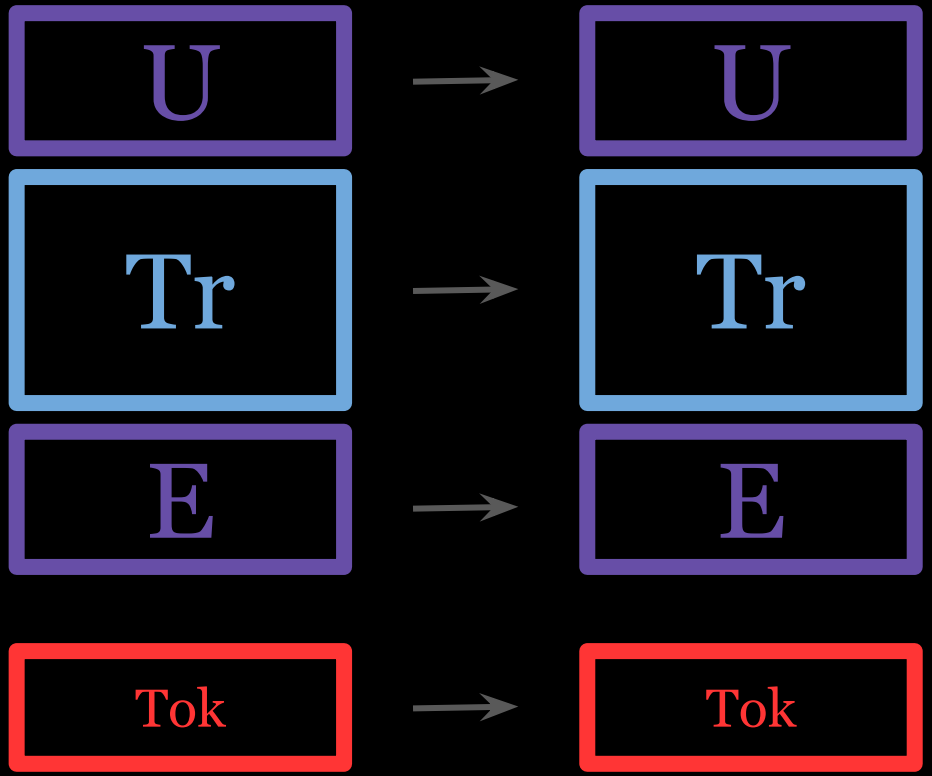
Language Adaptation



Language Adaptation



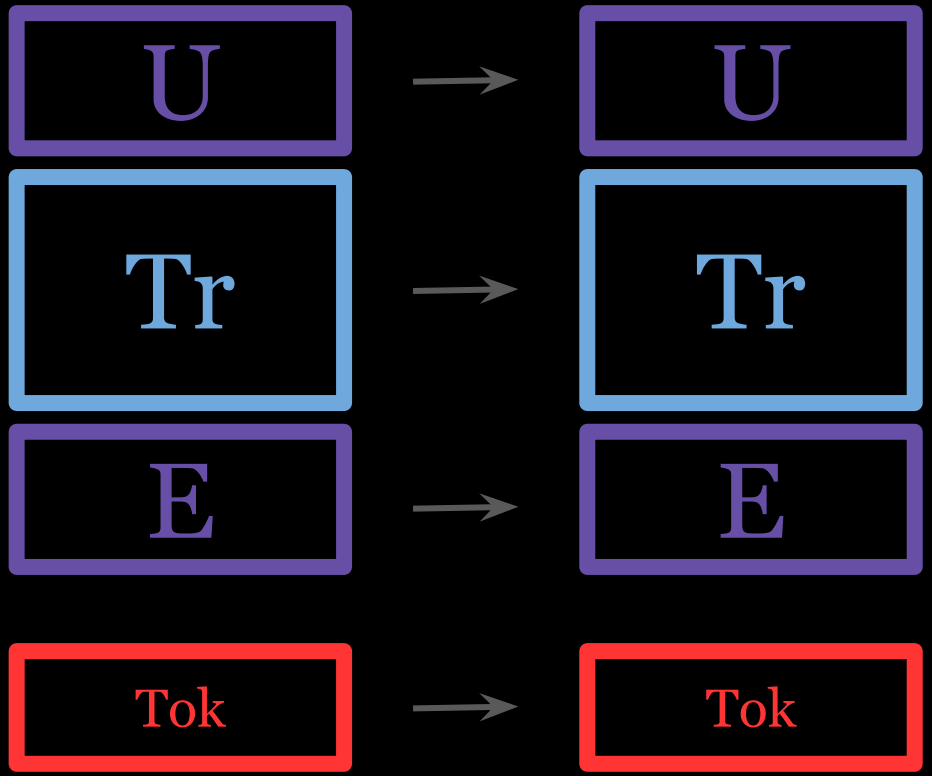
Language Adaptation



Ready?
TRAIN!

Continual Pre-Training

... or CPT



Ready?

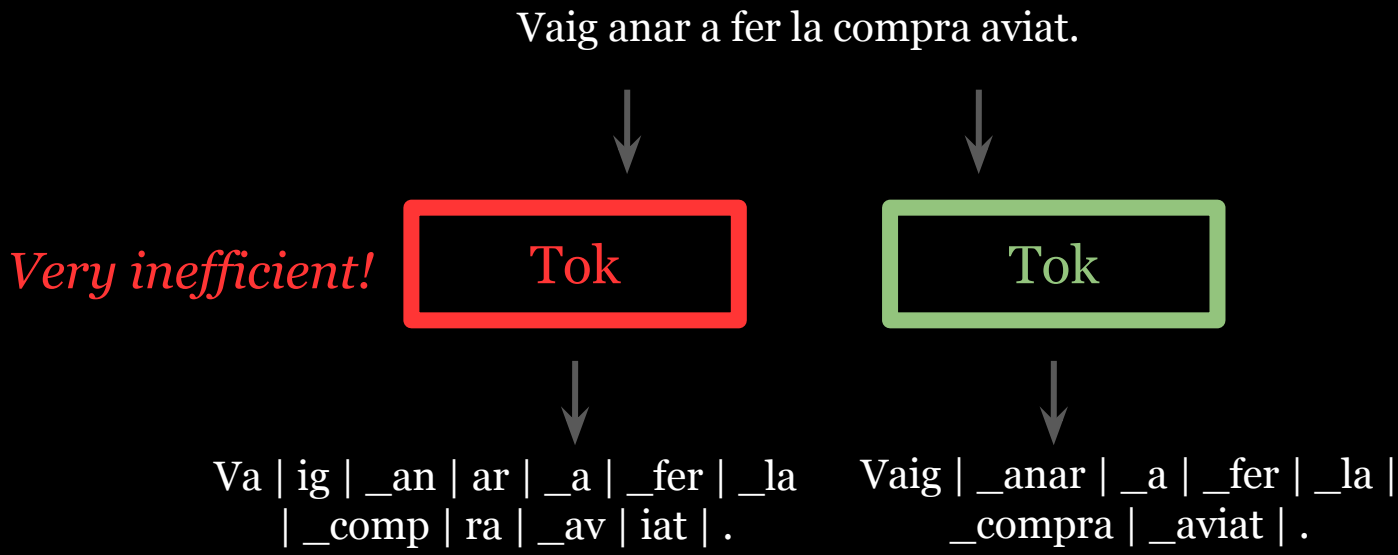
TRAIN!

Continual Pre-Training

Cedille: A large autoregressive French language model
By Martin Müller and Florian Laurent

Sabiá: Portuguese Large Language Models
By Ramon Pires et al.

Continual Pre-Training



Continual Pre-Training

Very inefficient!

Tok

Higher time and computational cost for training and inference

Shorter sequences of text can fit in the input context size of the model

Continual Pre-Training

Very inefficient!

Tok

Tok

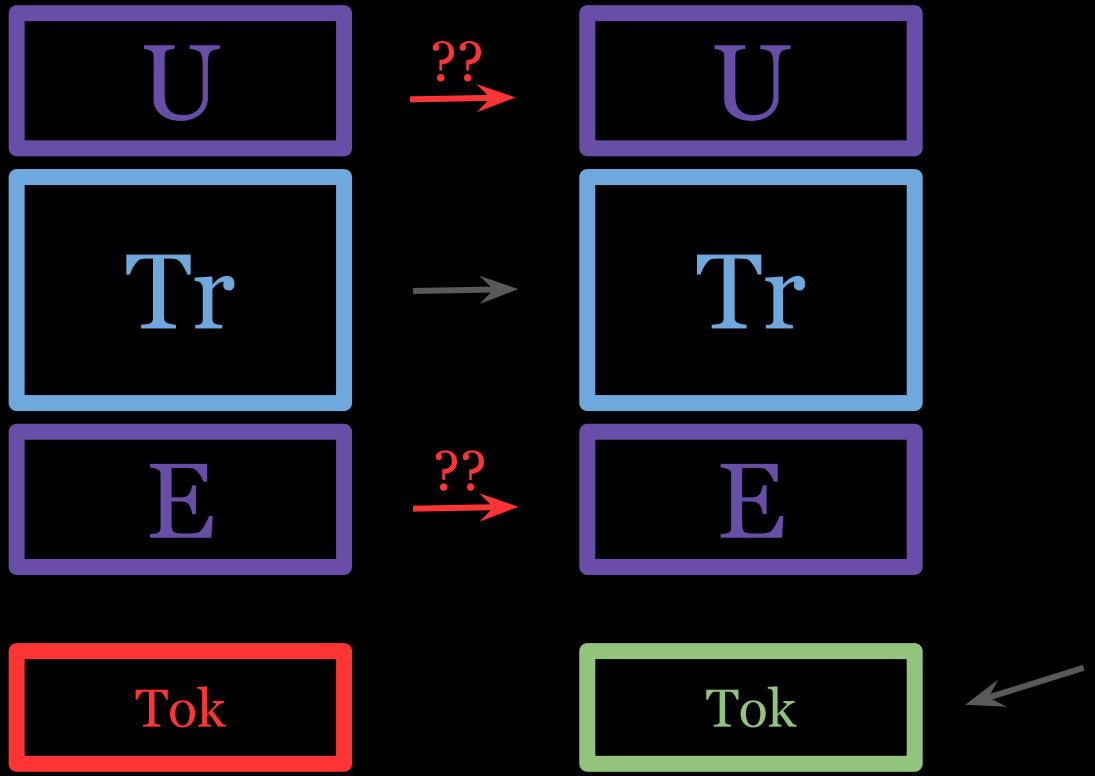
Higher time and computational cost for training and inference

Shorter sequences of text can fit in the input context size of the model

Research seems to suggest that more language-specific tokenization leads to better downstream task performance.

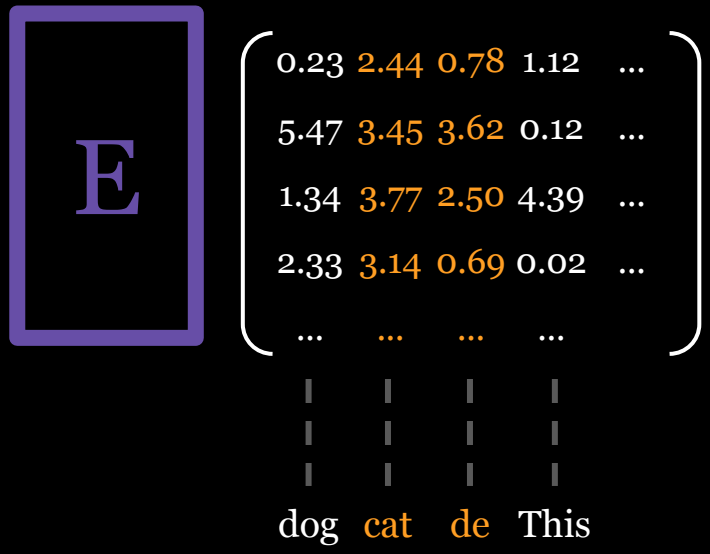
On the Cross-lingual Transferability of Monolingual Representations
By Mikel Artetxe, Sebastian Ruder, Dani Yogatama

Continual Pre-Training

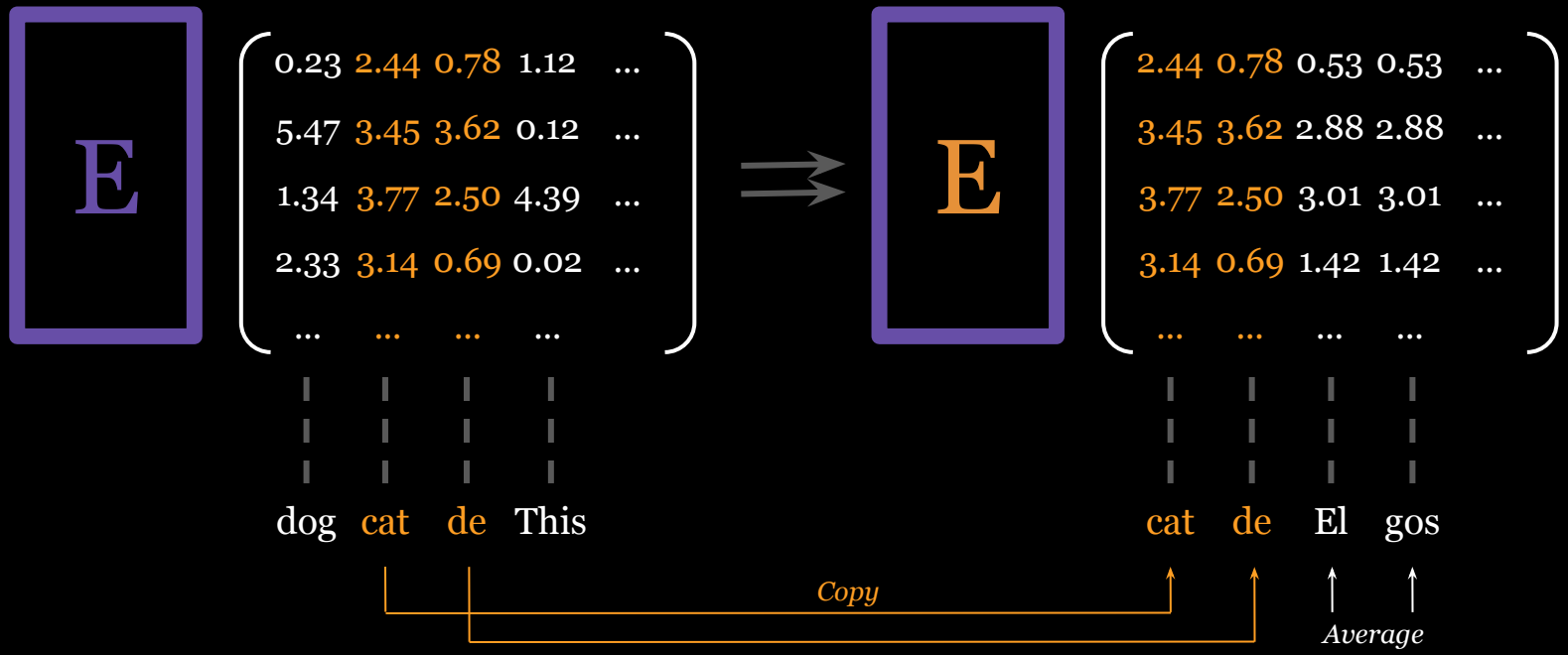


Train a new tokenizer in the target language

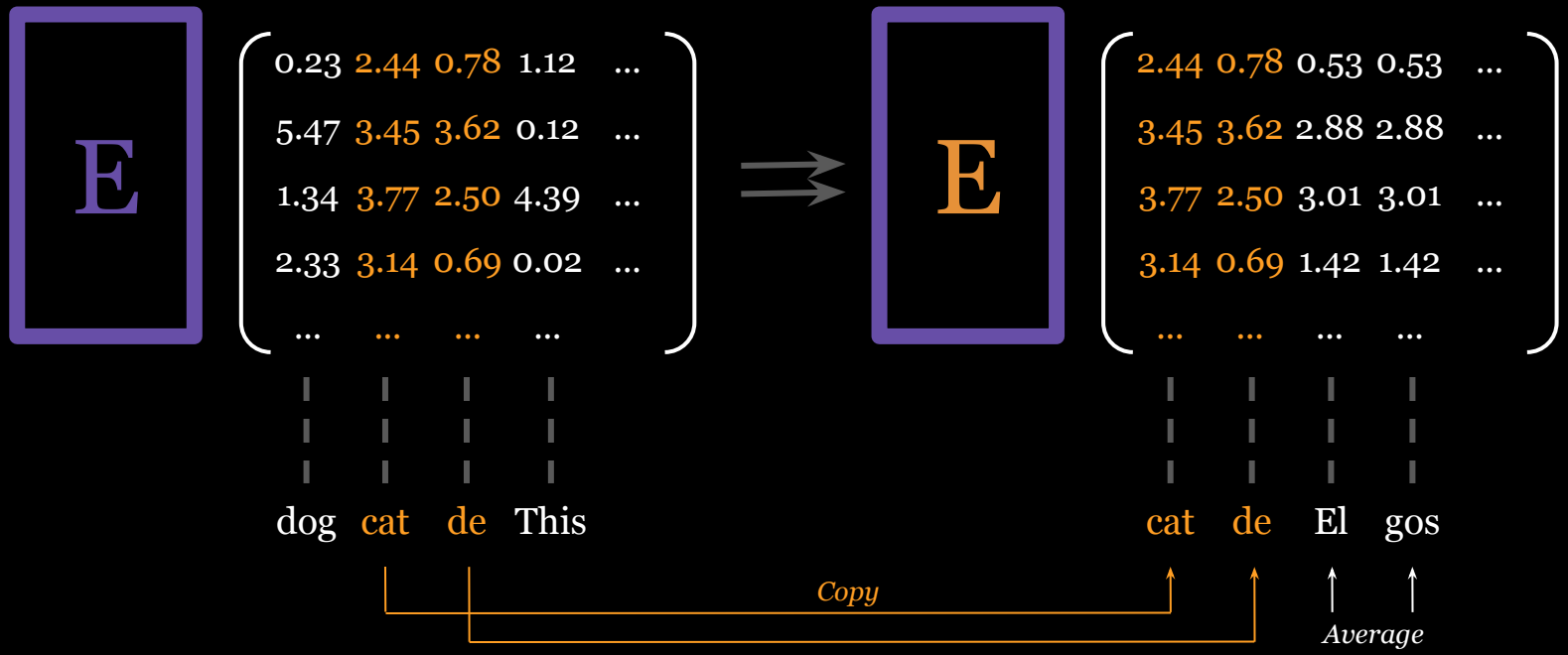
Continual Pre-Training



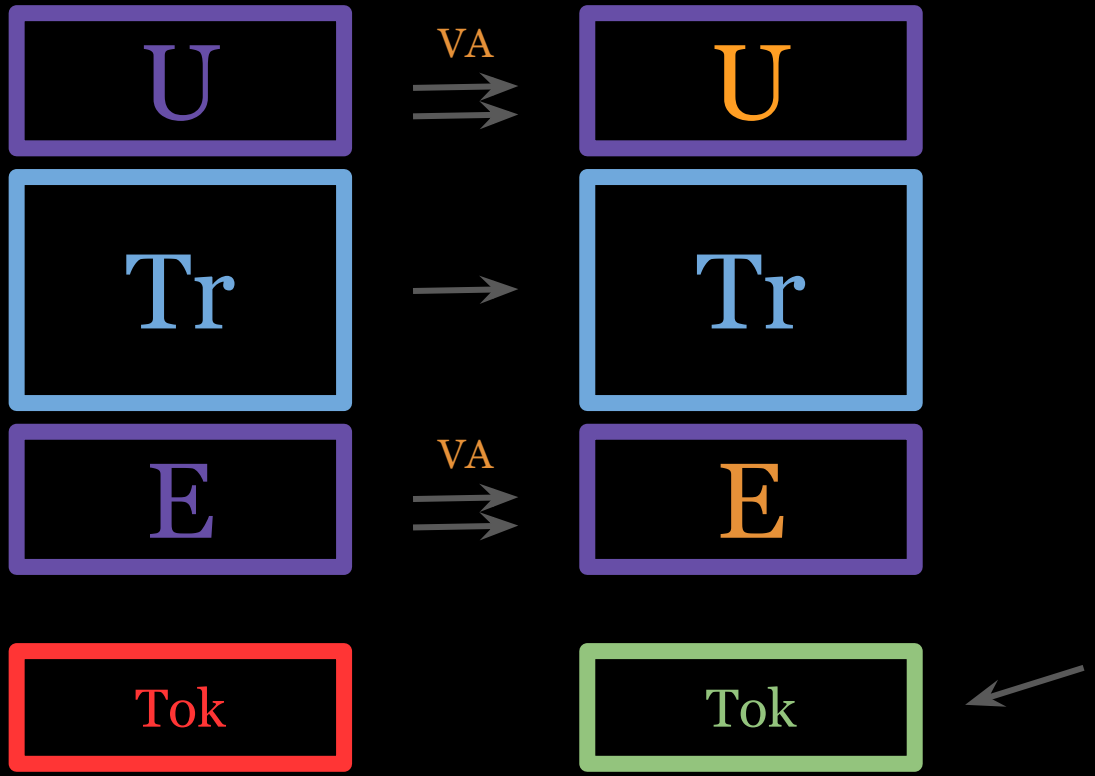
Continual Pre-Training



Vocabulary Adaptation

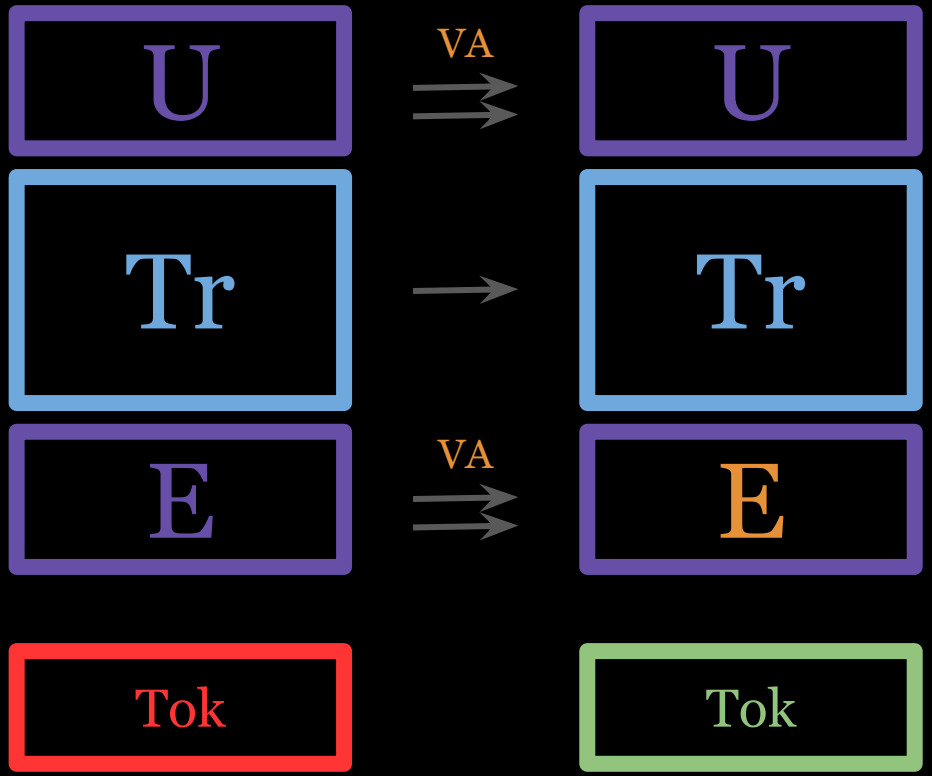


Vocabulary Adaptation



Train a new tokenizer in the target language

Vocabulary Adaptation



Ready?
TRAIN!

Train a new tokenizer in the target language

Continual Pre-Training

As Good as New. How to Successfully Recycle English GPT-2 to Make Models for Other Languages
Wietse de Vries et al.

WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models
Benjamin Minixhofer et al.

Transfer Learning in Multilingual Neural Machine Translation with Dynamic Vocabulary
Surafel M. Lakew et al.

Efficient Language Model Training through Cross-Lingual and Progressive Transfer Learning
Malte Ostendorff et al.

Continual Pre-Training

As Good as New. How to Successfully Recycle English GPT-2 to Make Models for Other Languages
Wietse de Vries et al.

WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models
Benjamin Minixhofer et al.

Transfer Learning in Multilingual Neural Machine Translation with Dynamic Vocabulary
Surafel M. Lakew et al.

Efficient Language Model Training through Cross-Lingual and Progressive Transfer Learning
Malte Ostendorff et al.

Evaluation

Reading Comprehension

Belebele (en, es, ca)

Question Answering

XQuAD (en, es, ca)
CatalanQA (ca)
CoQCat (ca)

NLI

XNLI (en, es, ca)
TE-ca (ca)

Paraphrase Identification

PAWS-X (en, es, ca)
Parafraseja (ca)

Commonsense Reasoning

XStoryCloze (en, es)
COPA (en, ca)

Translation

FLoRes (ca-es, ca-en, es-en)

Evaluation

Reading Comprehension

Belebele (en, es, ca)

Question Answering

XQuAD (en, es, ca)
CatalanQA (ca)
CoQCat (ca)

NLI

XNLI (en, es, ca)
TE-ca (ca)

Paraphrase Identification

PAWS-X (en, es, ca)
Parafraseja (ca)

Commonsense Reasoning

XStoryCloze (en, es)
COPA (en, ca)

Translation

FLoRes (ca-es, ca-en, es-en)

Model

Experiments

Model	Reading Comp.			Question Answering					Natural Language Inference			
	Belebele (acc)			XQuAD (f1)			CatQA (f1)	CoQCat (f1)	XNLI (acc)			TE-ca (acc)
	ca	es	en	ca	es	en	ca	ca	ca	es	en	ca
Random	25.00	25.00	25.00	-	-	-	-	-	33.33	33.33	33.33	33.33
Cerebras-GPT	33.44	31.89	**36.67**	8.56	19.98	**36.00**	10.87	14.12	36.83	38.88	**47.25**	35.62
From Scratch	33.44	31.00	29.00	8.93	8.47	4.19	13.58	18.86	43.77	42.24	38.40	38.78
CPT	39.22	35.11	33.56	26.67	28.17	25.22	34.99	31.93	45.55	44.01	41.48	40.53
Vocab. Adapt	**40.33**	**36.22**	35.33	**28.52**	**30.38**	28.27	**39.99**	**39.50**	**46.21**	**45.61**	43.35	**42.65**

Model	Paraphrase Identification				Commonsense Reasoning				Translation					
	PAWS-X (acc)			Paraf. (acc)	XStoryCloze (acc)		COPA (acc)		FLoRes (bleu)					
	ca	es	en	ca	es	en	ca	en	ca-es	es-ca	ca-en	en-ca	es-en	en-es
Random	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	-	-	-	-	-	-
Cerebras-GPT	52.40	**52.20**	**55.95**	52.05	49.11	**60.62**	51.40	**66.80**	2.42	1.81	2.69	0.82	3.36	1.77
From Scratch	51.45	51.35	53.55	54.15	55.20	53.54	61.40	59.60	2.72	2.22	1.36	1.14	1.51	1.08
CPT	**53.50**	51.40	50.35	53.95	56.45	57.38	61.60	60.80	9.19	14.88	18.23	12.14	13.10	7.71
Vocab. Adapt	49.95	50.85	51.25	**56.30**	**58.64**	59.10	**66.40**	61.60	**16.31**	**19.63**	**26.65**	**24.10**	**17.16**	**15.09**

**Evaluation Harness (by EleutherAI) was used for evaluation*

Experiments

Model	Reading Comp.			Question Answering					Natural Language Inference			
	Belebele (acc)			XQuAD (f1)			CatQA (f1)	CoQCat (f1)	XNLI (acc)			TE-ca (acc)
	ca	es	en	ca	es	en	ca	ca	ca	es	en	ca
Random	25.00	25.00	25.00	-	-	-	-	-	33.33	33.33	33.33	33.33
Cerebras-GPT	33.44	31.89	**36.67**	8.56	19.98	**36.00**	10.87	14.12	36.83	38.88	**47.25**	35.62
From Scratch	33.44	31.00	29.00	8.93	8.47	4.19	13.58	18.86	43.77	42.24	38.40	38.78
CPT	39.22	35.11	33.56	26.67	28.17	25.22	34.99	31.93	45.55	44.01	41.48	40.53
Vocab. Adapt	**40.33**	**36.22**	35.33	**28.52**	**30.38**	28.27	**39.99**	**39.50**	**46.21**	**45.61**	43.35	**42.65**

Model	Paraphrase Identification				Commonsense Reasoning				Translation					
	PAWS-X (acc)			Paraf. (acc)	XStoryCloze (acc)		COPA (acc)		FLoRes (bleu)					
	ca	es	en	ca	es	en	ca	en	ca-es	es-ca	ca-en	en-ca	es-en	en-es
Random	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	-	-	-	-	-	-
Cerebras-GPT	52.40	**52.20**	**55.95**	52.05	49.11	**60.62**	51.40	**66.80**	2.42	1.81	2.69	0.82	3.36	1.77
From Scratch	51.45	51.35	53.55	54.15	55.20	53.54	61.40	59.60	2.72	2.22	1.36	1.14	1.51	1.08
CPT	**53.50**	51.40	50.35	53.95	56.45	57.38	61.60	60.80	9.19	14.88	18.23	12.14	13.10	7.71
Vocab. Adapt	49.95	50.85	51.25	**56.30**	**58.64**	59.10	**66.40**	61.60	**16.31**	**19.63**	**26.65**	**24.10**	**17.16**	**15.09**

**Evaluation Harness (by EleutherAI) was used for evaluation*

Experiments

Model	Reading Comp.			Question Answering					Natural Language Inference			
	Belebele (acc)			XQuAD (f1)			CatQA (f1)	CoQCat (f1)	XNLI (acc)			TE-ca (acc)
	ca	es	en	ca	es	en	ca	ca	ca	es	en	ca
Random	25.00	25.00	25.00	-	-	-	-	-	33.33	33.33	33.33	33.33
Cerebras-GPT	33.44	31.89	**36.67**	8.56	19.98	**36.00**	10.87	14.12	36.83	38.88	**47.25**	35.62
From Scratch	33.44	31.00	29.00	8.93	8.47	4.19	13.58	18.86	43.77	42.24	38.40	38.78
CPT	39.22	35.11	33.56	26.67	28.17	25.22	34.99	31.93	45.55	44.01	41.48	40.53
Vocab. Adapt	**40.33**	**36.22**	35.33	**28.52**	**30.38**	28.27	**39.99**	**39.50**	**46.21**	**45.61**	43.35	**42.65**

Model	Paraphrase Identification				Commonsense Reasoning				Translation					
	PAWS-X (acc)			Paraf. (acc)	XStoryCloze (acc)		COPA (acc)		FLoRes (bleu)					
	ca	es	en	ca	es	en	ca	en	ca-es	es-ca	ca-en	en-ca	es-en	en-es
Random	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	-	-	-	-	-	-
Cerebras-GPT	52.40	**52.20**	**55.95**	52.05	49.11	**60.62**	51.40	**66.80**	2.42	1.81	2.69	0.82	3.36	1.77
From Scratch	51.45	51.35	53.55	54.15	55.20	53.54	61.40	59.60	2.72	2.22	1.36	1.14	1.51	1.08
CPT	**53.50**	51.40	50.35	53.95	56.45	57.38	61.60	60.80	9.19	14.88	18.23	12.14	13.10	7.71
Vocab. Adapt	49.95	50.85	51.25	**56.30**	**58.64**	59.10	**66.40**	61.60	**16.31**	**19.63**	**26.65**	**24.10**	**17.16**	**15.09**

**Evaluation Harness (by EleutherAI) was used for evaluation*

Experiments

Model	Reading Comp.			Question Answering					Natural Language Inference			
	Belebele (acc)			XQuAD (f1)			CatQA (f1)	CoQCat (f1)	XNLI (acc)			TE-ca (acc)
	ca	es	en	ca	es	en	ca	ca	ca	es	en	ca
Random	25.00	25.00	25.00	-	-	-	-	-	33.33	33.33	33.33	33.33
Cerebras-GPT	33.44	31.89	**36.67**	8.56	19.98	**36.00**	10.87	14.12	36.83	38.88	**47.25**	35.62
From Scratch	33.44	31.00	29.00	8.93	8.47	4.19	13.58	18.86	43.77	42.24	38.40	38.78
CPT	39.22	35.11	33.56	26.67	28.17	25.22	34.99	31.93	45.55	44.01	41.48	40.53
Vocab. Adapt	**40.33**	**36.22**	35.33	**28.52**	**30.38**	28.27	**39.99**	**39.50**	**46.21**	**45.61**	43.35	**42.65**

Model	Paraphrase Identification				Commonsense Reasoning				Translation					
	PAWS-X (acc)			Paraf. (acc)	XStoryCloze (acc)		COPA (acc)		FLoRes (bleu)					
	ca	es	en	ca	es	en	ca	en	ca-es	es-ca	ca-en	en-ca	es-en	en-es
Random	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	-	-	-	-	-	-
Cerebras-GPT	52.40	**52.20**	**55.95**	52.05	49.11	**60.62**	51.40	**66.80**	2.42	1.81	2.69	0.82	3.36	1.77
From Scratch	51.45	51.35	53.55	54.15	55.20	53.54	61.40	59.60	2.72	2.22	1.36	1.14	1.51	1.08
CPT	**53.50**	51.40	50.35	53.95	56.45	57.38	61.60	60.80	9.19	14.88	18.23	12.14	13.10	7.71
Vocab. Adapt	49.95	50.85	51.25	**56.30**	**58.64**	59.10	**66.40**	61.60	**16.31**	**19.63**	**26.65**	**24.10**	**17.16**	**15.09**

**Evaluation Harness (by EleutherAI) was used for evaluation*

Experiments

Model	Reading Comp.			Question Answering					Natural Language Inference			
	Belebele (acc)			XQuAD (f1)			CatQA (f1)	CoQCat (f1)	XNLI (acc)			TE-ca (acc)
	ca	es	en	ca	es	en	ca	ca	ca	es	en	ca
Random	25.00	25.00	25.00	-	-	-	-	-	33.33	33.33	33.33	33.33
Cerebras-GPT	33.44	31.89	**36.67**	8.56	19.98	**36.00**	10.87	14.12	36.83	38.88	**47.25**	35.62
From Scratch	33.44	31.00	29.00	8.93	8.47	4.19	13.58	18.86	43.77	42.24	38.40	38.78
CPT	39.22	35.11	33.56	26.67	28.17	25.22	34.99	31.93	45.55	44.01	41.48	40.53
Vocab. Adapt	**40.33**	**36.22**	35.33	**28.52**	**30.38**	28.27	**39.99**	**39.50**	**46.21**	**45.61**	43.35	**42.65**

Model	Paraphrase Identification				Commonsense Reasoning				Translation					
	PAWS-X (acc)			Paraf. (acc)	XStoryCloze (acc)		COPA (acc)		FLoRes (bleu)					
	ca	es	en	ca	es	en	ca	en	ca-es	es-ca	ca-en	en-ca	es-en	en-es
Random	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	-	-	-	-	-	-
Cerebras-GPT	52.40	**52.20**	**55.95**	52.05	49.11	**60.62**	51.40	**66.80**	2.42	1.81	2.69	0.82	3.36	1.77
From Scratch	51.45	51.35	53.55	54.15	55.20	53.54	61.40	59.60	2.72	2.22	1.36	1.14	1.51	1.08
CPT	**53.50**	51.40	50.35	53.95	56.45	57.38	61.60	60.80	9.19	14.88	18.23	12.14	13.10	7.71
Vocab. Adapt	49.95	50.85	51.25	**56.30**	**58.64**	59.10	**66.40**	61.60	**16.31**	**19.63**	**26.65**	**24.10**	**17.16**	**15.09**

**Evaluation Harness (by EleutherAI) was used for evaluation*

Catastrophic Forgetting!

Can we improve this?

Results

Model	Training tokens		Reading Comprehension			Question Answering				
			Belebele$_{acc}$			XQuAD$_{f1}$			CatalanQA$_{f1}$	CoQCat$_{f1}$
	Pre-train	Lang. adapt.	ca	es	en	ca	es	en	ca	ca
Random	-	-	25.00	25.00	25.00	-	-	-	-	-
mGPT-1.3B	440B	-	26.11	24.44	26.11	0.33	0.67	0.17	0.65	0.78
GPT-Neo-1.3B	380B	-	35.44	32.77	41.67	19.75	29.77	51.53	22.34	23.57
Pythia-1.4B	299.9B	-	35.78	35.11	41.56	26.19	34.13	52.98	27.47	25.38
OPT-1.3B	180B	-	35.22	33.56	43.78	23.53	31.85	52.95	26.58	20.18
Falcon-rw-1.3B	350B	-	34.33	33.56	**47.89**	5.93	19.25	**58.60**	6.91	15.61
Cerebras-GPT-1.3B	26B	-	33.44	31.89	36.67	8.56	19.98	36.00	10.87	14.12
BLOOM-1.1B	341B	-	39.89	37.22	39.33	36.81	36.98	44.10	44.65	34.57
From_scratch-1.3B	26B	-	33.44	31.00	29.00	8.93	8.47	4.19	13.58	18.86
Cerebras-GPT-continued_pre-training-1.3B	26B	38B	39.22	35.11	33.56	26.67	28.17	25.22	34.99	31.93
Cerebras-GPT-vocab_adapted-1.3B	26B	26B	40.33	36.22	35.33	28.52	30.38	28.27	39.99	39.50
BLOOM-vocab_adapted-760M	341B	26B	41.00	37.89	37.00	41.10	41.11	40.20	51.01	41.34
BLOOM-vocab_adapted-1.3B	341B	26B	**43.44**	**39.11**	40.44	**43.52**	**44.31**	44.11	**54.25**	**48.15**

Results

Model	Training tokens		Natural Language Inference				Paraphrase Identification			
			XNLI$_{acc}$			TE-ca$_{acc}$	PAWS-X$_{acc}$			Parafraseja$_{acc}$
	Pre-train	Lang. adapt.	ca	es	en	ca	ca	es	en	ca
Random	-	-	33.33	33.33	33.33	33.33	50.00	50.00	50.00	50.00
mGPT-1.3B	440B	-	40.06	43.81	45.67	37.03	51.00	52.30	56.15	51.32
GPT-Neo-1.3B	380B	-	41.44	45.57	49.92	35.38	54.65	53.40	54.60	51.70
Pythia-1.4B	299.9B	-	42.46	45.61	51.00	37.46	54.15	52.50	**57.70**	55.23
OPT-1.3B	180B	-	40.08	44.53	**52.48**	36.14	54.10	52.55	55.90	53.23
Falcon-rw-1.3B	350B	-	34.53	35.85	45.73	34.96	54.25	**54.05**	53.65	50.60
Cerebras-GPT-1.3B	26B	-	36.83	38.88	47.25	35.62	52.40	52.20	55.95	52.05
BLOOM-1.1B	341B	-	47.19	46.39	49.44	41.38	**55.05**	54.05	54.75	55.65
From_scratch-1.3B	26B	-	43.77	42.24	38.40	38.78	51.45	51.35	53.55	54.15
Cerebras-GPT-continued_pre-training-1.3B	26B	38B	45.55	44.01	41.48	40.53	53.50	51.40	50.35	53.95
Cerebras-GPT-vocab_adapted-1.3B	26B	26B	46.21	45.61	43.35	42.65	49.95	50.85	51.25	56.30
BLOOM-vocab_adapted-760M	341B	26B	46.93	46.03	46.11	42.14	52.35	52.50	54.85	56.55
BLOOM-vocab_adapted-1.3B	341B	26B	**49.20**	**48.82**	47.45	**42.89**	53.20	52.85	53.00	**57.43**

Results

Model	Training tokens		Commonsense Reasoning				Translation					
			$\text{XStoryCloze}_{\text{acc}}$		COPA_{acc}		$\text{FLoRes}_{\text{bleu}}$					
	Pre-train	Lang. adapt.	es	en	ca	en	ca→es	es→ca	ca→en	en→ca	es→en	en→es
Random	-	-	50.00	50.00	50.00	50.00	-	-	-	-	-	-
mGPT-1.3B	440B	-	55.33	60.09	52.20	63.40	3.25	2.96	9.25	3.79	17.75	15.34
GPT-Neo-1.3B	380B	-	51.42	66.58	53.40	74.80	3.27	3.80	17.77	5.49	17.70	12.04
Pythia-1.4B	299.9B	-	54.14	68.37	52.20	78.60	9.68	5.74	24.03	11.10	21.50	15.04
OPT-1.3B	180B	-	53.94	69.95	52.60	76.20	3.14	3.52	15.39	2.00	16.33	6.53
Falcon-rw-1.3B	350B	-	51.09	**71.34**	52.40	**79.60**	3.03	3.59	8.89	3.01	14.17	6.50
Cerebras-GPT-1.3B	26B	-	49.11	60.62	51.40	66.80	2.42	1.81	2.69	0.82	3.36	1.77
BLOOM-1.1B	341B	-	57.91	62.48	62.80	66.40	21.62	15.28	31.16	21.28	20.92	16.84
From_scratch-1.3B	26B	-	55.20	53.54	61.40	59.60	2.72	2.22	1.36	1.14	1.51	1.08
Cerebras-GPT-continued_pre-training-1.3B	26B	38B	56.45	57.38	61.60	60.80	9.19	14.88	18.23	12.14	13.10	7.71
Cerebras-GPT-vocab_adapted-1.3B	26B	26B	58.64	59.10	66.40	61.60	16.31	**19.63**	26.65	24.10	17.16	15.09
BLOOM-vocab_adapted-760M	341B	26B	61.42	61.42	65.40	64.20	**22.62**	15.77	32.26	26.04	20.91	18.08
BLOOM-vocab_adapted-1.3B	341B	26B	**64.06**	61.81	**68.00**	67.80	22.16	18.58	**33.95**	**29.31**	**23.09**	**20.30**

Takeaways

Do NOT train from scratch.

Consider adapting the vocabulary instead of CPT!

Use a pre-trained model as close as possible to your language as starting point.

Acknowledgments

Aina

PERTE
Nueva Economía
de la Lengua

Financiado por
la Unión Europea
NextGenerationEU

ILENIA

GOBIERNO
DE ESPAÑA

MINISTERIO
DE TRANSFORMACIÓN DIGITAL

Generalitat
de Catalunya

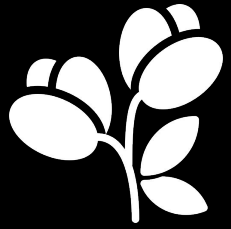
Plan de
Recuperación,
Transformación
y Resiliencia

red.es

Acknowledgments

This work has been promoted and financed by the Generalitat de Catalunya through the Aina project.

Este proyecto con referencia 2022/TL22/00215337, 2022/TL22/00215336, 2022/TL22/00215335 y 2022/TL22/00215334 está financiado por el Ministerio para la Transformación Digital y de la Función Pública y por el Plan de Recuperación, Transformación y Resiliencia - Financiado por la Unión Europea – NextGenerationEU.



Thank You!