Building and Evolving a Multilingual LLM from Scratch

Javier Aula-Blasco, PhD

/ CyberColombia
/ 8th Colombian HPC Summer School
/ June 17, 2025

Language Technologies Lab

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Infrastructure

MareNostrum 5

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

GPP based on Intel Sapphire Rapids, 6480 nodes, 45.9 PFlops peak.

ACC based on Intel Sapphire Rapids and NVIDIA Hopper GPUs, 1120 nodes with 4 Hopper GPUs each one, 260 PFlops peak.

GPP - Next Gen based on NVIDIA GRACE CPUs.

ACC - Next Generation coming soon.

MareNostrum 5

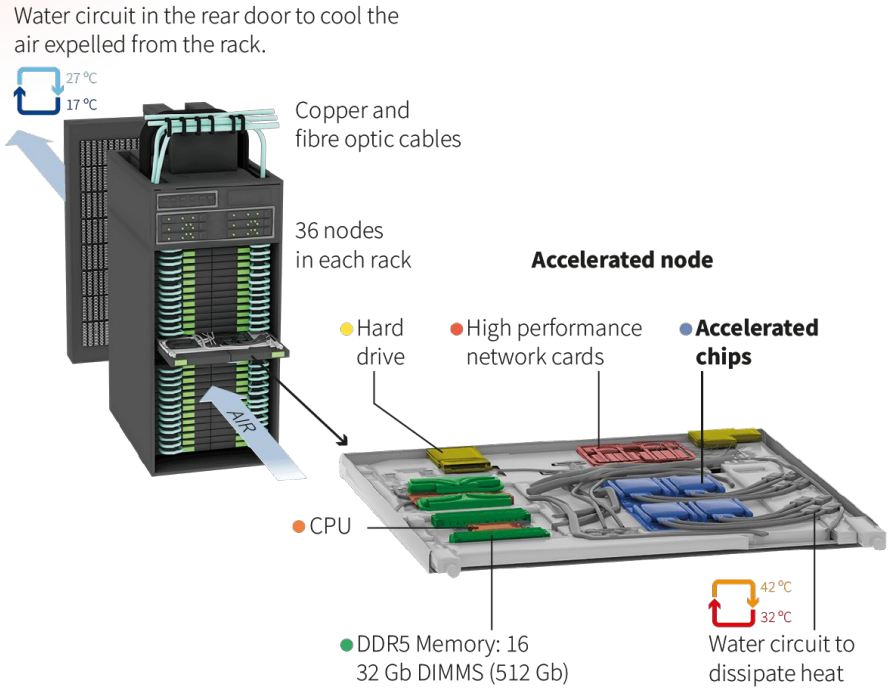
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Accelerated Partition

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Node configuration:

- 2x Intel Sapphire Rapids 8460Y+ at 2.3Ghz and 40c each
- 512 Gb of Main memory, using DDR5
- 4x NVIDIA Hopper GPU with 64Gb HBM2 memory
- 480Gb on NVMe storage
- 4x NDR200 (BW per node 800Gb/s)



Salamandra & ALIA models

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-LT/ALIA-40b
Text Generation • Updated Apr 3 • 331 • 75

BSC-LT/salamandra-7b-instruct
Text Generation • Updated Feb 20 • 11.3k • 59

BSC-LT/salamandra-7b
Text Generation • Updated Feb 20 • 432 • 27

BSC-LT/salamandra-2b-instruct
Text Generation • Updated Feb 20 • 970 • 21

BSC-LT/salamandra-2b
Text Generation • Updated Feb 20 • 1.06k • 23

BSC-LT/salamandra-7b-instruct-fp8
Text Generation • Updated Mar 26 • 26 • 1

BSC-LT/salamandra-7b-instruct-gptq
Text Generation • Updated Mar 26 • 66

BSC-LT/salamandra-2b-instruct-fp8
Text Generation • Updated Mar 26 • 10

BSC-LT/salamandra-2b-instruct-gptq
Text Generation • Updated Mar 26 • 5

BSC-LT/salamandra-7b-base-fp8
Text Generation • Updated Mar 26 • 17 • 1

BSC-LT/salamandra-7b-base-gptq
Text Generation • Updated Mar 26 • 52

BSC-LT/salamandra-2b-base-fp8
Text Generation • Updated Mar 26 • 19

BSC-LT/salamandra-2b-base-gptq
Text Generation • Updated Mar 18 • 9

Salamandra Technical Report

Language Technologies Unit

Barcelona Supercomputing Center

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Data

Pre-training data

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

In the pre-training phase, it is essential to have a massive volume of data. This volume is hardly achievable with conventional data sources:

3.7T

Chinchilla optimal for 175B parameter model (Hoffman et al)

vs

0.000911T

All Wikipedia in Spanish

vs

0.0000005T

Don Quixote

Larger sources (mainly mass scraping) are used, to which pre-processing tasks must be applied, including, depending on the case, the following layers: extraction, deduplication, language identification, filtering/quality assessment, removal of evaluation data or detection of inappropriate content.

Pre-training resources

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Parameters	Nodes (MN5)	Time (days)	Model	Tokens	Tokens in Spanish
~175B	-	?	ChatGPT3.5	?	?
176B	512	5	BLOOM	0,366 T	0,039 T (10,8%)
180B	512	43	Falcon-180B	3,5 T	0,059 T (1,68%)
40B	512	17	Falcon-40B	1 T	0,017 T (1,68%)
70B	512	29	Llama2-70B	2 T	0,002 T (0,13%)
7B	512	2	Llama2-7B	2 T	0,002 T (0,13%)
7B	512	5	Mistral-7B	8T?	?

Challenges

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

1. Language imbalance.
2. Traceability and control of the data used for training.
3. Data quality.
4. Legal aspects.
5. Hate speech, biases, toxic content, irrelevance, etc.

Approach

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Creating a **data infrastructure** together with the different actors in society.

Developing, contributing to, and identifying **data processing tools** that facilitate addressing these challenges.

Researching how to achieve high-performance models **with limited resources**.

Language imbalance

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

The main source of text data is Common Crawl (CC) or derivatives.

This crawling is done by domains and the main target is English.

In previous dumps, Catalan and Spanish have accounted for around 0.2% and 4.5% of the total respectively, against 45% for English.

crawl	CC-MAIN-2023-40	CC-MAIN-2023-50	CC-MAIN-2024-10
language	%	%	%
eng	46.4328	44.4285	46.4536
deu	5.8355	5.4499	5.3948
rus	5.5709	6.0303	5.8095
jpn	4.7475	5.1508	5.0934
fra	4.6351	4.3933	4.3263
spa	4.6199	4.5391	4.5462
cat	0.2096	0.2053	0.2036

% languages in last three CC dumps

Data agreements

HiTZ

RACÓ CATALÀ

Generalitat
de Catalunya
Agència Catalana
de Turisme

Generalitat de Catalunya
Entitat Autònoma del Diari Oficial
i de Publicacions

dfki
ai

Corporació Catalana
de Mitjans Audiovisuals, SA

CORTS VALENCIANES

UIB
Universitat
de les Illes Balears

CENID

elMon

amic
associació de mitjans d'informació i comunicació

INSTITUT D'ESTUDIS ARANESI
acadèmia aranesa dera lengua occitana

SciELO

ÒMNIUM

PGC

PARLAMENT DE CATALUNYA

Dialnet

WIKIPEDIA
The Free Encyclopedia

FECYT
INNOVACIÓN

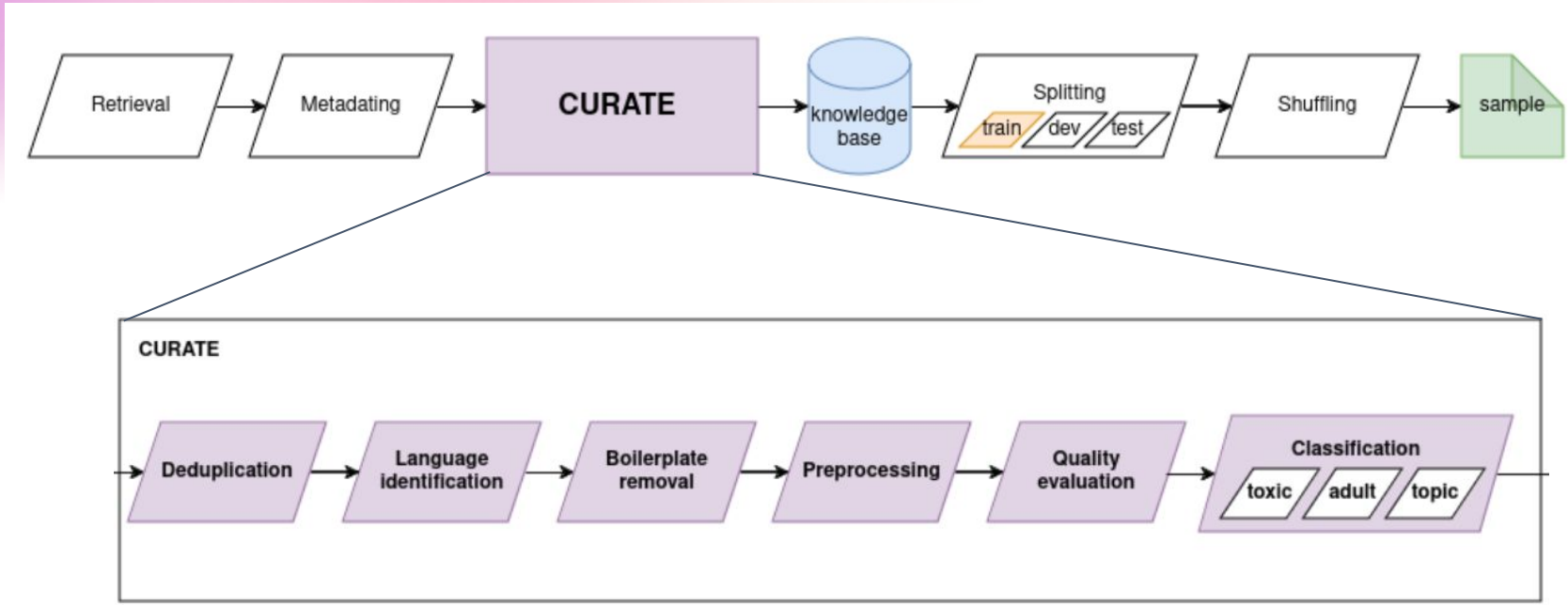
RECOLECTA
RECOLECTOR DE CIENCIA ABIERTA

OpenAIRE

CITIUS
Centro Singular de Investigación
en Tecnoloxías Intelixentes

CURATE Pipeline

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



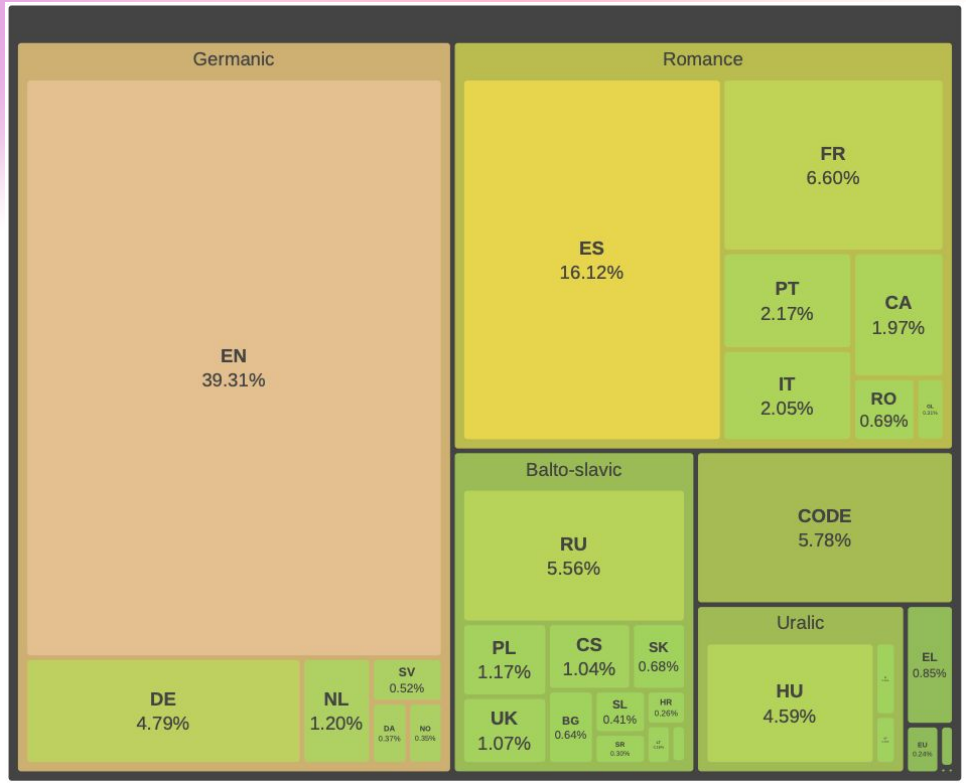
A CURATEd CATalog: Rethinking the Extraction of Pretraining Corpora for Mid-Resourced Languages (Palomar-Giner et al. 2024). *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

CURATE Pipeline

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

langtech-bsc/CURATE

Language distribution

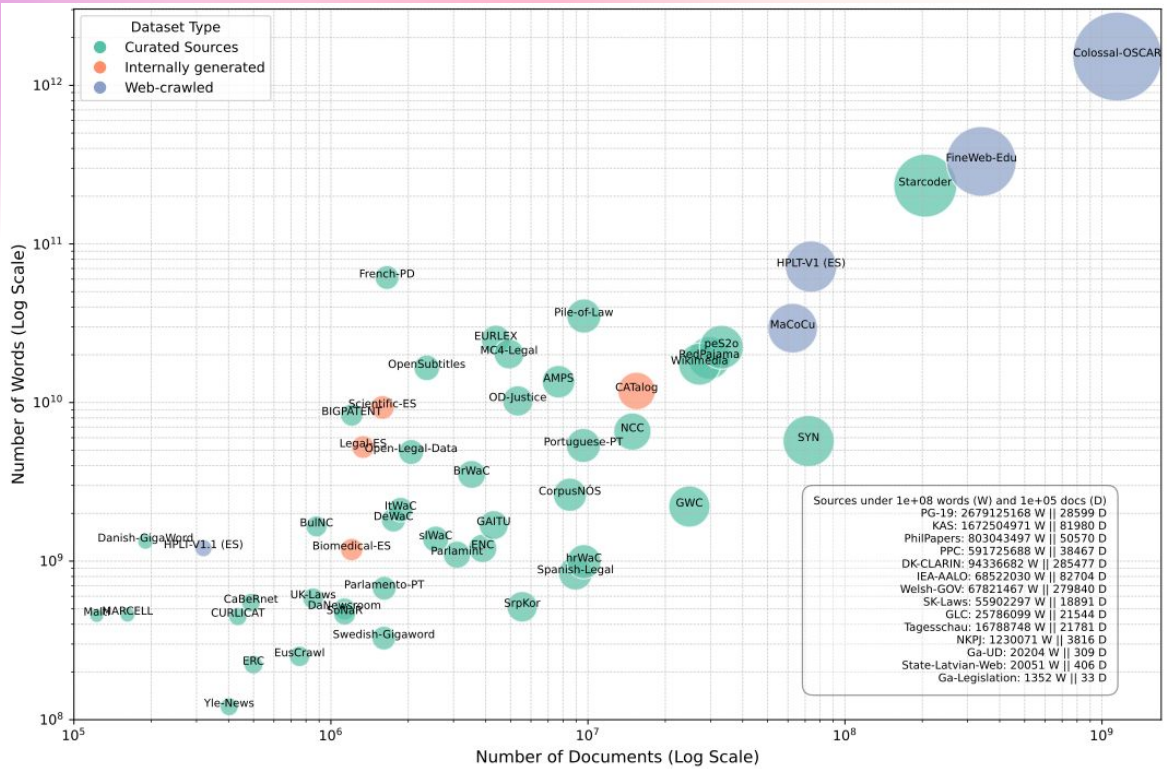


35 EU languages + code

Documents: 1.967,73 M

Words: 2.251.075,65 M

Source distribution



Ungoliant: An Optimized Pipeline for the Generation of a Very Large-Scale Multilingual Web Corpus

Julien Abadji[1] Pedro Javier Ortiz Suárez[1,2] Laurent Romary[1] Benoît Sagot[1]
[1]Inria, Paris, France
[2]Sorbonne Université, Paris, France

Colossal OSCAR curates 20 **multilingual** CommonCrawl **snapshots**.

The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale

Guilherme Penedo Hynek Kydlíček Loubna Ben allal Anton Lozhkov
Margaret Mitchell Colin Raffel Leandro Von Werra Thomas Wolf
🤗 Hugging Face

StarCoder contains data in **86 programming languages**.

CATalog corpus

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

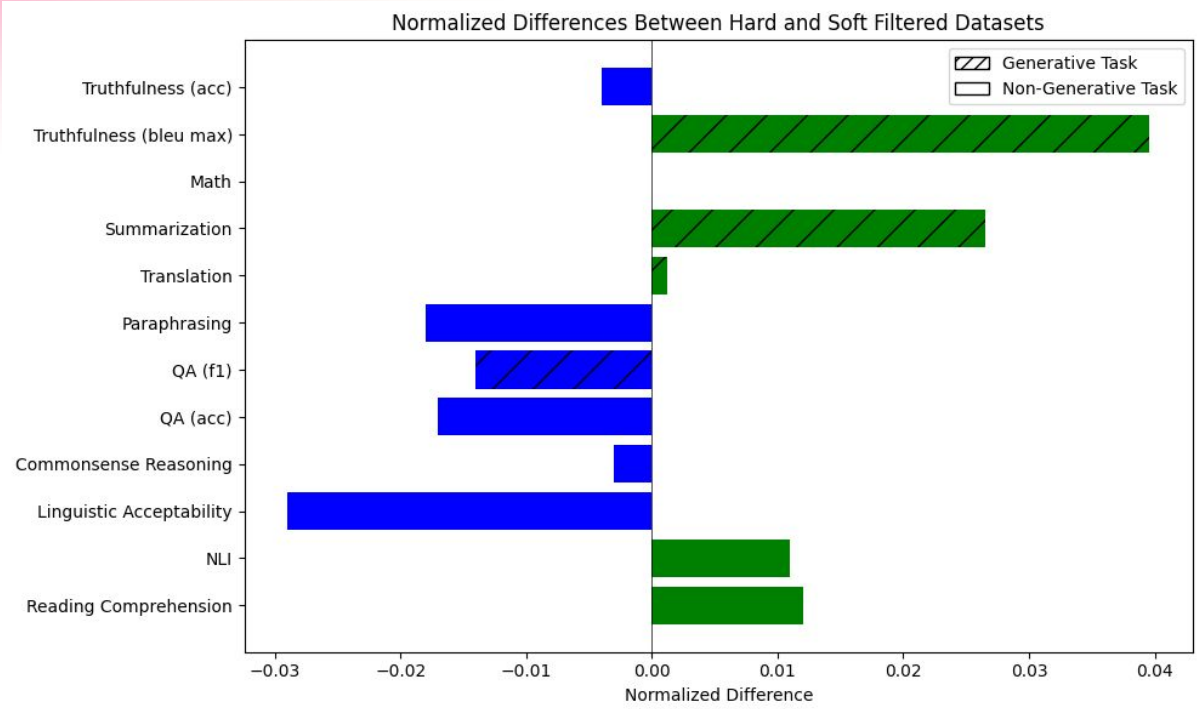
projecte-aina/CATalog

Pre-training data

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Two 7B with two different thresholds on data quality. On the left, soft filtering; on the right, hard filtering.

Multilingual setup.

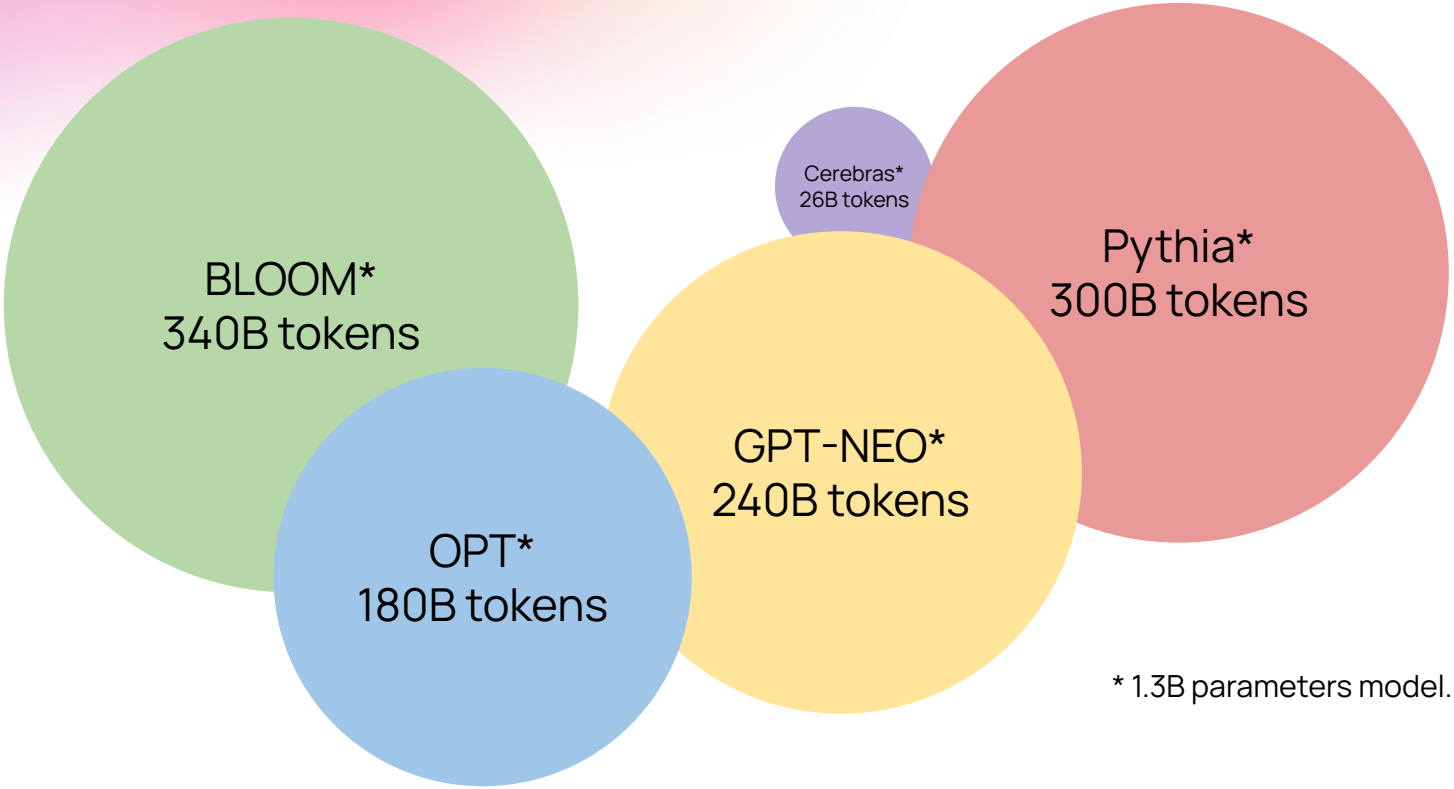


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Pre-training

Number of tokens

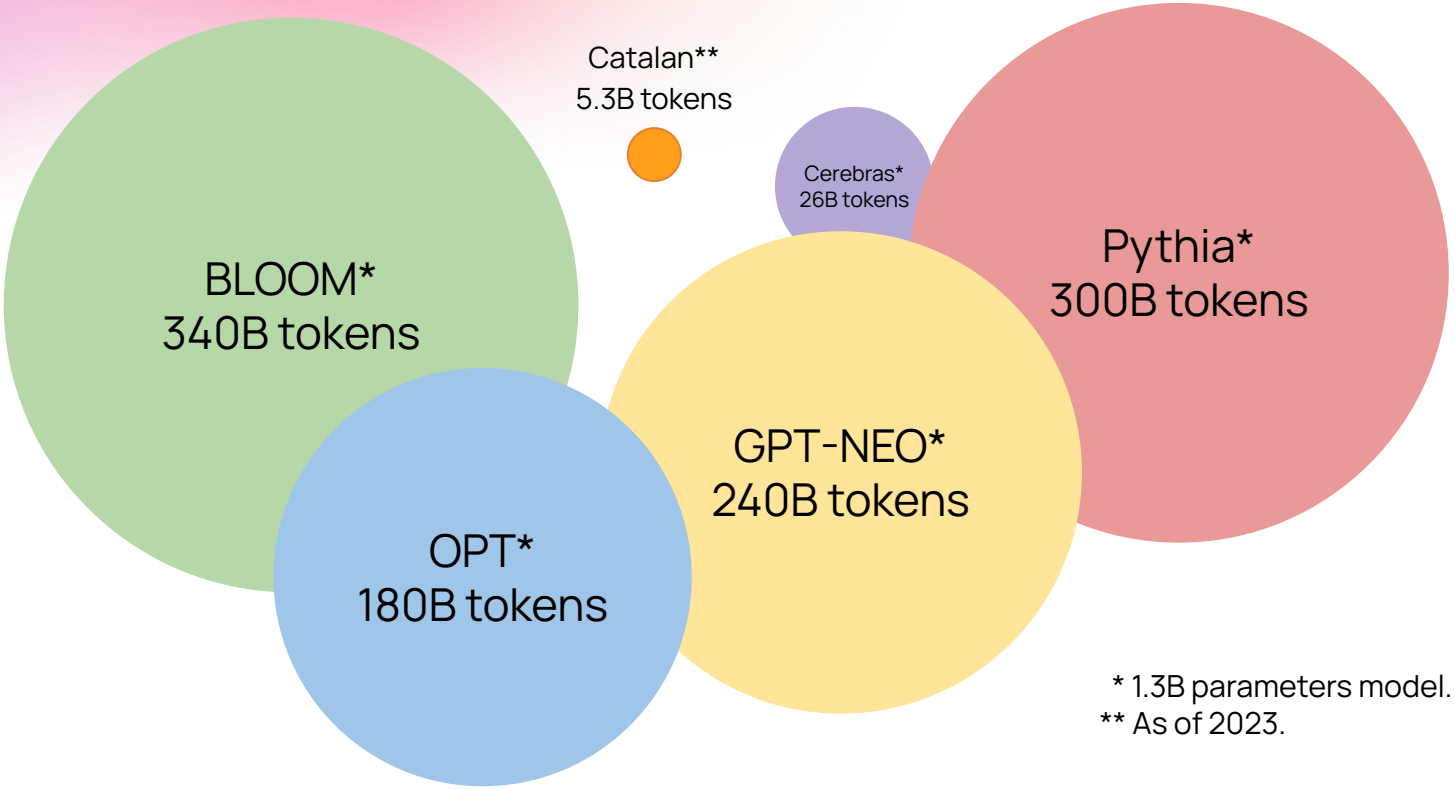
Barcelona Supercomputing Center
Centro Nacional de Supercomputación



* 1.3B parameters model.

Number of tokens

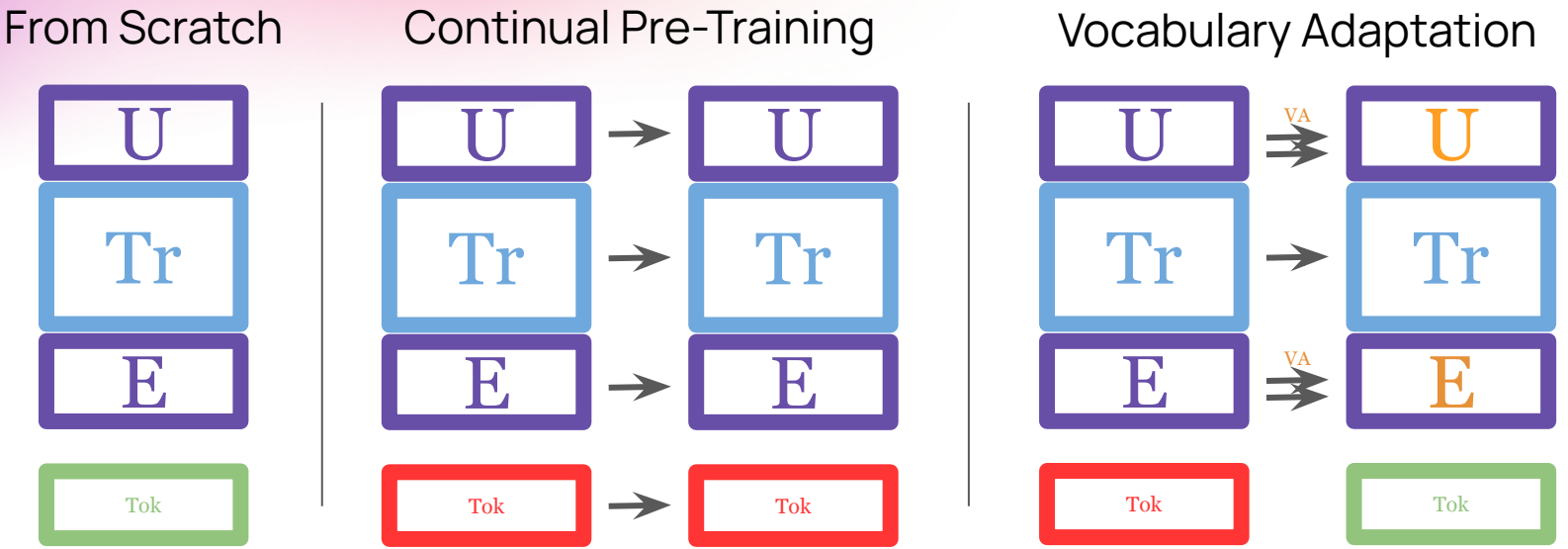
Barcelona Supercomputing Center
Centro Nacional de Supercomputación



* 1.3B parameters model.
** As of 2023.

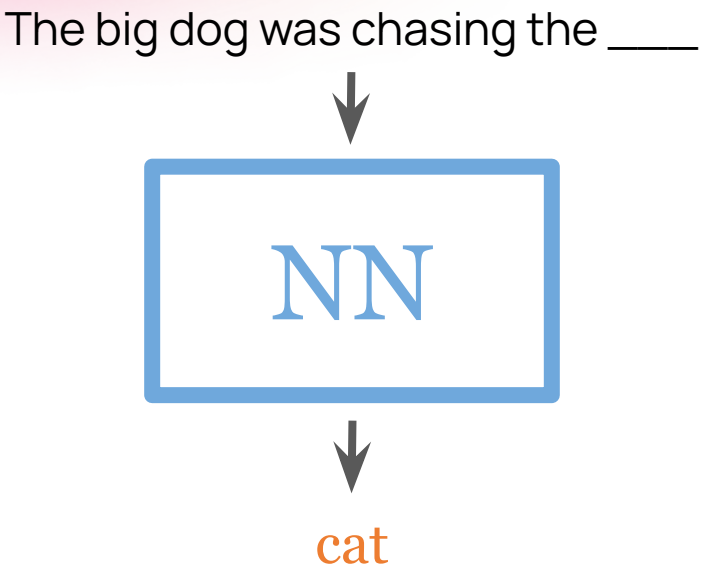
Strategies

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



Decoders

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



Decoders

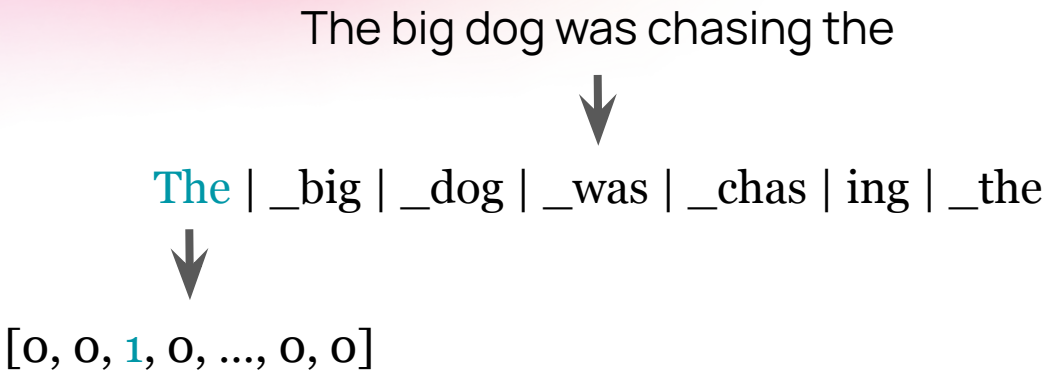
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

The big dog was chasing the

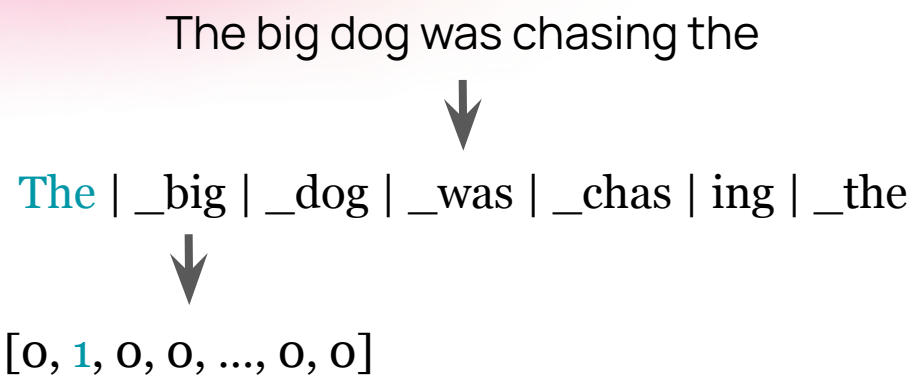
The | _big | _dog | _was | _chas | ing | _the

Decoders

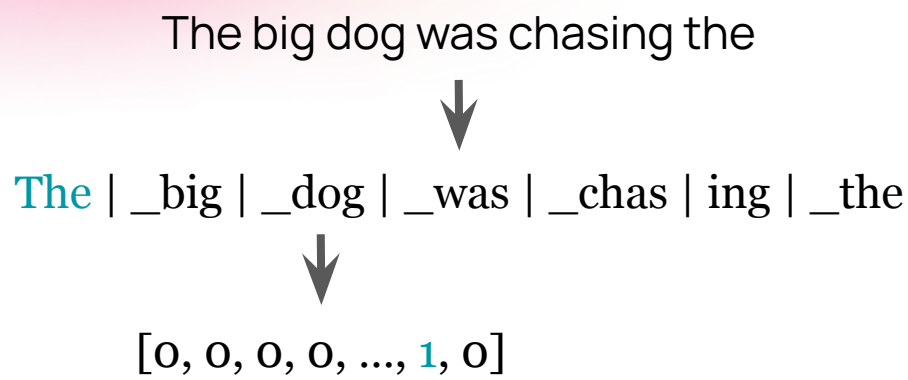
Barcelona Supercomputing Center
Centro Nacional de Supercomputación



Decoders


The big dog was chasing the
The | _big | _dog | _was | _chas | ing | _the
[0, 1, 0, 0, ..., 0, 0]

Decoders


The big dog was chasing the
The | _big | _dog | _was | _chas | ing | _the
[0, 0, 0, 0, ..., 1, 0]

Decoders

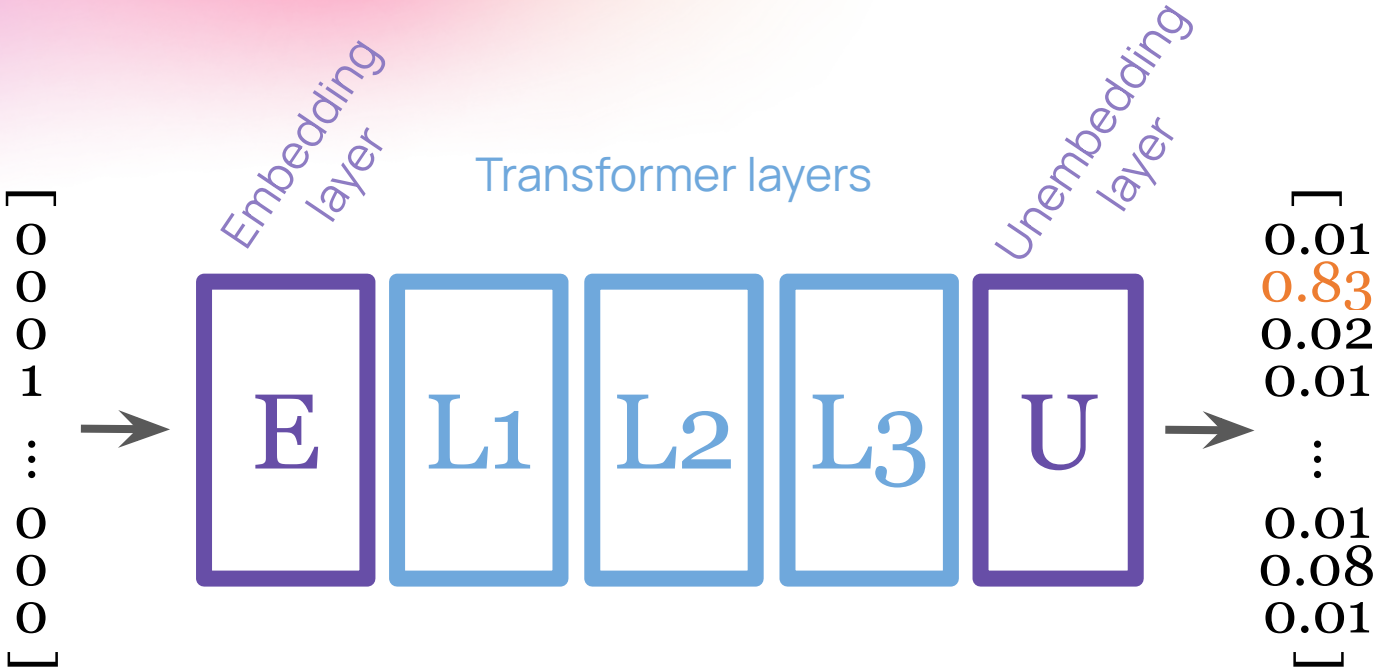
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

The big dog was chasing the

The | _big | _dog | _was | _chas | ing | _the

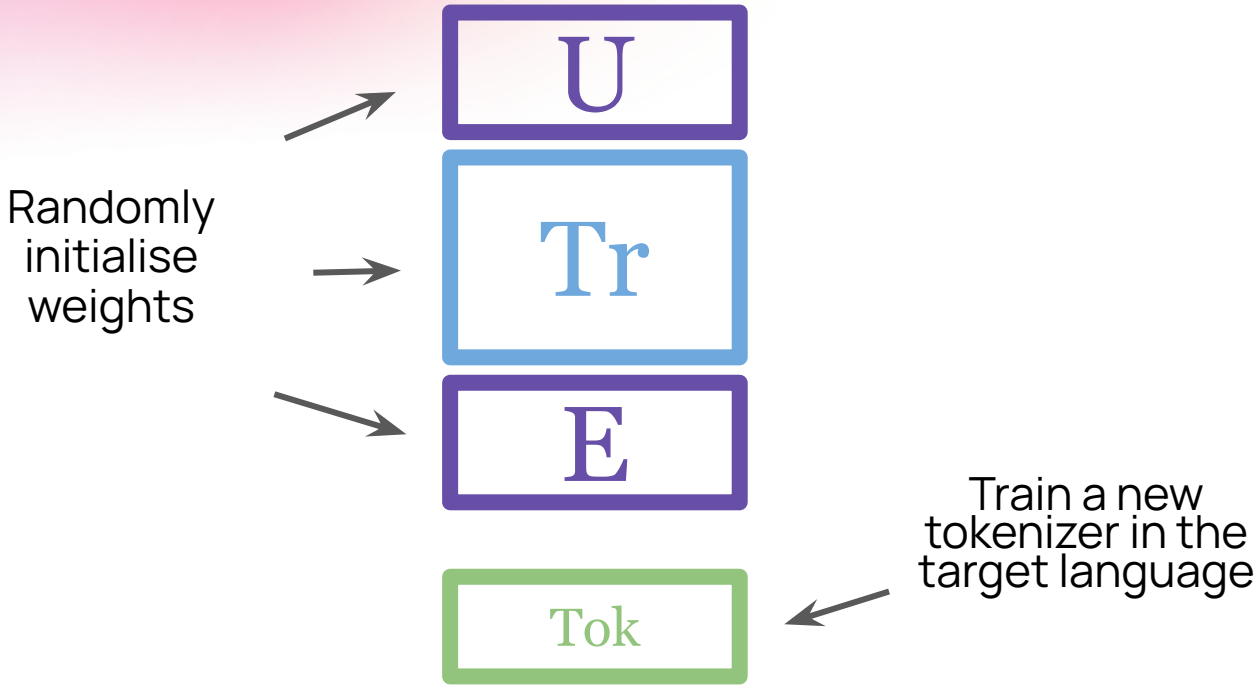
[0, 0, 0, ..., 1, 0, 0, 0]
[0, 1, 0, ..., 0, 0, 0, 0]
[1, 0, 0, ..., 0, 0, 0, 0]
[0, 0, 1, ..., 0, 0, 0, 0]
[0, 0, 0, ..., 0, 0, 0, 1]
[0, 0, 0, ..., 0, 1, 0, 0]
[0, 0, 0, ..., 0, 0, 1, 0]

Decoders



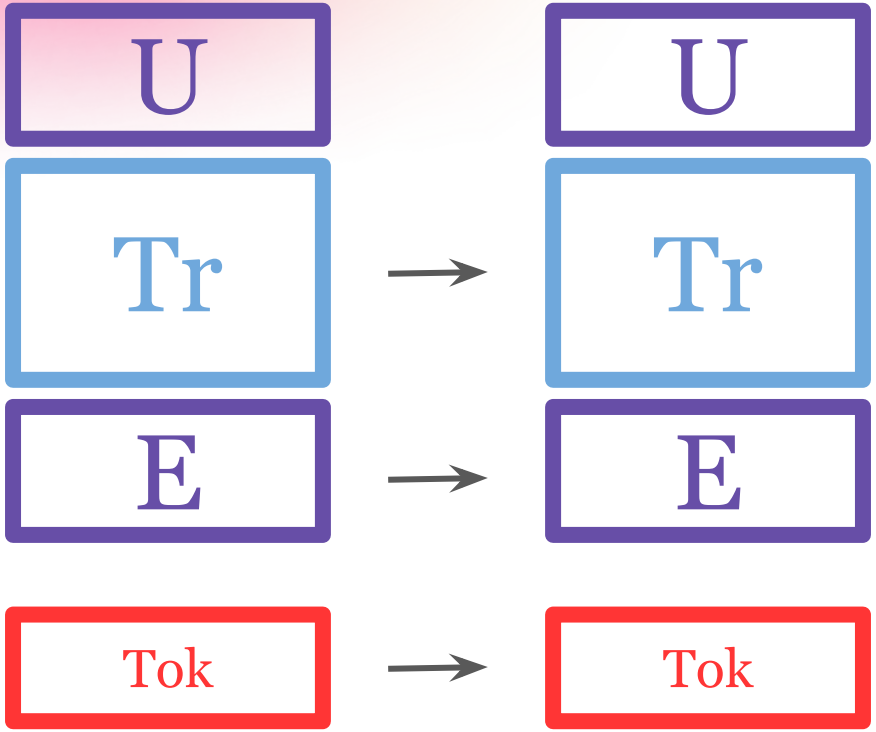
* This is an oversimplification.

From scratch



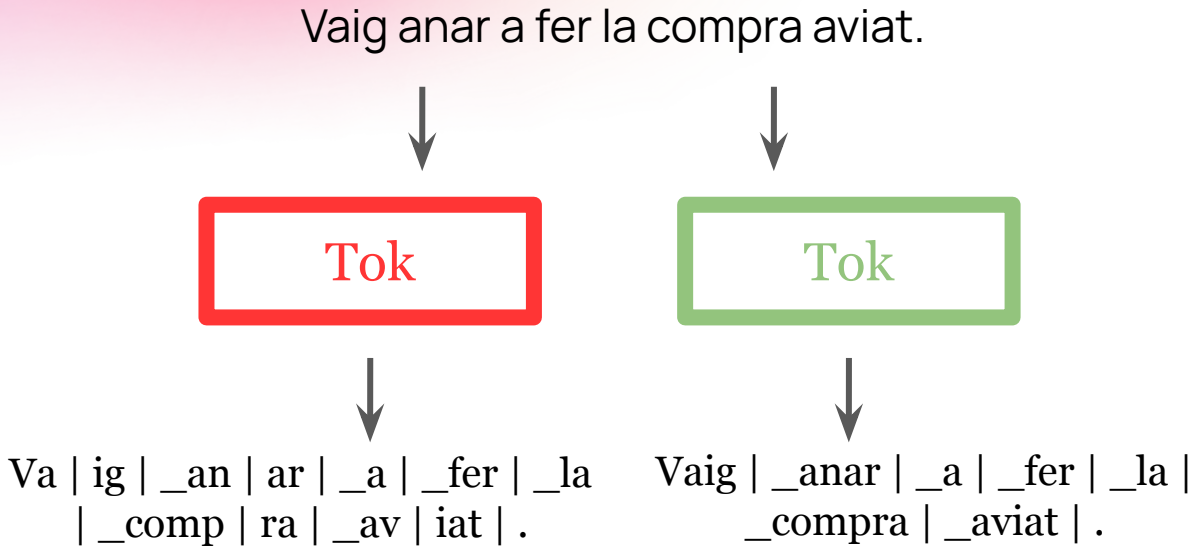
CPT

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



CPT

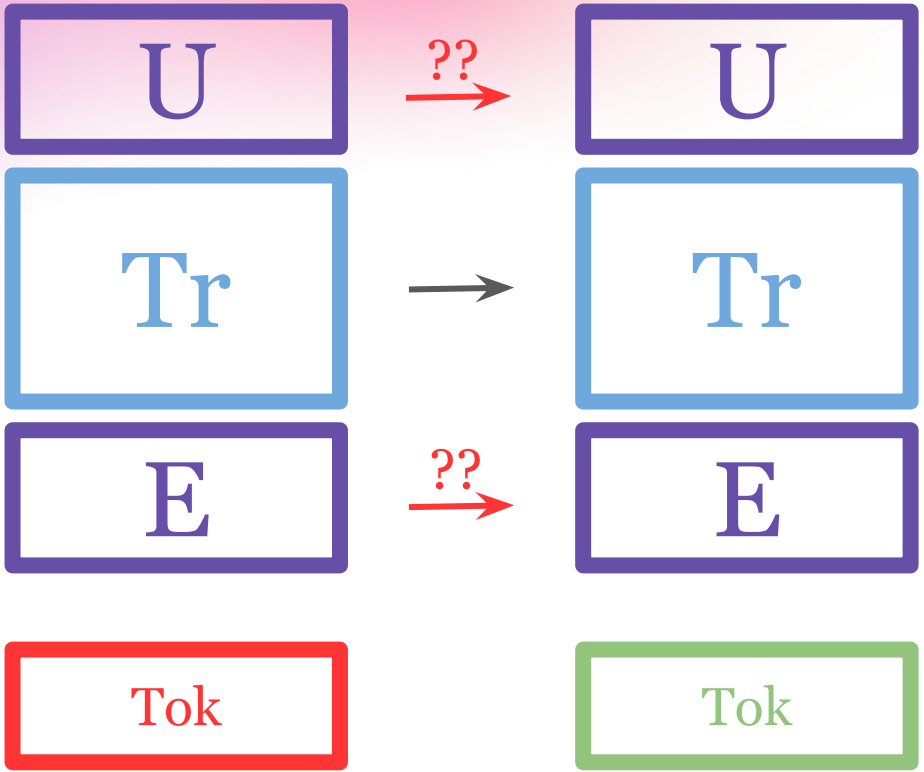
Barcelona Supercomputing Center
Centro Nacional de Supercomputación



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

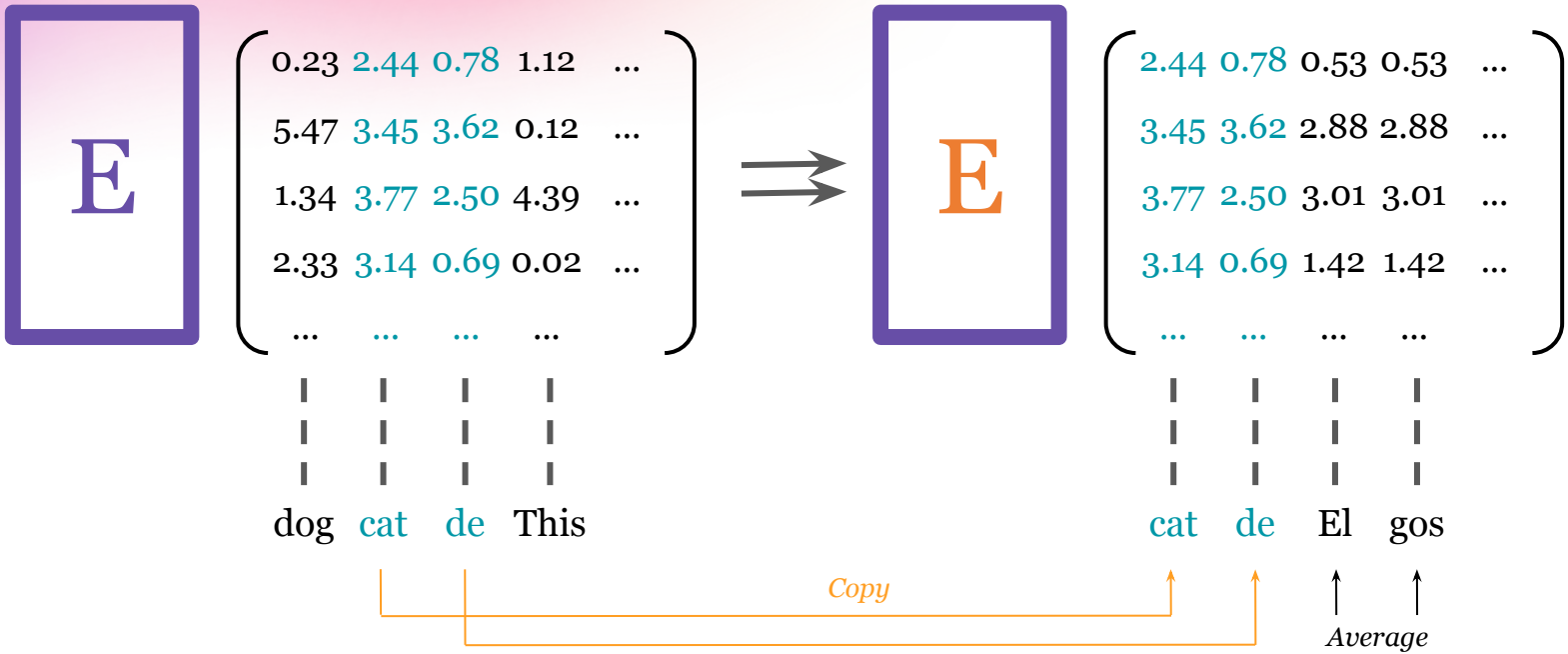
Vocabulary Adaptation

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



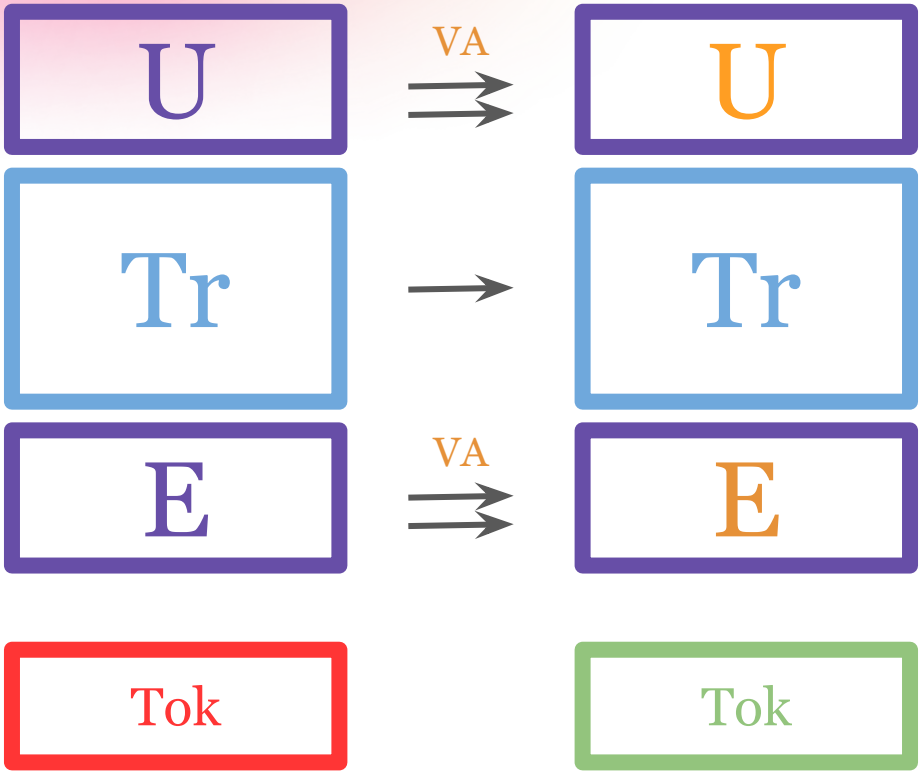
Train a new tokenizer in the target language

Vocabulary Adaptation

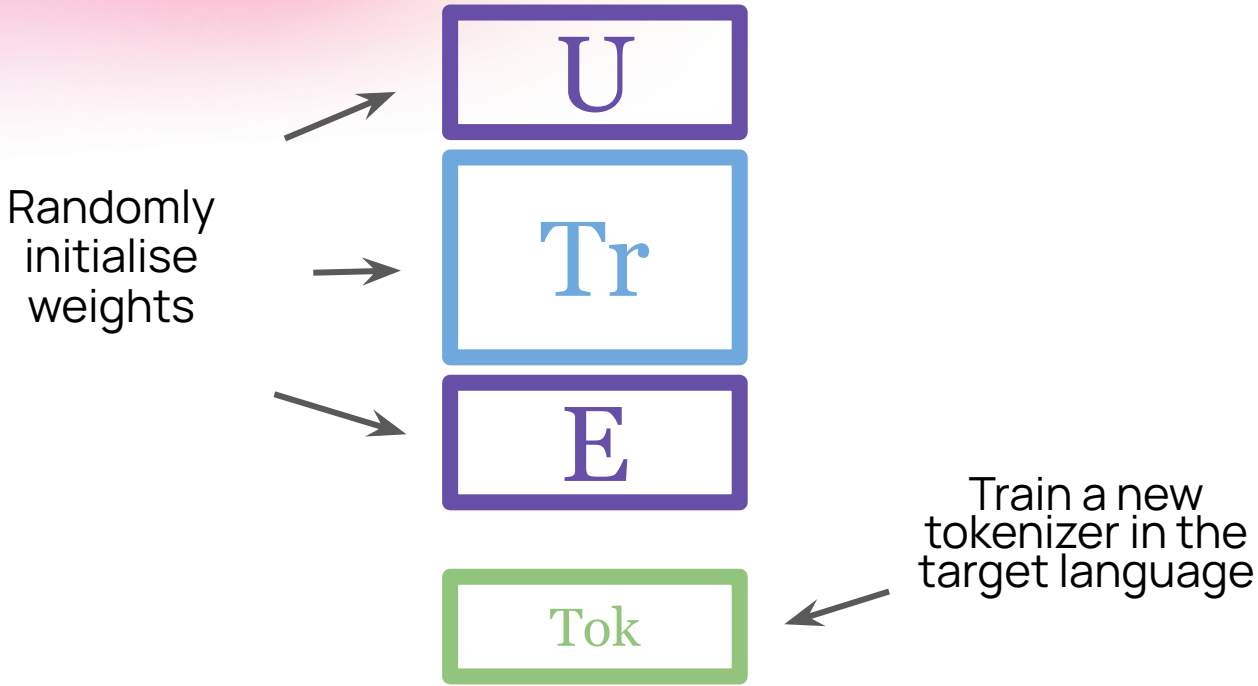


Vocabulary Adaptation

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



From scratch



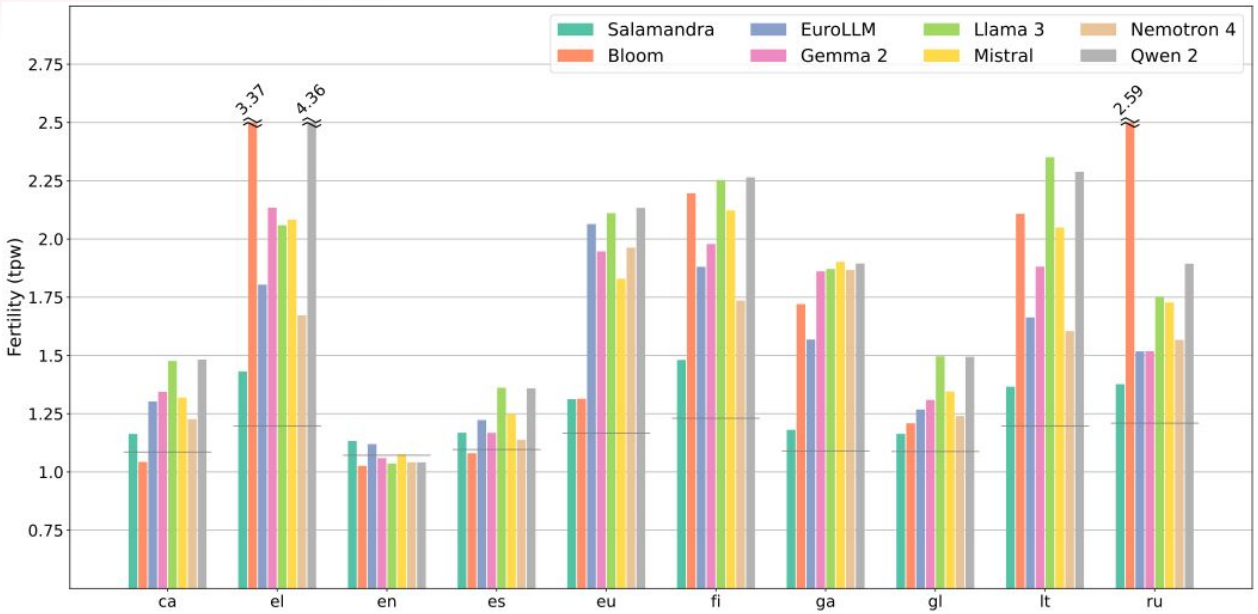
Tokenizer

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

SentencePiece implementation of **Byte-Pair Encoding** (BPE) algorithm.

256,000 multilingual tokens. 100 reserved for adaptations.

Byte fallback, digit splitting, normalization + more in Tech Report.



Distribution

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Models trained using **NVIDIA NeMo Framework** from a random initialization of weights and **Adam optimizer**.

Checkpointing was performed every 2,000 or 5,000 steps.

Model	Nodes	GPUs	TP	PP	DP	MBS	GBS	Context	Batch Size	Tokens
2B	64	256	1	1	256	1	512	8,192	~4M	12.9T
7B	128	512	4	1	128	2	512	8,192	~4M	12.9T
40B	512	2,048	4	2	128	1	1,024	4,096	~4M	~9T

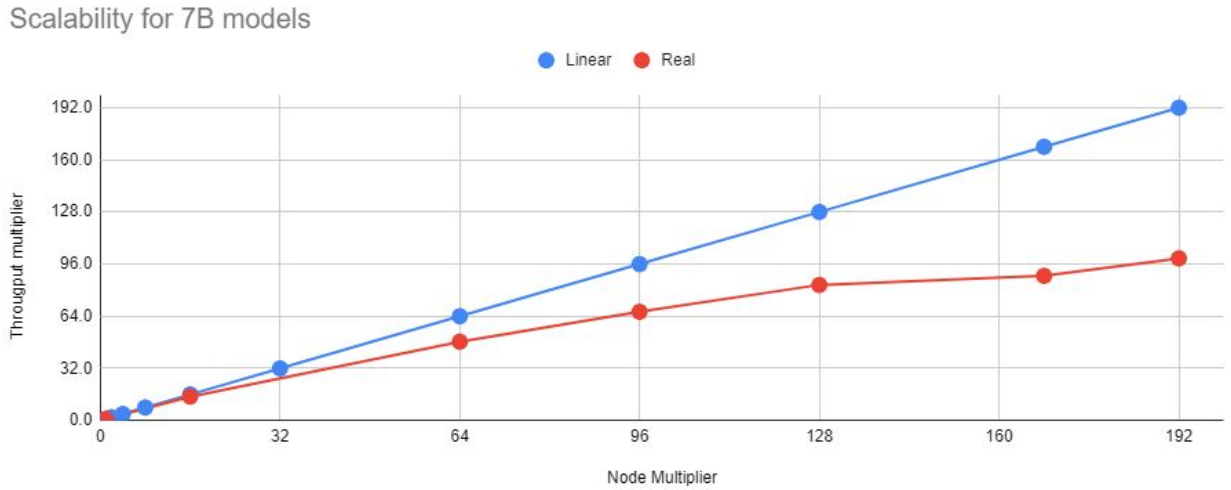
Extended to 32k

Scalability

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

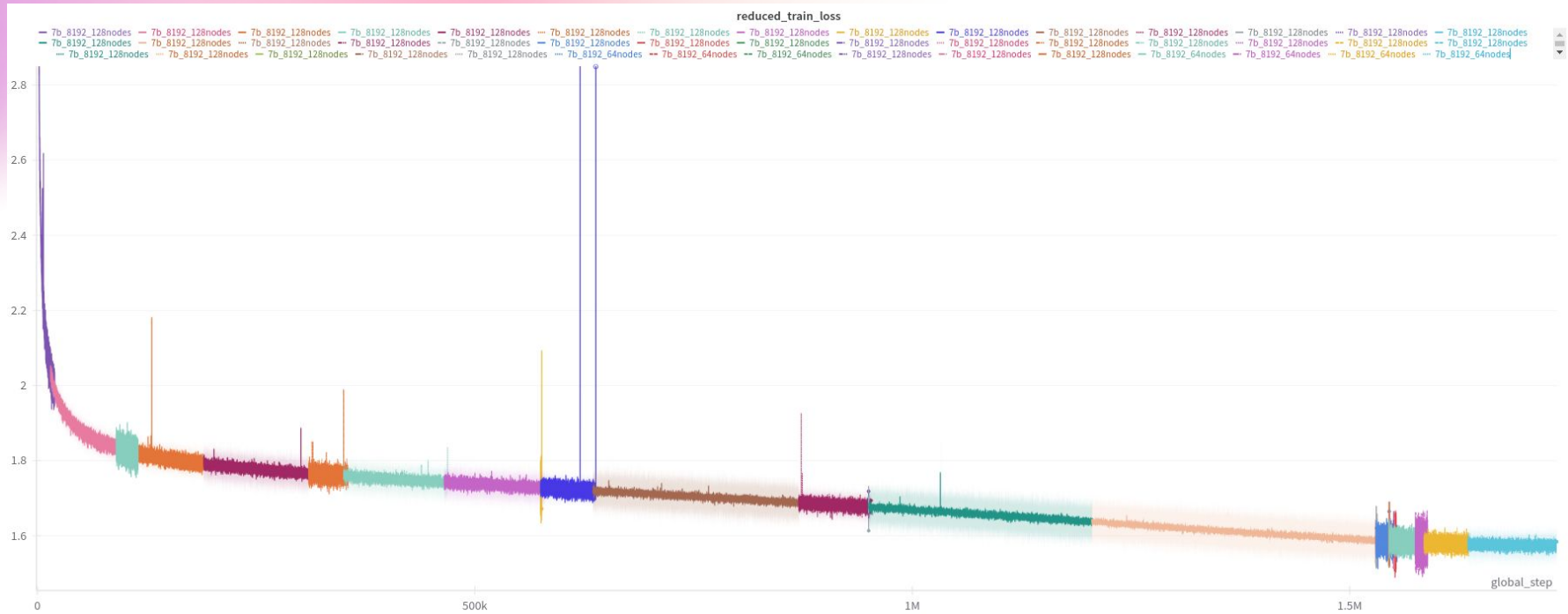
For the 7B model, going from 64 nodes to 128 resulted in a throughput increase of 86%, but going from 64 nodes to 192 this increase was reduced to 68%.

For the 40B model, scaling from 256 nodes to 512 yields an increase of 73%.



Training loss

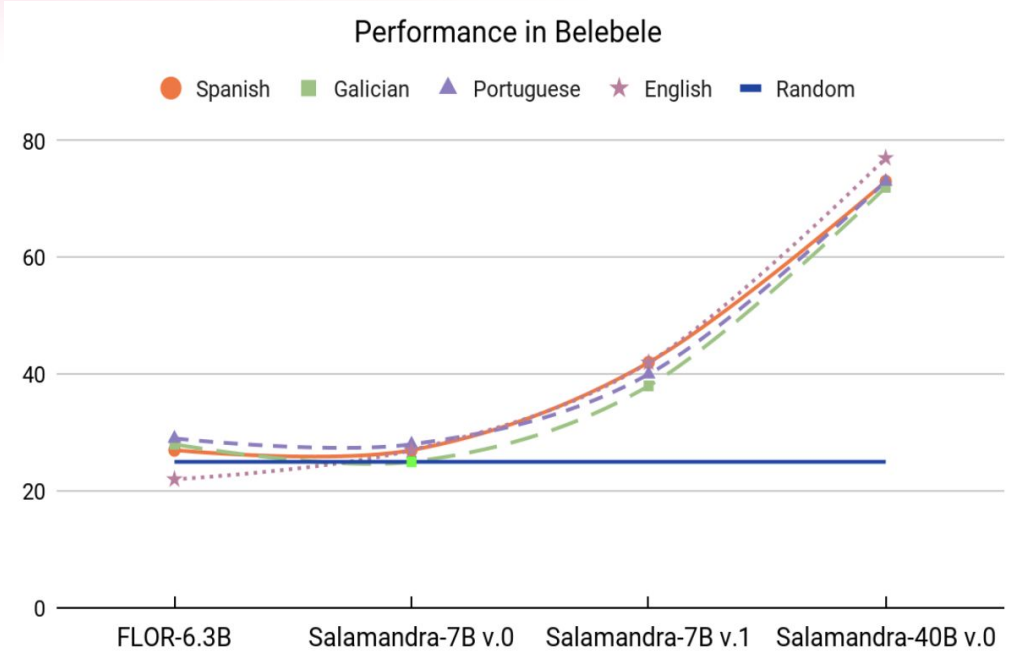
Barcelona Supercomputing Center
Centro Nacional de Supercomputación



Annealing

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Successful **targeted solutions** to models' limitations.



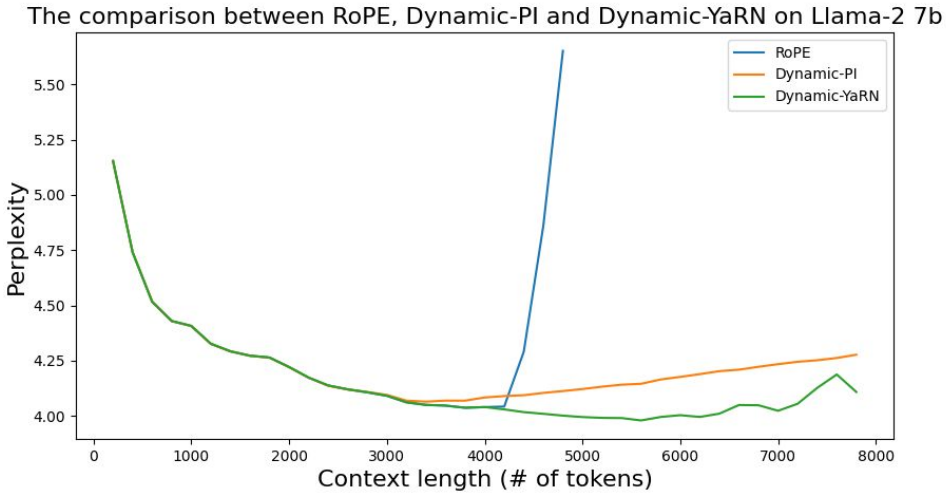
Just scale context?

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Decoder-only transformers (like GPT) have no hard architectural limit on the context window size.

But if we naively increase the context window (e.g., from 4K to 8K tokens), the model's performance collapses.

This isn't due to hardware, it's a failure of positional understanding.



YaRN: Efficient Context Window Extension of Large Language Models (Peng et al. 2023).

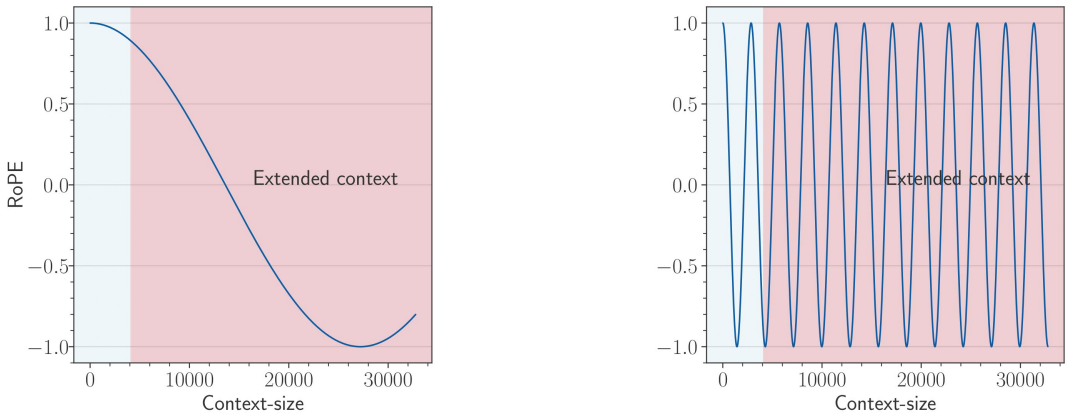
Waves of position

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Transformers use positional encodings to understand where tokens appear.

These encodings act like waves, mapping position to a signal the model reasons over.

Extending context length means we need to stretch or scale these waves, but doing this wrong leads to loss of resolution or generalization.



Base of RoPE Bounds Context Length (Men et al. 2024).

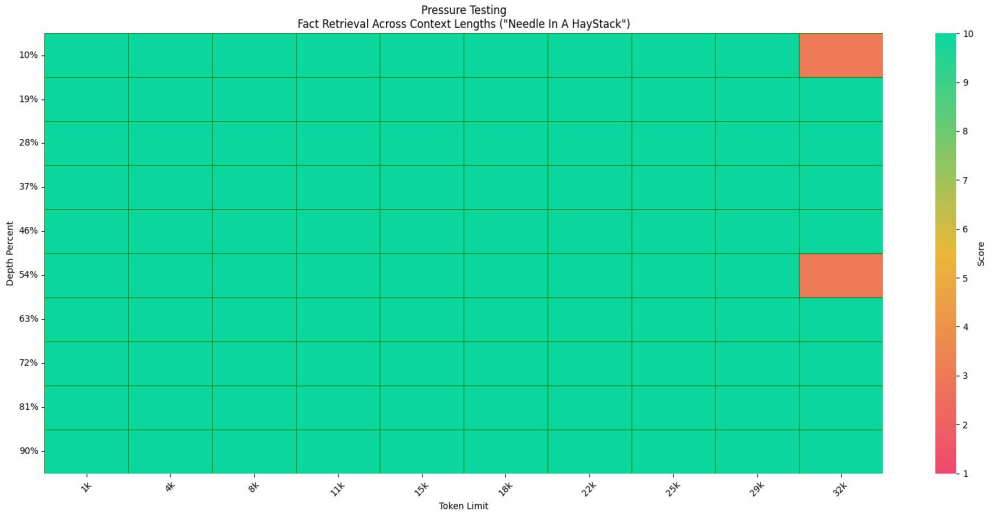
Evaluating extension

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Needle in a Haystack: A benchmark where a short "needle" is inserted into a long sequence. The task: recover the needle.

In standard models (trained on short contexts), performance drops sharply as the haystack gets longer.

Needle: The best thing to do in San Francisco is eat a sandwich and sit in Dolores Park on a sunny day.



Evaluating extension

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

RULER (Hsieh et al., 2024) benchmarks how model performance degrades across increasing context lengths on standard NLP tasks.

Long Context is still not solved. Even the best long-context models show gradual degradation.

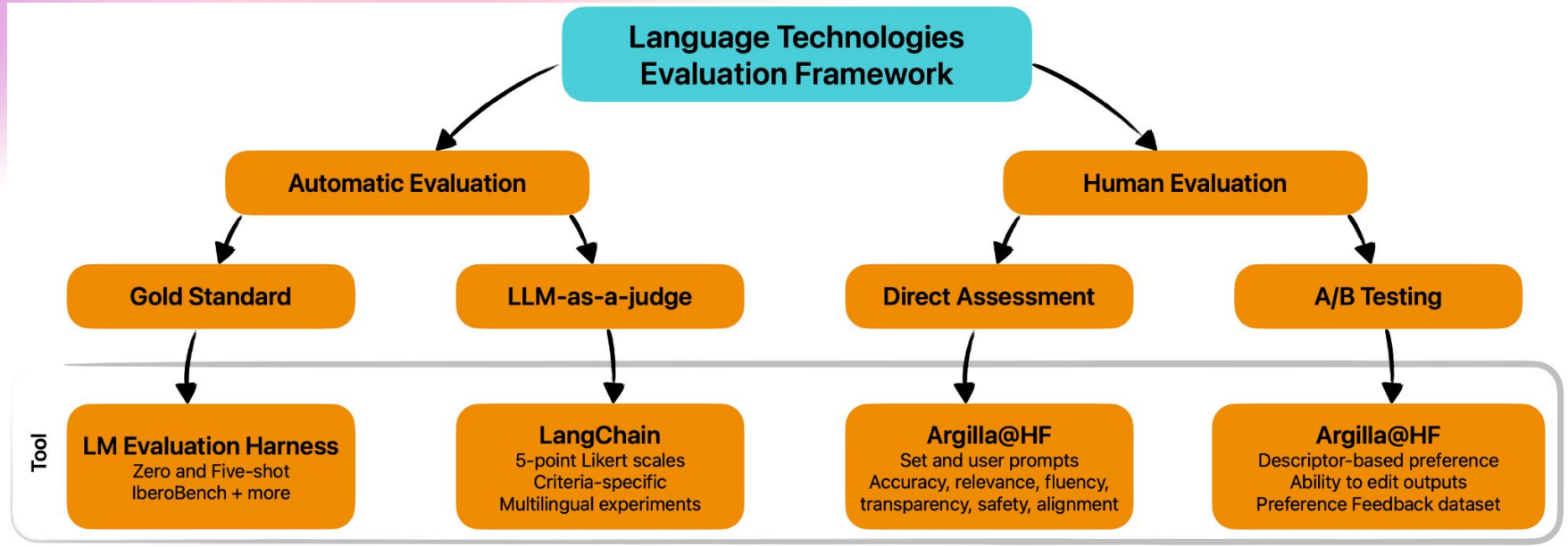
Models	Claimed Length	Effective Length	4K	8K	16K	32K	64K	128K	Avg.	wAvg. (inc)	wAvg. (dec)
Llama2 (7B)	4K	-	85.6								
Gemini-1.5-Pro	1M	>128K	96.7	95.8	96.0	95.9	95.9	94.4	95.8	95.5 (1st)	96.1 (1st)
GPT-4	128K	64K	96.6	96.3	95.2	93.2	87.0	81.2	91.6	89.0 (2nd)	94.1 (2nd)
Llama3.1 (70B)	128K	64K	96.5	95.8	95.4	94.8	88.4	66.6	89.6	85.5 (4th)	93.7 (3rd)
Qwen2 (72B)	128K	32K	96.9	96.1	94.9	94.1	79.8	53.7	85.9	79.6 (9th)	92.3 (4th)
Command-R-plus (104B)	128K	32K	95.6	95.2	94.2	92.0	84.3	63.1	87.4	82.7 (7th)	92.1 (5th)
GLM4 (9B)	1M	64K	94.7	92.8	92.1	89.9	86.7	83.1	89.9	88.0 (3rd)	91.7 (6th)
Llama3.1 (8B)	128K	32K	95.5	93.8	91.6	87.4	84.7	77.0	88.3	85.4 (5th)	91.3 (7th)
GradientAI/Llama3 (70B)	1M	16K	95.1	94.4	90.8	85.4	80.9	72.1	86.5	82.6 (8th)	90.3 (8th)
Mixtral-8x22B (39B/141B)	64K	32K	95.6	94.9	93.4	90.9	84.7	31.7	81.9	73.5 (11th)	90.3 (9th)
Yi (34B)	200K	32K	93.3	92.2	91.3	87.5	83.2	77.3	87.5	84.8 (6th)	90.1 (10th)
Phi3-medium (14B)	128K	32K	93.3	93.2	91.1	86.8	78.6	46.1	81.5	74.8 (10th)	88.3 (11th)
Mistral-v0.2 (7B)	32K	16K	93.6	91.2	87.2	75.4	49.0	13.8	68.4	55.6 (13th)	81.2 (12th)
LWM (7B)	1M	<4K	82.3	78.4	73.7	69.1	68.1	65.0	72.8	69.9 (12th)	75.7 (13th)
DBRX (36B/132B)	32K	8K	95.1	93.8	83.6	63.1	2.4	0.0	56.3	38.0 (14th)	74.7 (14th)
Together (7B)	32K	4K	88.2	81.1	69.4	63.0	0.0	0.0	50.3	33.8 (15th)	66.7 (15th)
LongChat (7B)	32K	<4K	84.7	79.9	70.8	59.3	0.0	0.0	49.1	33.1 (16th)	65.2 (16th)
LongAlpaca (13B)	32K	<4K	60.6	57.0	56.6	43.6	0.0	0.0	36.3	24.7 (17th)	47.9 (17th)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Evaluation

Framework

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Evaluation

Gold Standard

Benchmark

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Category	ca	es	eu	gl	pt
Commonsense Reasoning	copa_ca **xstorycloze_ca**	**copa_es** xstorycloze_es	**xcopa_eu** xstorycloze_eu		
Linguistic Acceptability	catcola	escola		galcola	
Math	**mgsm_direct_ca**	mgsm_direct_es	**mgsm_direct_eu**	**mgsm_direct_gl**	
NLI	xnli_ca wnli_ca teca	xnli_es wnli_es	xnli_eu **wnli_eu** qnli_eu		assin_entailment
Paraphrasing	parafraseja **paws_ca**	paws_es		**parafrases_gl** **paws_gl**	assin_paraphrase
QA	**arc_ca** catalanqa coqcat **openbookqa_ca** **piqa_ca** **siqa_ca** xquad_ca	xquad_es **openbookqa_es**	**piqa_eu** eus_exams eus_proficiency eus_trivia	**openbookqa_gl**	
Reading Comprehension	belebele_cat_Latn	belebele_spa_Latn	belebele_eus_Latn eus_reading	**belebele_glg_Latn**	belebele_por_Latn
Summarization	cabreu	xlsum_es		**summarization_gl**	
Translation / Adaptation	flores_ca **phrases_va**	flores_es **phrases_es**	flores_eu	flores_gl	flores_pt
Truthfulness	veritasqa_ca	veritasqa_es		veritasqa_gl **truthfulqa_gl**	

Table 1: Tasks included in IberoBench. Newly introduced datasets are marked in bold.

IberoBench: A Benchmark for LLM Evaluation in Iberian Languages

Gold Standard

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Odd labels in datasets

Subset (2)
ARC-Challenge · 2.59k rows

Split (3)
test · 1.17k rows

Search this dataset

id string · *lengths* 8 22	question string · *lengths* 13 831	choices sequence	answerKey string · *classes* 8 values
MCAS_2010_8_12016	Company X makes 100 custom buses each year. Company Y makes 10,000 of one type of bus each year. Which of the following is the most likely reason a customer…	{ "text": ["to keep the cost of the bus low", "to ensure that the bus will be easy to replace", "to provide ideas about how the bus will be built", "to…	C
Mercury_SC_400324	Students go on a class field trip to collect insects. Which of these is an important safety rule for the students to follow?	{ "text": ["Pick up insects with bare hands.", "Stay with a classmate at all times.", "Reach under large rocks without looking first.", "Work away from th…	B
Mercury_SC_LBS10662	A hot rock is dropped into a pail of cool water. Heat energy transferred from the rock to the water by	{ "text": ["boiling.", "evaporation.", "conduction.", "radiation."], "label": ["A", "B", "C", "D"] }	C
VASoL_2009_3_8	Which of these is needed in all stages of the butterfly's life cycle?	{ "text": ["Wings", "Eyes", "Soil", "Air"], "label": ["A", "B", "C", "D"] }	D
Mercury_SC_401185	Trail mix is a type of snack. One kind of trail mix is made of raisins, chocolate pieces, peanuts, and sunflower seeds. Which statement describes why…	{ "text": ["The mix contains four ingredients.", "The different ingredients clump together.", "Each component maintains its original properties.",…	C
NYSEDREGENTS_2015_8_29	Which change is the best example of a physical change?	{ "text": ["a cookie baking", "paper burning", "ice cream melting", "a nail rusting"], "label": ["1", "2", "3", "4"] }	3
Mercury_7234378	At which type of margin does mountain building produce mountains composed entirely of preexisting crust material?	{ "text": ["divergent boundary between two oceanic plates", "divergent boundary between an oceanic plate and a continental plate", "convergent…	D

Gold Standard

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Errors in datasets


```
{
    "id":"CSZ_2009_8_CSZ30651",
    "question":{
        "stem":"Which of the following compounds is most likely to be part of living organisms?",
        "choices":[
            {"text":"C?H??O?","label":"A"},
            {"text":"BF?","label":"B"},
            {"text":"MoCl?","label":"C"},
            {"text":"CsI","label":"D"}]},
    "answerKey":"A"
}
```

```
{
    "id":"Mercury_400198",
    "question":{
        "stem":"What is the dependent variable in this experiment?",
        "choices":[
            {"text":"type of soil used","label":"A"},
            {"text":"amount of water used","label":"B"},
            {"text":"length of sunlight exposure","label":"C"},
            {"text":"growth of the bean plants","label":"D"}]},
    "answerKey":"D"
}
```

Gold Standard

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Lack of prompt-awareness and pre-processing

lm-evaluation-harness / lm_eval / tasks / xnli / **xnli_en.yaml**

lintangsutawika update

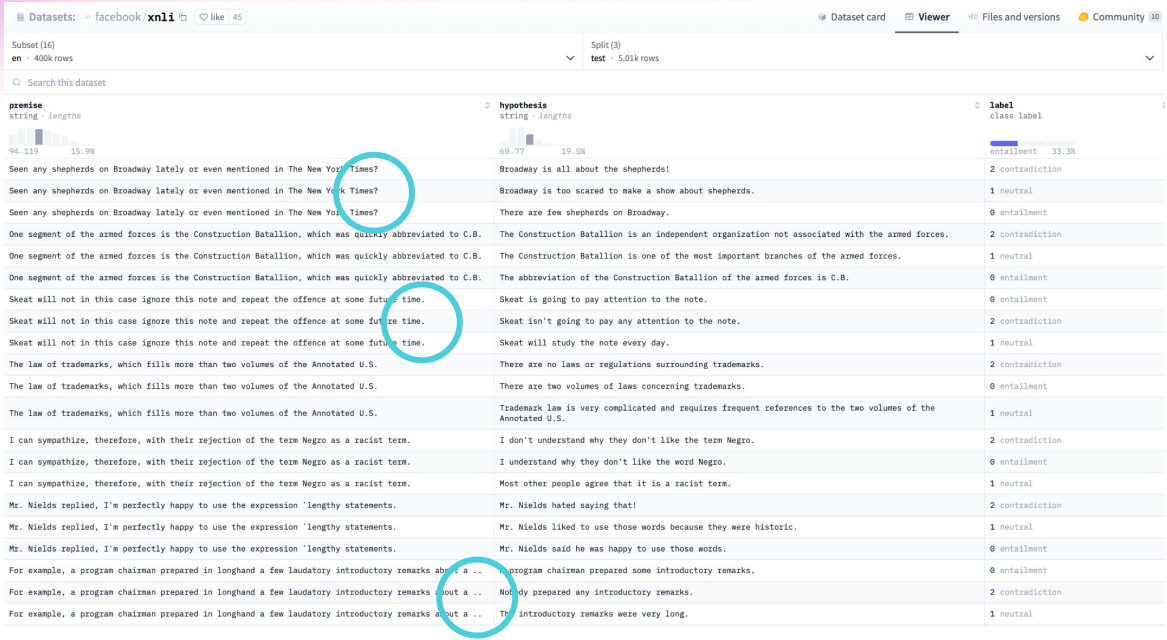
Code | Blame | 7 lines (7 loc) · 230 Bytes ·

```
1  # Generated by utils.py
2  dataset_name: en
3  doc_to_choice: '{{[premise+", right? Yes, "+hypothesis,premise+", right? Also, "+hypothesis,premise+",
4    right? No, "+hypothesis]}}'
5  doc_to_text: ''
6  include: xnli_common_yaml
7  task: xnli_en
```

Gold Standard

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Lack of prompt-awareness and pre-processing



Datasets: facebook/xnli | like 45 | Dataset card | Viewer | Files and versions | Community 10

Subset (16): en · 400k rows | Split (3): test · 5.01k rows

Search this dataset

premise string · lengths 94..119 15.9%	hypothesis string · lengths 60..77 19.5%	label class label entailment 33.3%
Seen any shepherds on Broadway lately or even mentioned in The New York Times?	Broadway is all about the shepherds!	2 contradiction
Seen any shepherds on Broadway lately or even mentioned in The New York Times?	Broadway is too scared to make a show about shepherds.	1 neutral
Seen any shepherds on Broadway lately or even mentioned in The New York Times?	There are few shepherds on Broadway.	0 entailment
One segment of the armed forces is the Construction Batallion, which was quickly abbreviated to C.B.	The Construction Batallion is an independent organization not associated with the armed forces.	2 contradiction
One segment of the armed forces is the Construction Batallion, which was quickly abbreviated to C.B.	The Construction Batallion is one of the most important branches of the armed forces.	1 neutral
One segment of the armed forces is the Construction Batallion, which was quickly abbreviated to C.B.	The abbreviation of the Construction Batallion of the armed forces is C.B.	0 entailment
Skeat will not in this case ignore this note and repeat the offence at some future time.	Skeat is going to pay attention to the note.	0 entailment
Skeat will not in this case ignore this note and repeat the offence at some future time.	Skeat isn't going to pay any attention to the note.	2 contradiction
Skeat will not in this case ignore this note and repeat the offence at some future time.	Skeat will study the note every day.	1 neutral
The law of trademarks, which fills more than two volumes of the Annotated U.S.	There are no laws or regulations surrounding trademarks.	2 contradiction
The law of trademarks, which fills more than two volumes of the Annotated U.S.	There are two volumes of laws concerning trademarks.	0 entailment
The law of trademarks, which fills more than two volumes of the Annotated U.S.	Trademark law is very complicated and requires frequent references to the two volumes of the Annotated U.S.	1 neutral
I can sympathize, therefore, with their rejection of the term Negro as a racist term.	I don't understand why they don't like the term Negro.	2 contradiction
I can sympathize, therefore, with their rejection of the term Negro as a racist term.	I understand why they don't like the word Negro.	0 entailment
I can sympathize, therefore, with their rejection of the term Negro as a racist term.	Most other people agree that it is a racist term.	1 neutral
Mr. Nields replied, I'm perfectly happy to use the expression 'lengthy statements.	Mr. Nields hated saying that!	2 contradiction
Mr. Nields replied, I'm perfectly happy to use the expression 'lengthy statements.	Mr. Nields liked to use those words because they were historic.	1 neutral
Mr. Nields replied, I'm perfectly happy to use the expression 'lengthy statements.	Mr. Nields said he was happy to use those words.	0 entailment
For example, a program chairman prepared in longhand a few laudatory introductory remarks about a ..	A program chairman prepared some introductory remarks.	0 entailment
For example, a program chairman prepared in longhand a few laudatory introductory remarks about a ..	Nobody prepared any introductory remarks.	2 contradiction
For example, a program chairman prepared in longhand a few laudatory introductory remarks about a ..	The introductory remarks were very long.	1 neutral

Gold Standard

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

```
'{{[[premise+", right? Yes, "+hypothesis,premise+", right? Also, "+hypothesis,premise+", right? No, "+hypothesis]]}'
```

“Seen any shepherds on Broadway lately or even mentioned in The New York Times?, right? Yes, There are few shepherds on Broadway.”

“For example, a program chairman prepared in longhand a few laudatory introductory remarks about a .., right? Also, The introductory remarks were very long.”

“Pulse-tone is not a technical term., right? No, The official technical manual states that pulse-tone is the correct term in this case.”

Gold Standard

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Pre-processing is a must for uncontrolled datasets.

lm-evaluation-harness / lm_eval / tasks / spanish_bench / utils.py

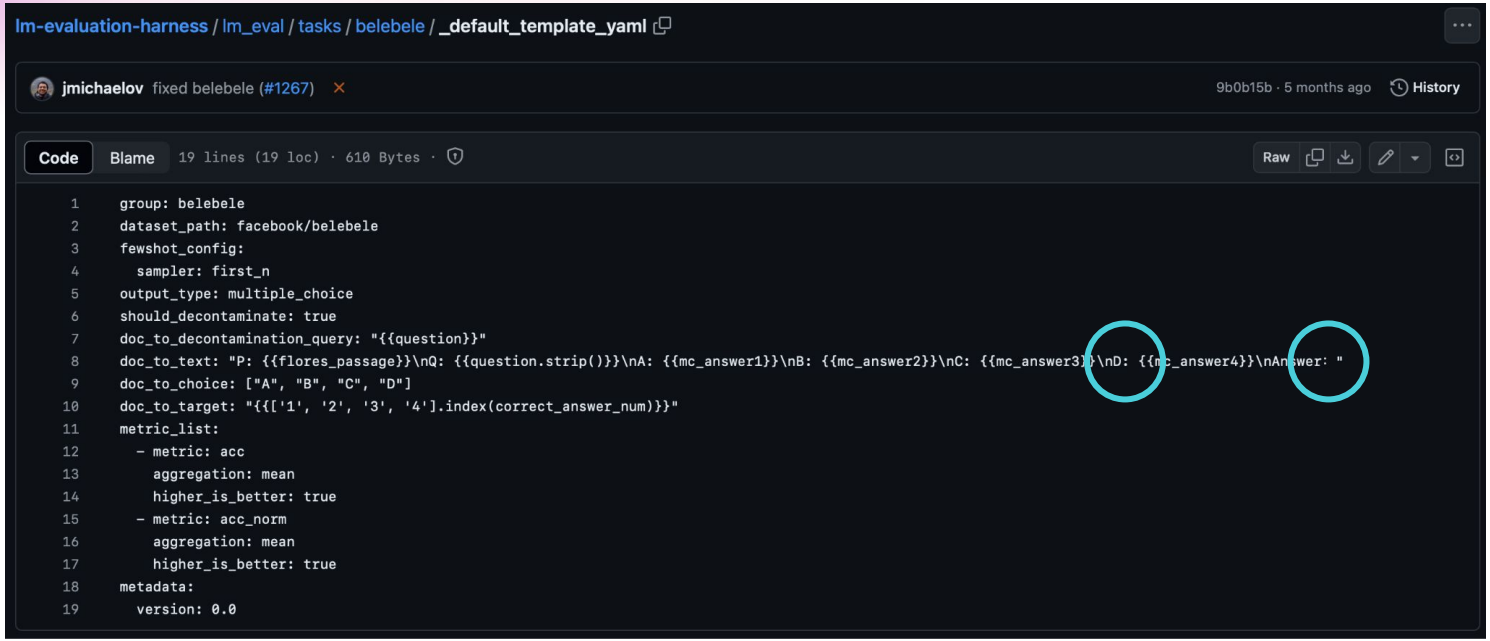
Code | Blame | Executable File · 83 lines (70 loc) · 3.02 KB | Your organization can pay for GitHub Copilot | Raw

```python
import re
from itertools import product
import evaluate
import transformers.data.metrics.squad_metrics as squad_metrics
from lm_eval.utils import general_detokenize


def lowercase_first_letter(text):
        return text[0].lower() + text[1:]

def process_doc_nli(dataset):
    def process_fn(doc):
        # Detokenize(remove extra whitespaces)
        doc["premise"] = general_detokenize(doc["premise"]).strip()
        doc["hypothesis"] = general_detokenize(doc["hypothesis"]).strip()
            # Remove last punctuation mark in the premise
        doc["premise"] = doc["premise"][:-1] if doc["premise"].endswith((".", ",", "!", "?")) else doc["premise"]
            # Lowercase the first letter in the hypothesis
        doc["hypothesis"] = lowercase_first_letter(doc["hypothesis"])
            # Ensure that the hypothesis ends with a dot
        doc["hypothesis"] = (doc["hypothesis"] + ".") if not doc["hypothesis"].endswith(".") else doc["hypothesis"]
        return doc
    return dataset.map(process_fn)
```

Gold Standard

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Non-UTF-8 characters in prompts



Gold Standard

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Non-UTF-8 characters in prompts

```
{{mc_answer2}}\nC: {{mc_answer3}}\nD: {{mc_answer4}}\nAnswer: "
```

```
>>> flor6b_tokenizer.tokenize(": ")
['ï', '¼', 'ļ']
```

“Binary digits are also referred to as what? A: Bits (...) D: Formsï¼ļ”

Gold Standard

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Type of task	Metric
Commonsense Reasoning	Accuracy
Linguistic Acceptability	Matthews Correlation Coefficient
Math	Exact Match
Natural Language Inference	Accuracy
Paraphrasing	F1 Score
Question Answering	Acc. for binary or MC tasks + F1 for generative tasks
Reading Comprehension	Accuracy
Summarization	ROUGE1 or BLEU
Translation	BLEU
Truthfulness	Accuracy + BLEU

Gold Standard

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Differences in runs using tensor parallelism.

Run Name	Created	Duration	Experiment Name	score	model
arc_challenge	16 hours ago	117ms	fourth_epoch_mix1_no-tp	49.66	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo
arc_challenge	1 day ago	30ms	fourth_epoch_mix1	50.34	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo

Run Name	Created	Duration	Experiment Name	score	model
teca	21 hours ago	117ms	fourth_epoch_mix1_no-tp	51.72	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo
teca	1 day ago	27ms	fourth_epoch_mix1	48.84	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo

Run Name	Created	Duration	Experiment Name	score	model
cabreu_extreme	20 hours ago	151ms	fourth_epoch_mix1_no-tp	41.94	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo
cabreu_extreme	1 day ago	38ms	fourth_epoch_mix1	40.72	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo

Run Name	Created	Duration	Experiment Name	score	model
catalanqa	19 hours ago	118ms	fourth_epoch_mix1_no-tp	82.23	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo
catalanqa	20 hours ago	34ms	fourth_epoch_mix1	81.7	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo

Gold Standard

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Differences in runs using tensor parallelism.

Run Name	Created	Duration	Experiment Name	score	model
catalanqa	19 hours ago	118ms	fourth_epoch_mix1_no-tp	82.23	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo
catalanqa	20 hours ago	34ms	fourth_epoch_mix1	81.7	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo

Run Name	Created	Duration	Experiment Name	score	model
catalanqa	19 hours ago	118ms	fourth_epoch_mix1_no-tp	82.23	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo
catalanqa	20 hours ago	34ms	fourth_epoch_mix1	81.7	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo
catalanqa	20 hours ago	39ms	fourth_epoch_mix1	81.7	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo
catalanqa	20 hours ago	36ms	fourth_epoch_mix1	81.7	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo
catalanqa	20 hours ago	43ms	fourth_epoch_mix1	81.7	fourth_epoch_bsc_7b_restart_lr3e-5_lr3e-6_step52500_decompressed_nemo

Gold Standard

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

+ more issues:

Double or no spaces when concatenating in prompt.

Limitations with metrics for open-ended generative tasks.

Multiple gold standards for some tasks.

Errors in datasets (both content and form).

1% variability across evaluations with same setup in a SingularityCE.

First-token generation bias.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Evaluation

LLM-as-a-Judge

LLM-as-a-judge

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Prompts and criteria for LLM-Evaluator need to be in English.

```
SYSTEM_MESSAGE_PROMETHEUS_2 = "You are a fair judge assistant tasked with providing clear, objective feedback based on \
specific criteria, ensuring each assessment reflects the absolute standards set for performance."
SCORING_TEMPLATE_WITH_REFERENCE_1_TO_5 = ChatPromptTemplate.from_messages(
    [
        ("system", SYSTEM_MESSAGE_PROMETHEUS_2),
        (
            "human",
            "###Task Description:\nAn instruction (might include an Input inside it), a response to evaluate, and a score rubric representing a \
            evaluation criteria are given.\
            \n1. Write detailed feedback that assess the quality of the response strictly based on the given score rubric, not evaluating in general.\
            \n2. After writing feedback, write a score that is an integer between 1 and 5. You should refer to the score rubric.\
            \n3. The output format should look as follows: 'Feedback: (write feedback for criteria) [RESULT] (an integer number between 1 and 5)'\
            \n4. Please do not generate any other opening, closing, and explanations.\
            \n###The instruction to evaluate: {input}\
            \n###Response to evaluate: {prediction}\
            \n###Score Rubrics:\n{criteria}\
            \n###Feedback:"
        ),
    ]
)
```

LLM-as-a-judge

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Prompts and criteria for LLM-Evaluator need to be in English.

```
conciseness_criteria = {
    "conciseness": """
Score 1: The answer contains too much information and includes unnecessary details that could confuse the reader or detract from the main point.
Score 2: The answer is overly long-winded and could be condensed to convey the same information more efficiently.
Score 3: The answer is generally brief and to the point, but there are still some words or phrases that could be removed to improve clarity and brevity.
Score 4: The answer is clear, succinct, and effectively communicates the necessary information without unnecessary elaboration.
Score 5: The answer is exceptionally brief, yet it effectively conveys all the required information with clarity and precision."""
}

relevance_criteria = {
    "relevance": """
Score 1: The answer is not relevant to the question and does not address the topic at hand.
Score 2: The answer is somewhat relevant but lacks focus and may include tangential information.
Score 3: The answer is relevant but could be more focused on addressing the specific question or topic.
Score 4: The answer is highly relevant and directly addresses the question or topic with appropriate detail.
Score 5: The answer is perfectly relevant, providing precise and comprehensive information directly related to the question or topic."""}

correctness_criteria = {
    "correctness": """
Score 1: The answer contains significant inaccuracies or errors that mislead the reader.
Score 2: The answer has several inaccuracies or errors that affect its credibility and may confuse the reader.
Score 3: The answer is generally correct but contains minor inaccuracies or errors that could be corrected for improved accuracy.
Score 4: The answer is mostly correct, with few inaccuracies or errors that do not significantly impact its overall correctness.
Score 5: The answer is entirely correct, providing accurate and reliable information without any inaccuracies or errors."""
}
```

LLM-as-a-judge

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Beware of the FT prompt and the output parser.

```
from langchain_core.prompts.chat import ChatPromptTemplate

SYSTEM_MESSAGE = "You are a helpful assistant."
SCORING_TEMPLATE_WITH_REFERENCE_1_TO_5_a = ChatPromptTemplate.from_messages(
    [
        ("system", SYSTEM_MESSAGE),
        (
            "human",
            "[Instruction]\nPlease act as an impartial judge \
and evaluate the quality of the response provided by an AI \
assistant to the user question displayed below. {criteria}"
            'Begin your evaluation \
by providing a short explanation. Be as objective as possible. \
After providing your explanation, you must rate the response on a scale of 1 to 5 \
by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".\n\n\
[Question]\n{input}[Actual Verified Answer]\n{reference}\n\n[The Start of Assistant\'s Answer]\n{prediction}\n\
[The End of Assistant\'s Answer].',
        ),
    ]
)
```

LLM-as-a-judge

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Beware of the FT prompt and the output parser.

```
SYSTEM_MESSAGE_PROMETHEUS_2 = "You are a fair judge assistant tasked with providing clear, objective feedback based on \
specific criteria, ensuring each assessment reflects the absolute standards set for performance."
SCORING_TEMPLATE_WITH_REFERENCE_1_TO_5_CONTEXT = ChatPromptTemplate.from_messages(
    [
        ("system", SYSTEM_MESSAGE_PROMETHEUS_2),
        (
            "human",
            "###Task Description:\nAn instruction (might include an Input inside it), a response to evaluate, and a score rubric representing a \
            evaluation criteria are given.\
            \n1. Write detailed feedback that assess the quality of the response strictly based on the given score rubric, not evaluating in general.\
            \n2. After writing feedback, write a score that is an integer between 1 and 5. You should refer to the score rubric.\
            \n3. The output format should look as follows: 'Feedback: (write feedback for criteria) [RESULT] (an integer number between 1 and 5)'\
            \n4. Please do not generate any other opening, closing, and explanations.\
            \n###The instruction to evaluate: {input}\
            \n###Verified correct answer: {reference}\
            \n###Response to evaluate: {prediction}\
            \n###Score Rubrics:\n{criteria}\
            \n###Feedback:"
        ),
    ]
)
```

LLM-as-a-judge

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

To try to fix this, we initially changed point 3 of the prompt to:

```
3. The output format should look as follows: 'Feedback: (write feedback for criteria) [[RESULT]] (an integer number between 1 and 5, for example, [[5]])'
```

Which led to answers like:

```
I do like this answer very much. [[RESULT]] 5
```

LLM-as-a-judge

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Let's change it to:

3. The output format should look as follows: 'Feedback: (write feedback for criteria) [[RESULT GOES HERE]] (an integer number between 1 and 5, for example, [[5]])'

This led to answers like:

I do like this answer very much. [[RESULT]] [[5]]

LLM-as-a-judge

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

```
 2    from langchain.schema.output_parser import BaseOutputParser
10    class PrometheusOutputParser(BaseOutputParser):
11
12        def __init__(self):
13            super().__init__()
14
15        def parse(self, output):
29            parts = output.split('[RESULT]')
30            assert len(parts) == 2
31            score = int(parts[-1].strip())
32
33            return {'score': score}
```

LLM-as-a-judge

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Prompt preference for each LLM-Evaluator.

Empty strings as best scores.

LLMs can distinguish themselves from other LLMs and humans, and there is a direct correlation between self-recognition capability and self-preference bias.

The “emergent ability” of Likert-scale granular understanding.

Reproducibility issues.

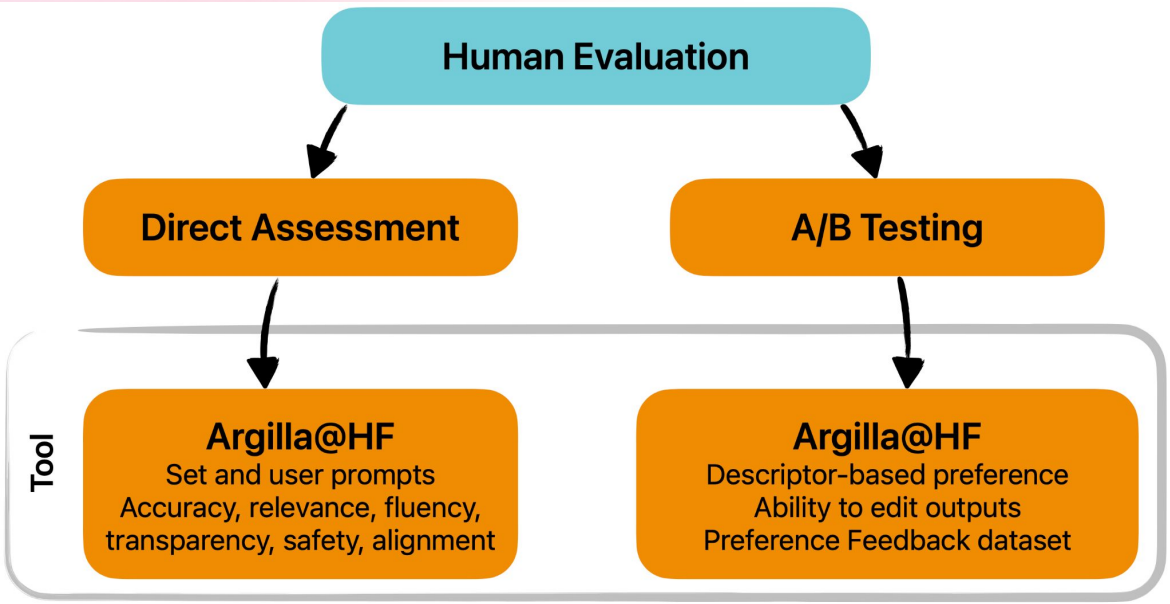
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Evaluation

Human Evaluation

Human Evaluation

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



Human Evaluation

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

High cost and challenging scalability.

Granularity of criteria and scales.

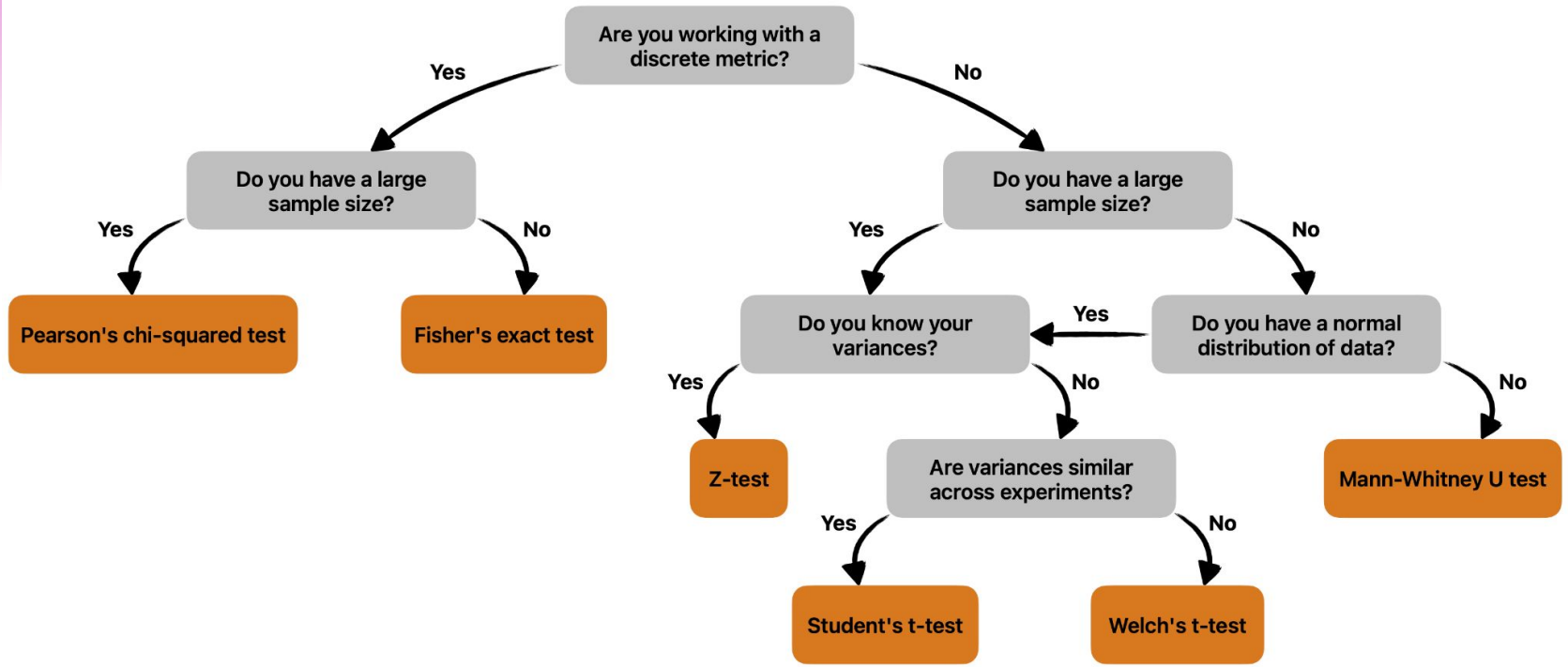
Overcoming human cognitive biases.

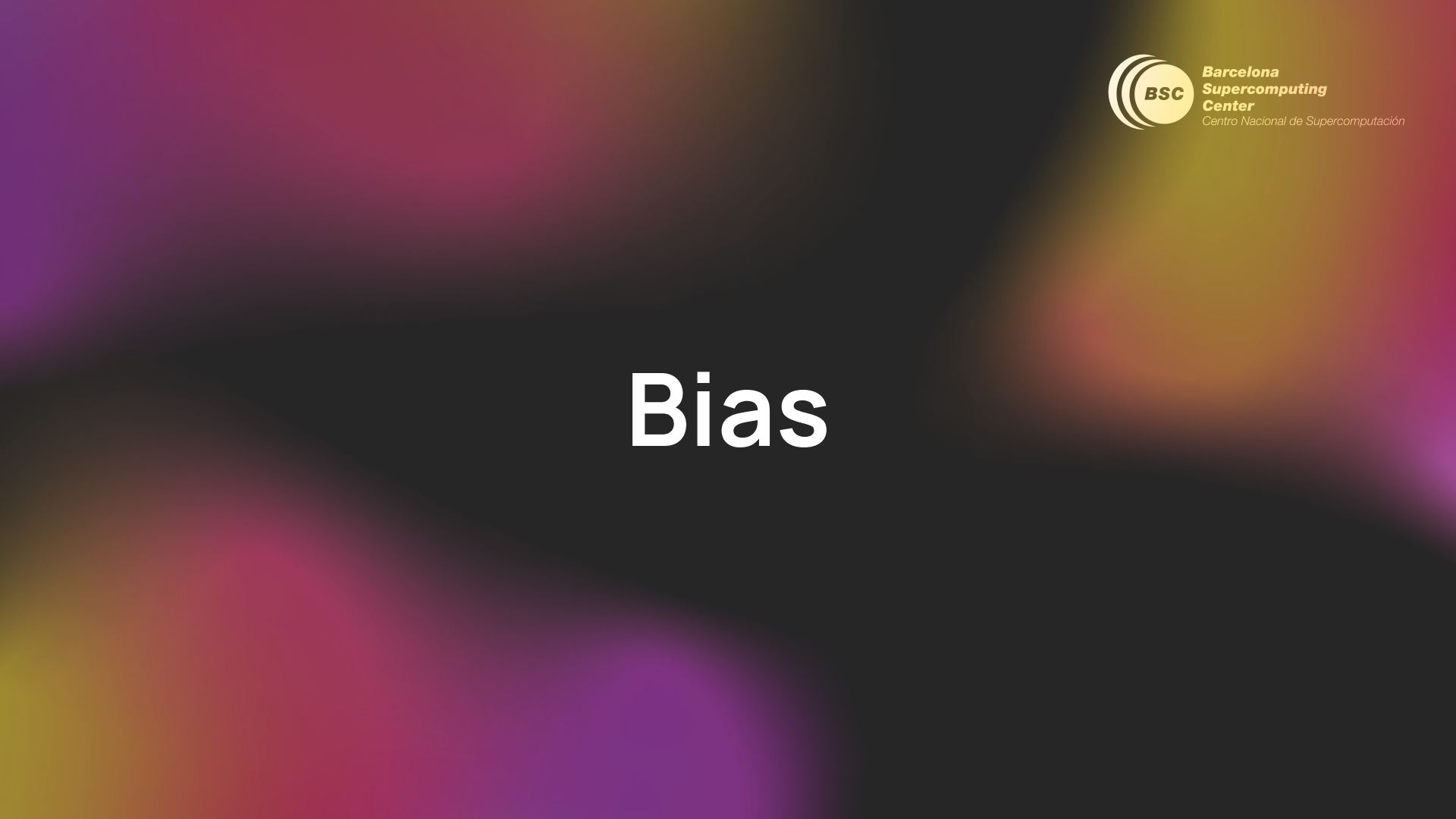
Trusting annotators.

Measuring reliability and using statistics.

Human Evaluation

Barcelona Supercomputing Center
Centro Nacional de Supercomputación




BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación
Bias

Concerns

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Social bias

Misrepresentation
Stereotypes
Derogatory language

Cognitive bias

Primacy/recency bias
Majority class bias
Common token bias

Challenges

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Varying definitions of “bias”, “toxicity”, “harmful content” etc.

No clear metrics for automatic evaluation.

Most benchmarks built around USA socio-cultural context.

Human annotation is challenging and IAA tends to be low.

Keyword bias affects dataset curation and evaluation.

Full “debiasing” is impossible; focus on *mitigation*.

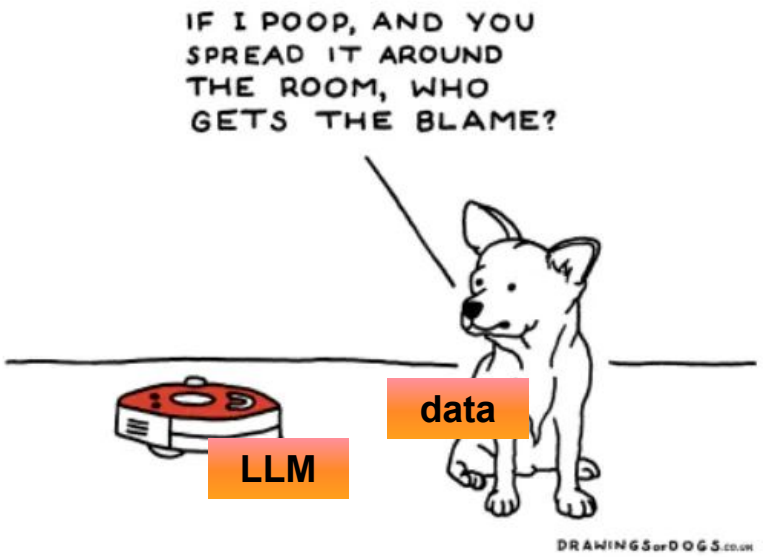
Identification approaches

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

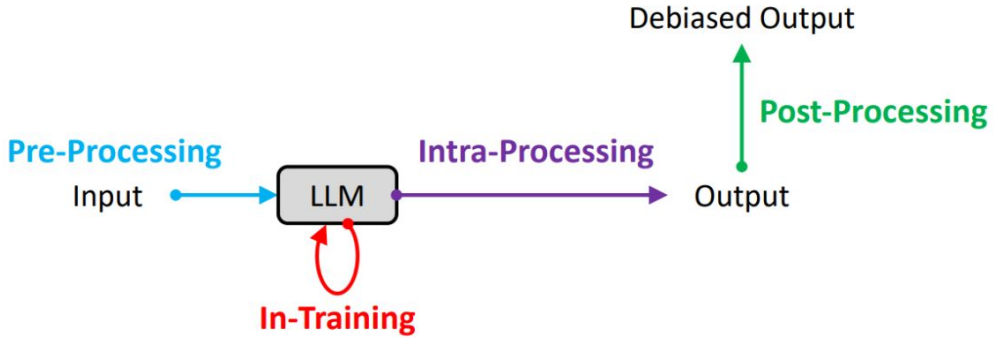
Bias can **come from the training data** and also be **modulated by** the LLM training/alignment **procedures**.

We work on:

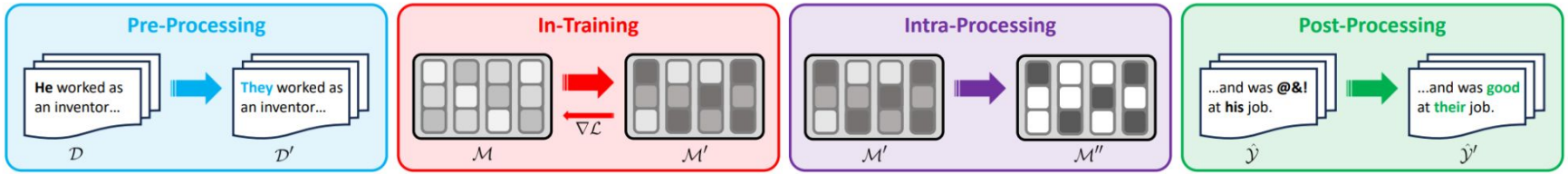
- Creating, adapting and/or translating benchmarks.
- Researching new/better ways of measuring bias in our data and model structure.



Mitigation approaches



(a)



Change model inputs (training data or prompts) | Modify parameters via gradient-based updates | Modify inference behavior without further training | Rewrite model output text generations

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Post-training

Instruction tuning

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Multilingual instructions available:

- Licensing issues.
- Domain, style and format limitations.
- Lack of context awareness.
- “Identity” theft.
- Circa 419k pairs for our pre-training languages (one year ago).

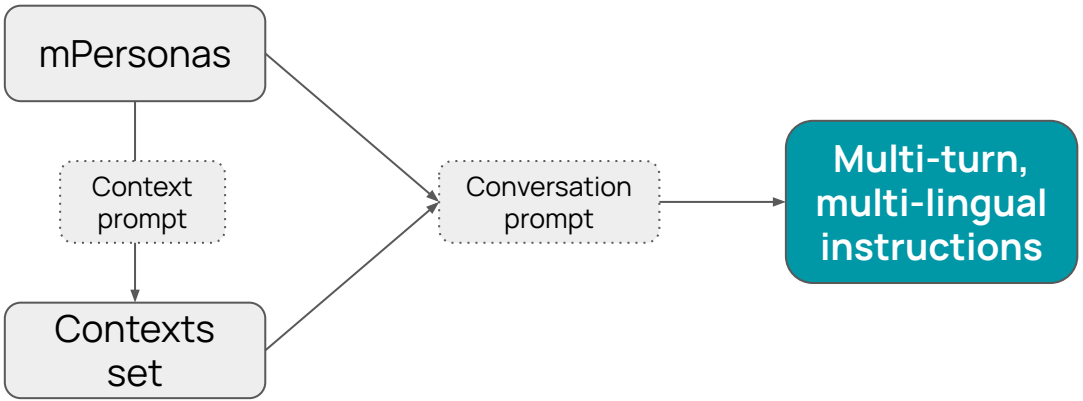
Use **synthetic data responsively** and **edit existing data**.

Instruction tuning

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Follow the **strategy from PersonaHub**, replacing tools and data subject to viral license with **open** ones.

Generated **personas follow the same language as the source document** being used, making them more culturally relevant for our context.



Context prompt

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

```
You are a helpful assistant.

  ===Example 1===

  Persona: A curious and analytical individual, likely with a background in mathematics or science, who enjoys exploring intriguing "what if" scenarios and is fascinated by the intersection of population demographics and geography.

  Prompt: Is it possible for the global population to stand on Jeju Island?

  ===Example 2===

  Persona: An astronomy enthusiast or a professional astronomer, likely with a strong interest in peculiar galaxy structures and a good understanding of celestial objects, seeking to gather specific information about the unique Hoag's object galaxy.

  Prompt: Given the unique ring structure of Hoag's Object and its mysterious formation, what are the leading theories explaining the origin of the central core galaxy? Additionally, is there any evidence supporting the hypothesis that the ring may have formed through a gravitational lensing effect or past galactic interactions?

  ___

  Your task: Guess a prompt (i.e., instruction) that the following person may ask you to do: <S_REPL>persona<E_REPL>

  Generate it in <S_REPL>lang<E_REPL>.

  Return only the prompt inside <prompt> and </prompt>.
```

Conversation prompt

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

```
You are an AI assistant specialized in creating diverse but specific conversations between a user and an AI assistant.
Create a conversation with <S_REPL>turns<E_REPL> turn(s) based on the persona and prompt below:
       Persona: <S_REPL>persona<E_REPL>
       Prompt: <S_REPL>prompt<E_REPL>
Return the conversation in JSON format in <S_REPL>lang<E_REPL>. The format must be inside <conversation> and </conversation> and
with the format:
    <conversation>
    {
      "lang": "en",
      "conversations": [
        {
          "from": "human",
          "value": "What is the capital of Spain?"
        },
        {
          "from": "gpt",
          "value": "The capital is Madrid."
        }
      ]
    }
    </conversation>
```

Persona examples

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Generic persona

Un periodista especializado en economía y política europea, que sigue de cerca las crisis y tendencias en España y utiliza plataformas digitales como Orbyt para mantenerse informado sobre análisis de medios como The Economist.

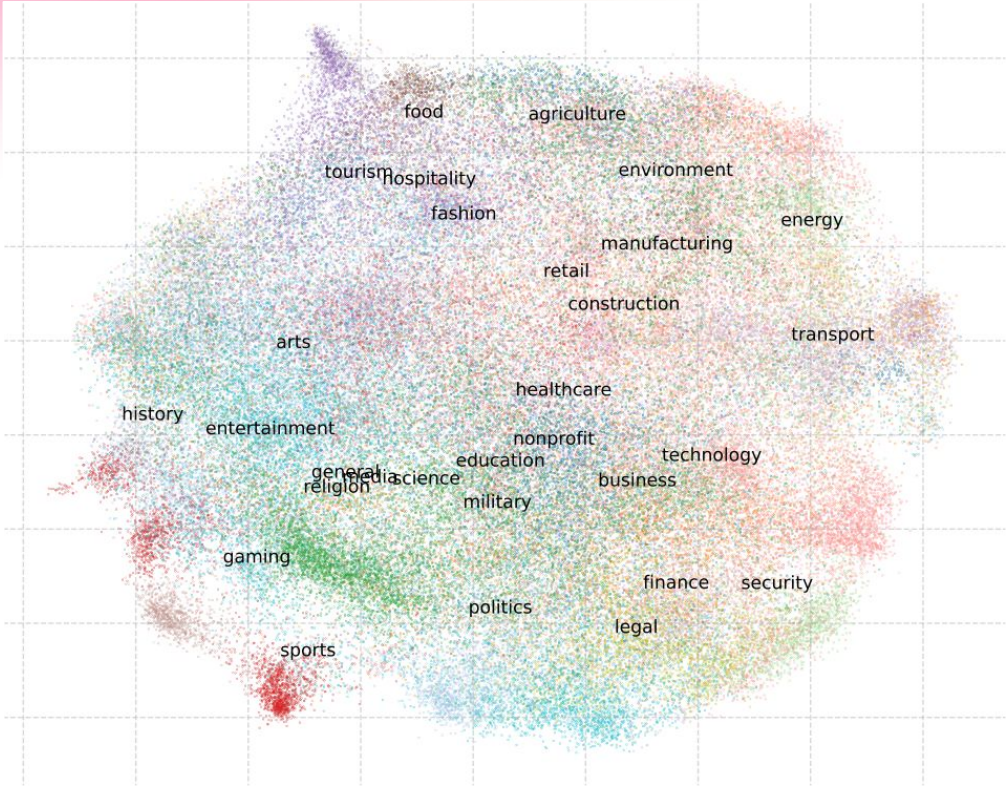
Specific persona

Un historiador especializado en cartografía antigua, con un profundo interés en la evolución de las técnicas de grabado y la producción de atlas entre los siglos XVI y XIX. Trabaja como investigador en el Departamento de Bellas Artes y Cartografía de la Biblioteca Nacional de España, donde estudia y cataloga colecciones de mapas raros, con especial enfoque en obras flamencas, holandesas y españolas. Domina el análisis de estilos artísticos en frontispicios alegóricos y valora la precisión histórica de mapas de exploración. Es experto en figuras como Abraham Ortelio, Gerard Mercator y Tomás López, y frecuentemente colabora con instituciones culturales para exposiciones sobre patrimonio cartográfico.

Instruction tuning

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Domains in the first sample of ~126k personas



Direct Preference Optimization

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

DPO is an optimization algorithm that takes both the chosen instruction-response pairs and their respective rejected pairs, as chosen by human evaluators, to minimize a preference loss.

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y_c,y_r)\sim\mathcal{D}_R}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_c \mid x)}{\pi_{\text{ref}}(y_c \mid x)} - \beta\log\frac{\pi_\theta(y_r \mid x)}{\pi_{\text{ref}}(y_r \mid x)}\right)\right]$$

Proximal Policy Optimization

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

PPO is a two-step approach that first trains a separate reward model to emulate human preferences by minimizing a preference loss,

$$\mathcal{L}_R(\psi) = -\mathbb{E}_{(x,y_c,y_r)\sim\mathcal{D}_R}\left[\log\sigma\left(R_\psi(x,y_c) - R_\psi(x,y_r)\right)\right]$$

and then uses this reward model to give a reward score to the outputs of the instructed model in order to align it with the desired behavior.

DPO vs. PPO

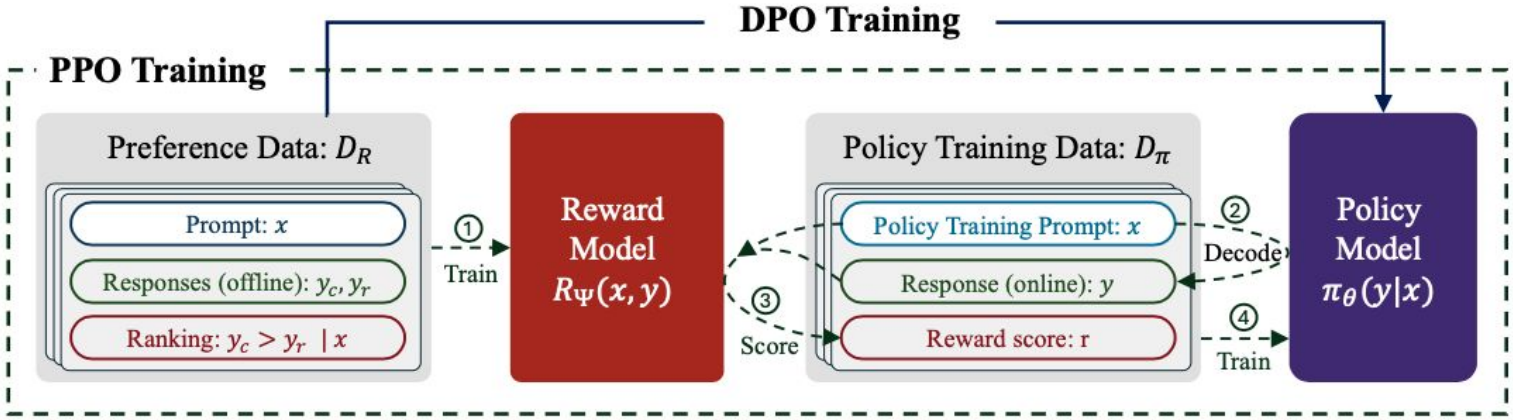


Figure 2: The core aspects of learning from preference feedback. For DPO (solid line), preference data is directly used to train a policy model. For PPO (dashed line), preference data is used to train a reward model, which is then used to score model-generated responses during PPO training.

Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preference Feedback. Ivison et al. (2024). *NeurIPS'24*

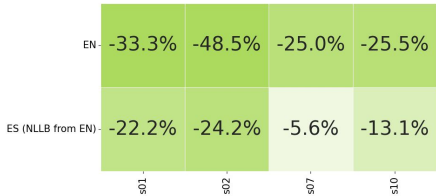
Safety

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Pipeline **simulates malicious use**. Moderator model classifies the type of attack and if it was successful or not.

Focus on **four axes**: risk type, attack strategy, language, and number of turns.

Basic RLHF alignment with **DPO improves model resistance** to attacks.

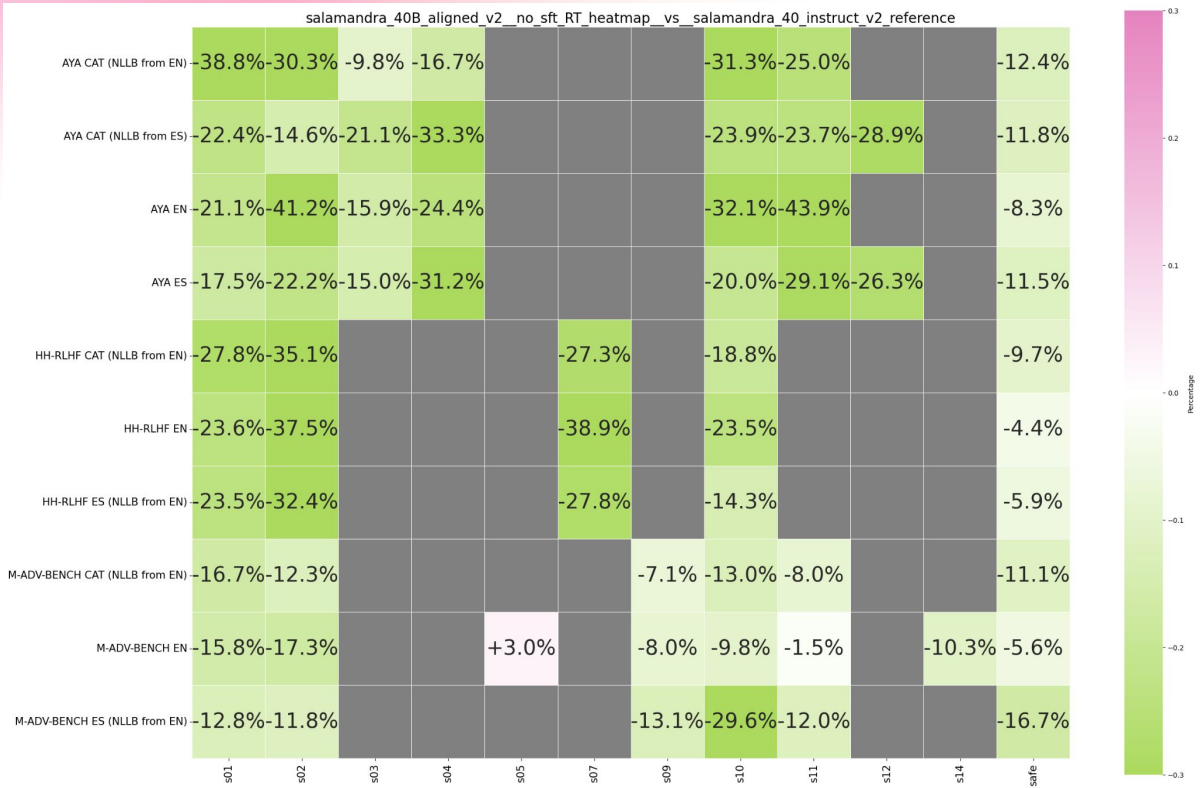


Use a combination of **synthetic data and human annotation** to expand to languages other than English.

Focus on **proven methodologies but using open-license data**.

Safety

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Coming soon

Coming soon

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

More **capable models** (all sizes) with **larger contexts** and **new languages**.

Refined **instructed and aligned** versions.

Multimodal, multilingual models.

Multimodal, multilingual **evaluation**.

Multilingual agents and judges.

Multilingual **judge evaluation**.

Domain-specific models.

In-house **quantization**.

+ more.

Building and Evolving a Multilingual LLM from Scratch

Javier Aula-Blasco, PhD

/ CyberColombia
/ 8th Colombian HPC Summer School
/ June 17, 2025

Acknowledgements

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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

This work is funded by the *Ministerio para la Transformación Digital y de la Función Pública* and *Plan de Recuperación, Transformación y Resiliencia* - Funded by EU – NextGenerationEU within the framework of the project **ILENIA** with reference 2022/TL22/00215337, 2022/TL22/00215336, 2022/TL22/00215335 and 2022/TL22/00215334, and within the framework of the project **Desarrollo Modelos ALIA**.

Javier Aula-Blasco, PhD

Head of Data and Model Evaluation

Language Technologies Lab

javier.aulablasco@bsc.es

/javieraulablasco

20 BSC YEARS