EuroBERT: Scaling Multilingual Encoders for European Languages



ARTEFACT, TECHDREDI
APRIL 25, 2025

The Team



Nicolas Boizard



Hippolyte Gisserot-B.



Duarte M. Alves



Manuel Faysse



Ricardo Rei



Maxime Peyrard



Kevin El Haddad



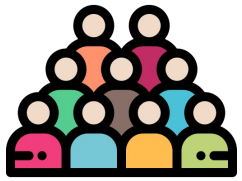
André Martins



Céline Hudelot



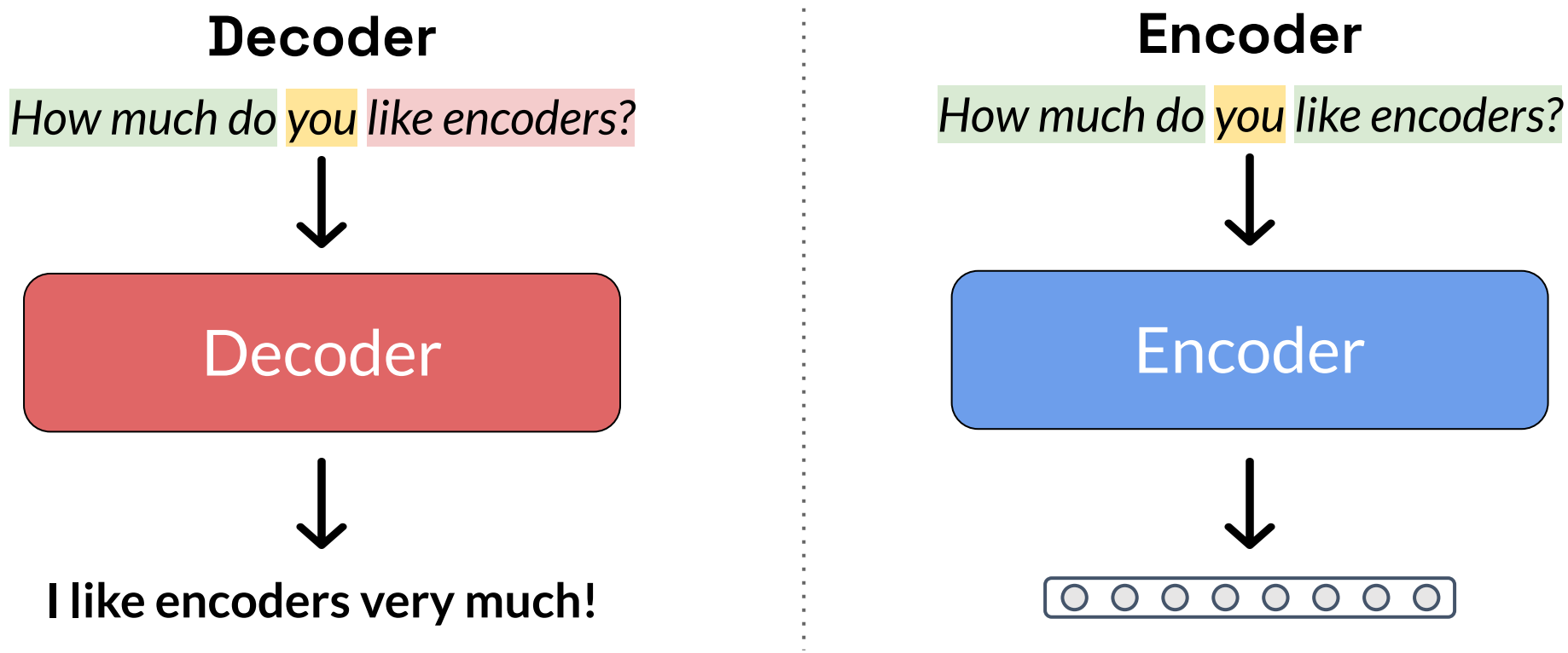
Pierre Colombo



Many more

Encoders vs. Decoders

In the modern era of NLP, two primary types of models dominate: encoders and decoder-only generative models.



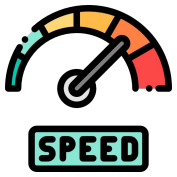
Why Encoder Models?



With the recent explosion of generative AI models, researchers tend to forget encoders are still by far the most widely used deep learning text models in the industry!



Versatility
Encoder representations can be used for tasks like classification, retrieval, NER, extractive Q&A...



Efficiency
Encoder models are inherently more compact than decoder models, making them significantly faster for both fine-tuning and inference



Opportunities
With recent advances in training decoder models (improvements in data quality, scale, and architecture) there's strong potential to drive similar progress in encoder models

EuroBERT Models at a Glance

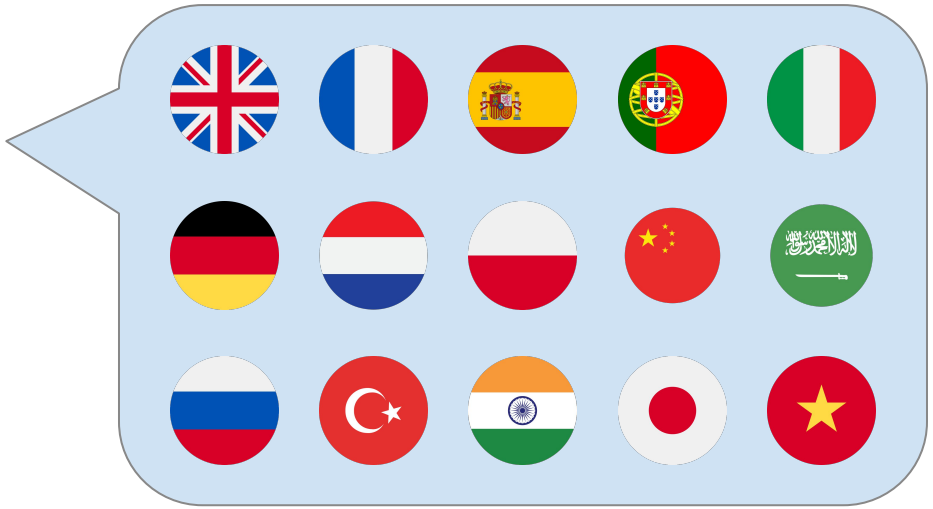
- ➔ **3 sizes** suitable for various constraints
- ➔ **5T tokens** seen during training
- ➔ **15 languages** supported
- ➔ **Long context** (8k tokens)
- ➔ **Fully open-source**



EuroBERT 210M

EuroBERT 610M

EuroBERT 2.1B



End-to-End Training Recipe

Pre-training

Data
> Large quantity of unsupervised data
> Tokens masked randomly

Learning objective
> MLM: model trained to retrieve the masked tokens

Annealing

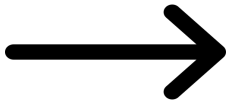
Data
> Smaller quantity of unsupervised data
> Higher quality

Learning objective
> MLM
> Learning rate decreased to 0

Fine-tuning

Data
> Depending on use case (classification, regression...)

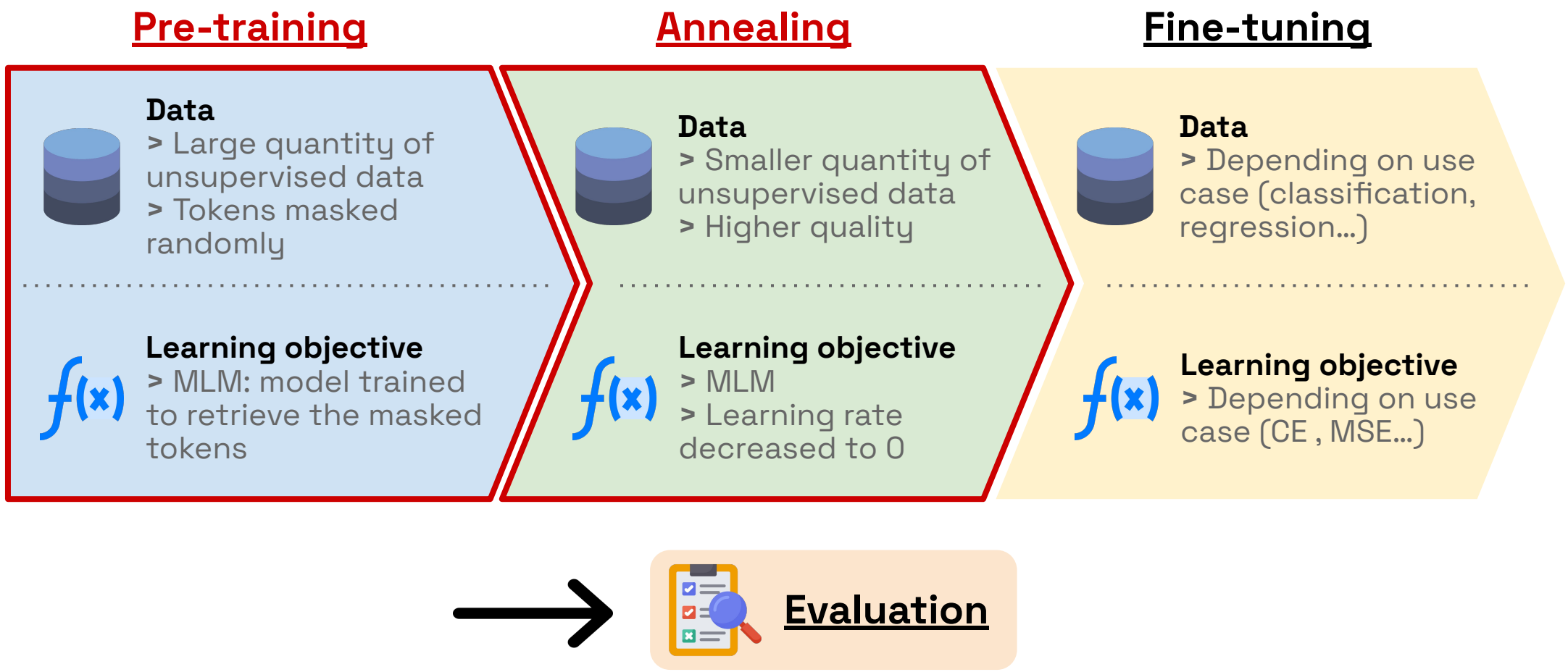
Learning objective
> Depending on use case (CE , MSE...)





Evaluation

Pre-training

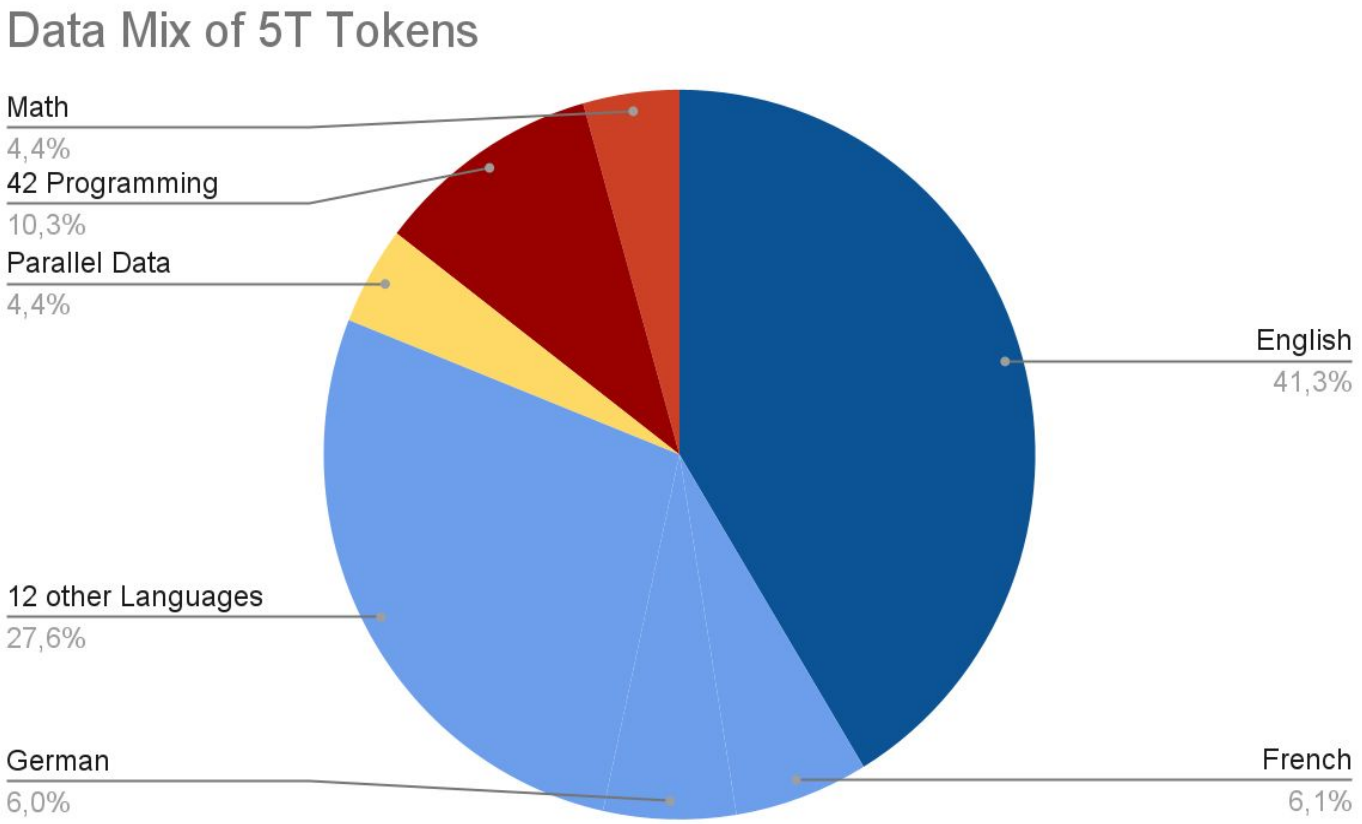


Training Data



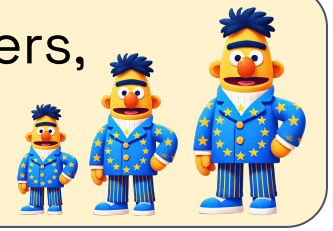
Data mix is key

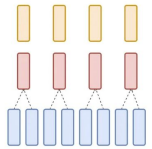
> **Mix:** a model is good on the domains it has seen
> **Math & code:** for reasoning purposes
> **Parallel data:** to transfer knowledge between languages



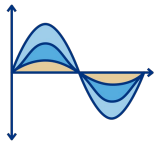
Model Architecture

EuroBERT models range from 210 million to 2.1 billion parameters, all trained on the same number of tokens and using identical architectural configurations.





GQA
Grouped Query Attention is faster and uses less memory



RoPE
Rotary Positional Embeddings achieve better performance notably for long context



GLU activation
Brings stability and performance



RMS normalisation
Brings stability and performance

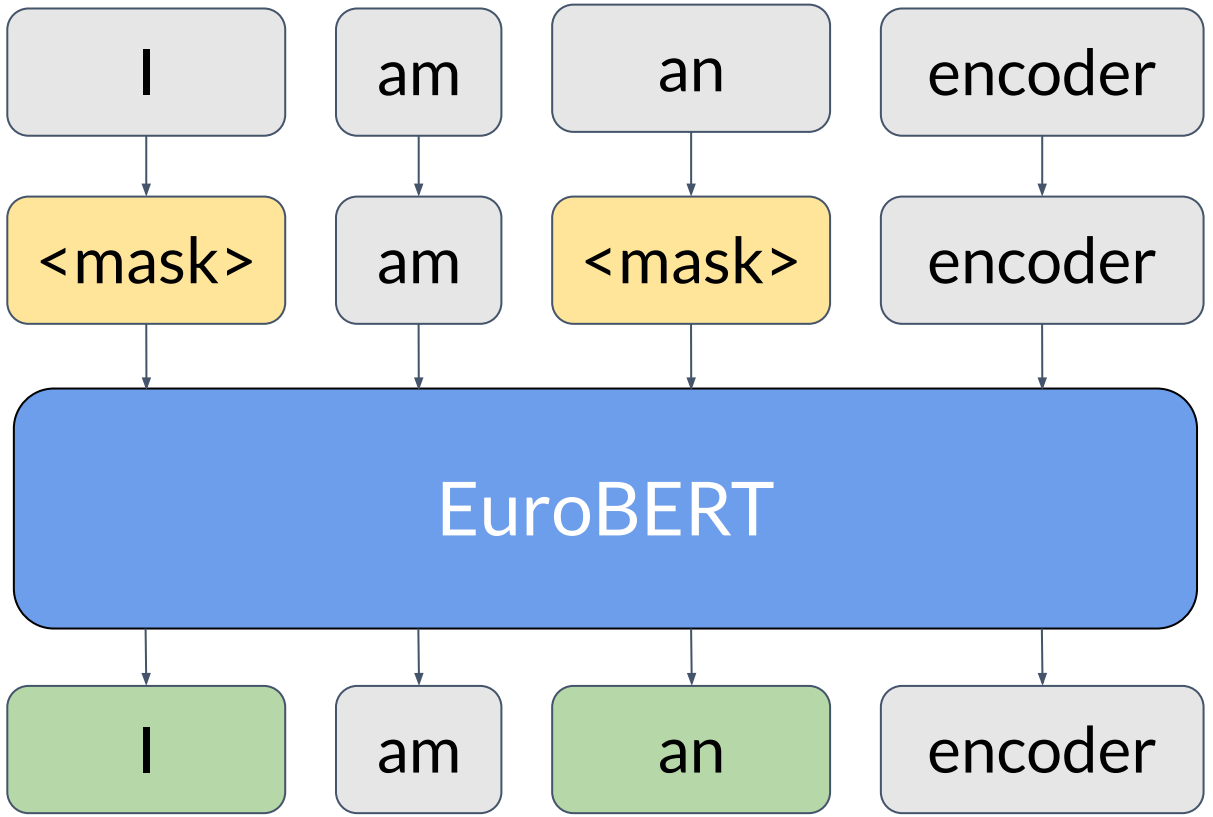


Long context
Performance and task oriented

Masking Strategy

Masking strategy

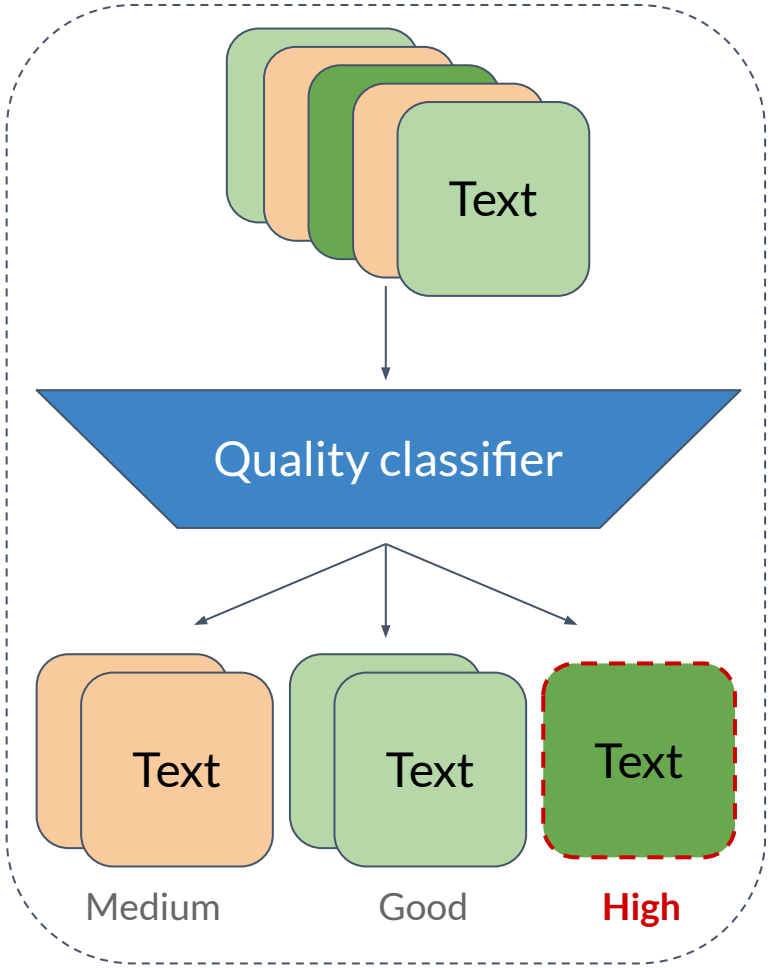
> 50% of the sentence is masked
> Scaling law: this ratio yields limit stability
> Only masking: no replacement with random tokens (cf. BERT-like training)



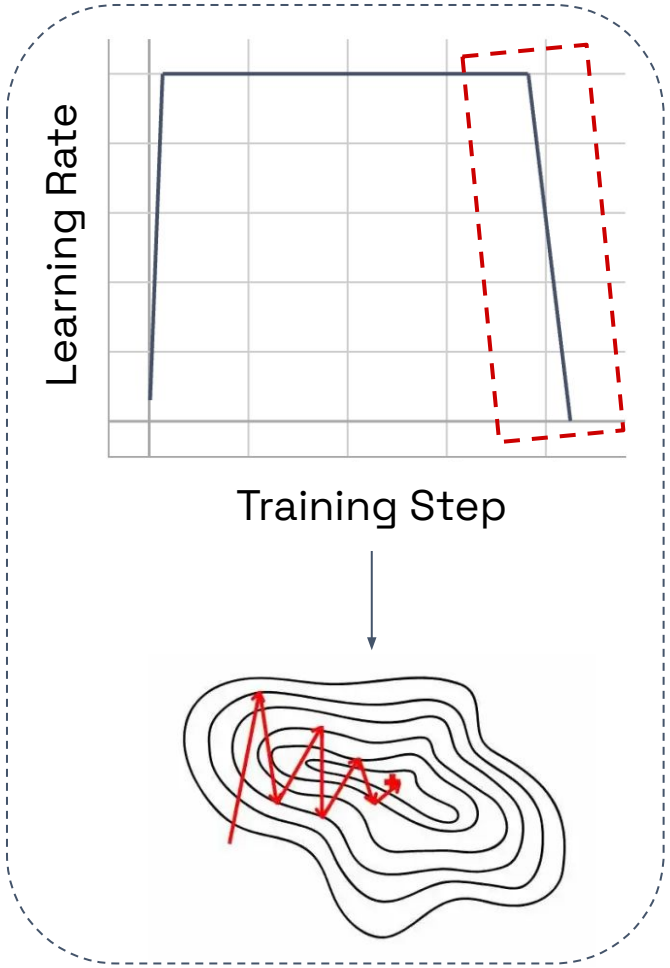
$f(x)$: *Cross-entropy on vocabulary space*

Annealing

Quality Filtering



Learning Rate Decay

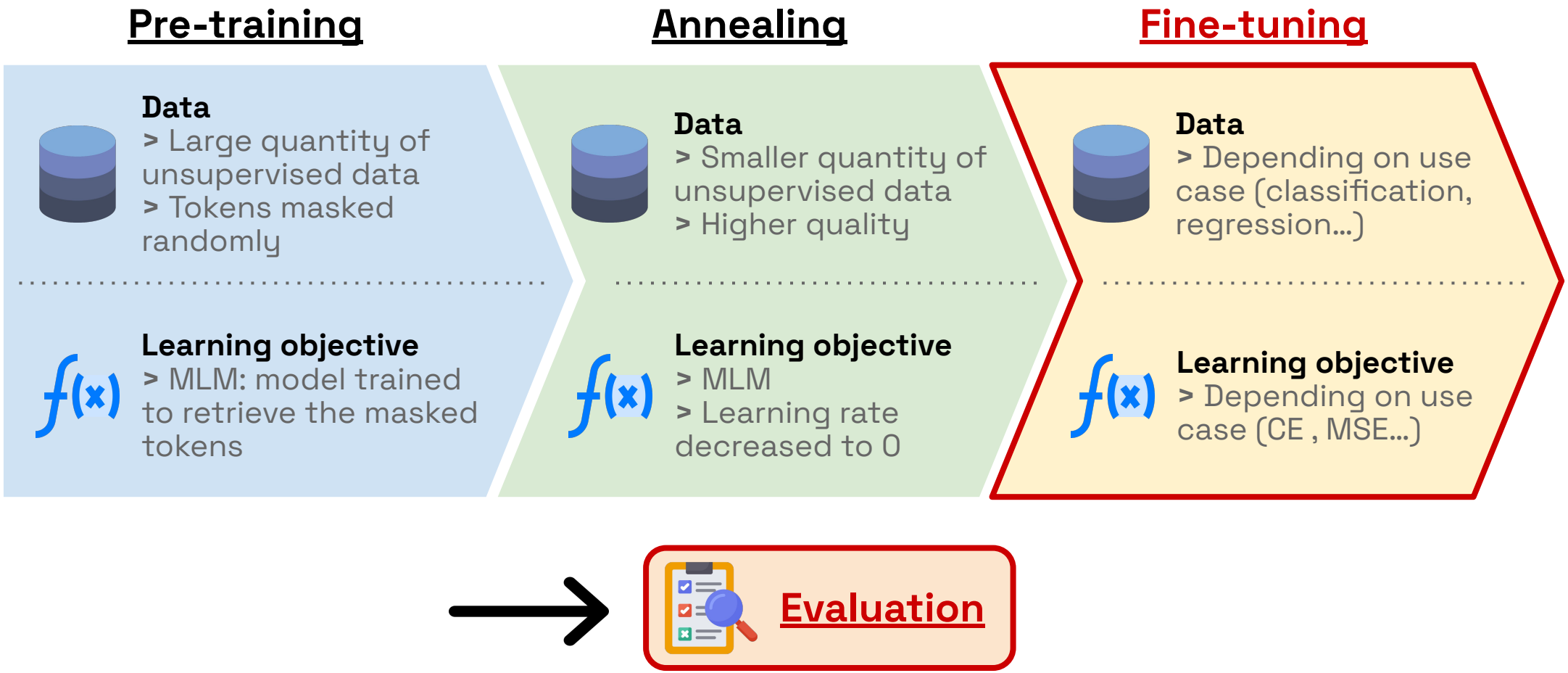


Improved performance

> Improved alignment with real world data

> More effective convergence toward a local minimum

Fine-tuning & Evaluation



Evaluation Protocol



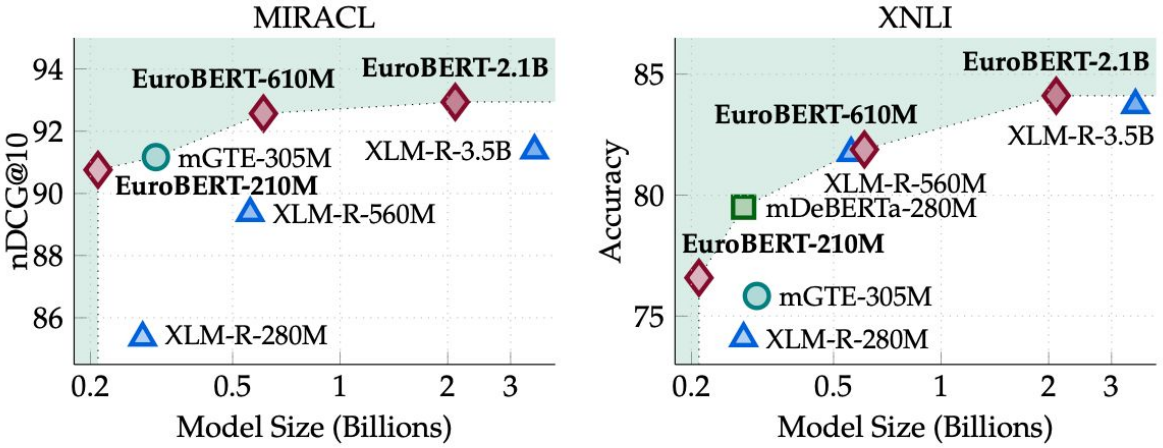
Evaluation protocol

> **Tasks**: retrieval, sequence classification & regression, token classification
> **Baselines**: XLM-RoBERTa, mDeBERTa, mGTE, ModernBERT
> **Fine-tuning**: all models are applied the same protocol to ensure a fair comparison

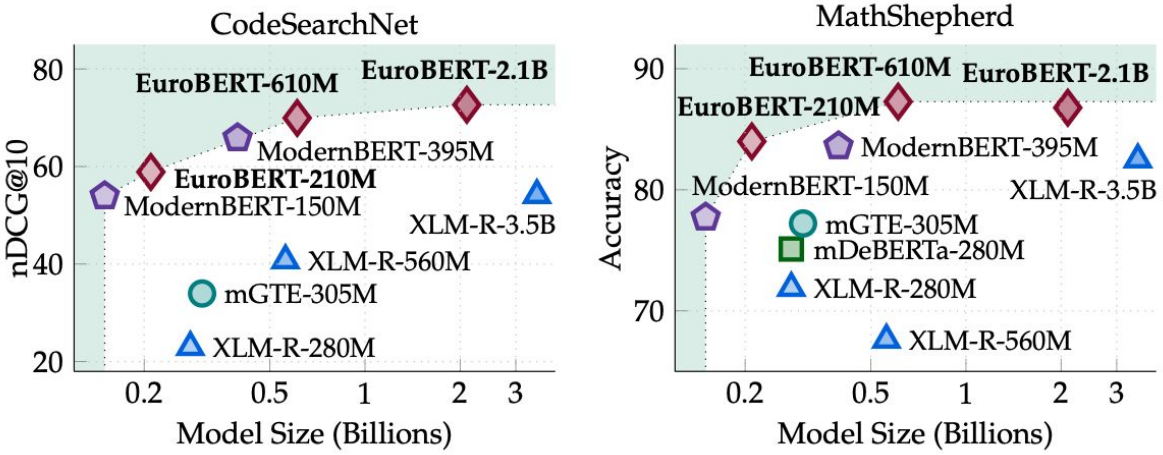
	European Languages								Extra-European Languages							Code	Math
Task	en	de	es	fr	it	nl	pl	pt	ar	hi	ja	ru	tr	vi	zh		
Information Retrieval																	
MIRACL	✓		✓	✓						✓	✓	✓	✓			✓	
MLDR	✓	✓	✓	✓	✓			✓		✓	✓	✓				✓	
Wikipedia	✓	✓			✓	✓		✓				✓					
CC-News	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
CodeSearchNet	✓															✓	
DupStackMath	✓															✓	
MathFormula	✓																✓
Sequence Classification																	
XNLI	✓	✓	✓	✓					✓	✓		✓	✓	✓	✓		
PAWS-X	✓	✓	✓	✓													
AmazonReviews	✓	✓	✓	✓							✓				✓		
MassiveIntent	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
CodeDefect	✓															✓	
CodeComplexity	✓															✓	
MathShepherd	✓																✓
Sequence Regression																	
WMT	✓	✓		✓			✓				✓	✓	✓		✓		
SeaHorse	✓	✓	✓										✓	✓			
Token Classification																	
NER	✓	✓	✓			✓											

Evaluation Results

Multilingual Tasks



Code and Math Tasks



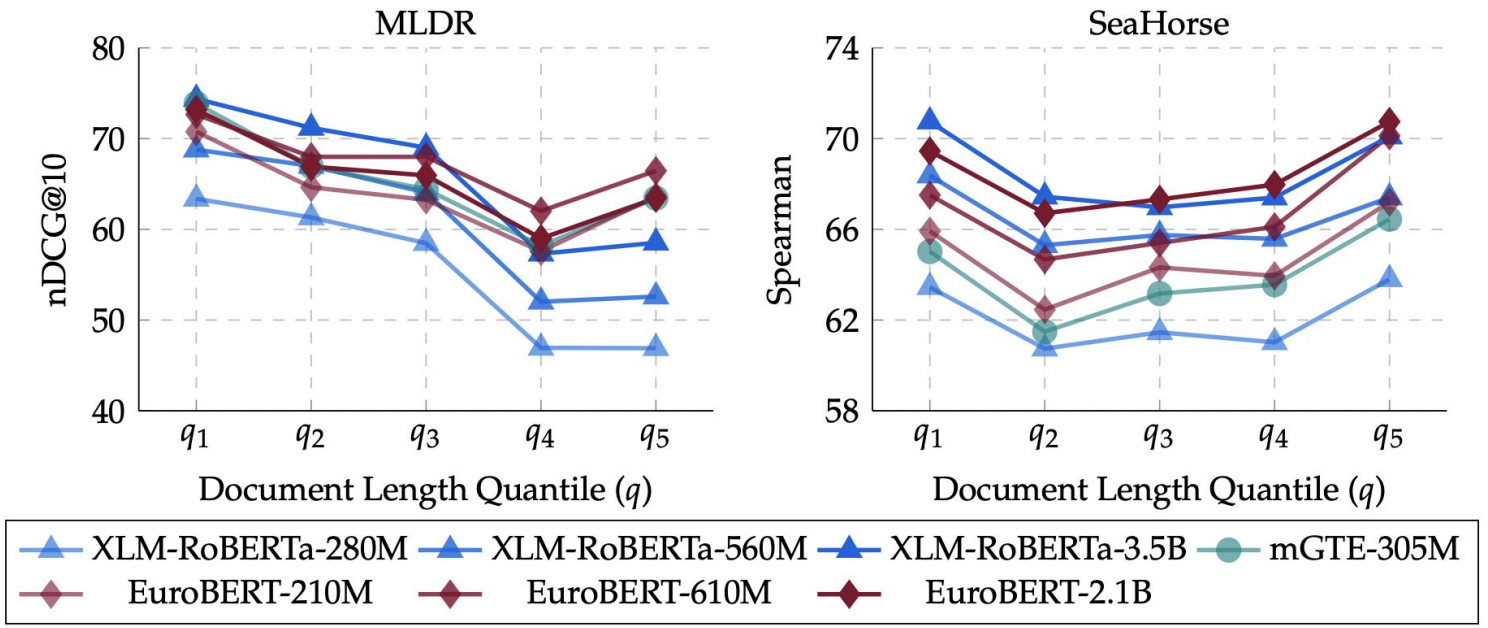


EuroBERT performance

> **Multilingual strength:** EuroBERT matches or surpasses similarly sized models on multilingual benchmarks

> **Math & Code:** EuroBERT achieves SOTA results on mathematical and code-related tasks

Long Context Performance



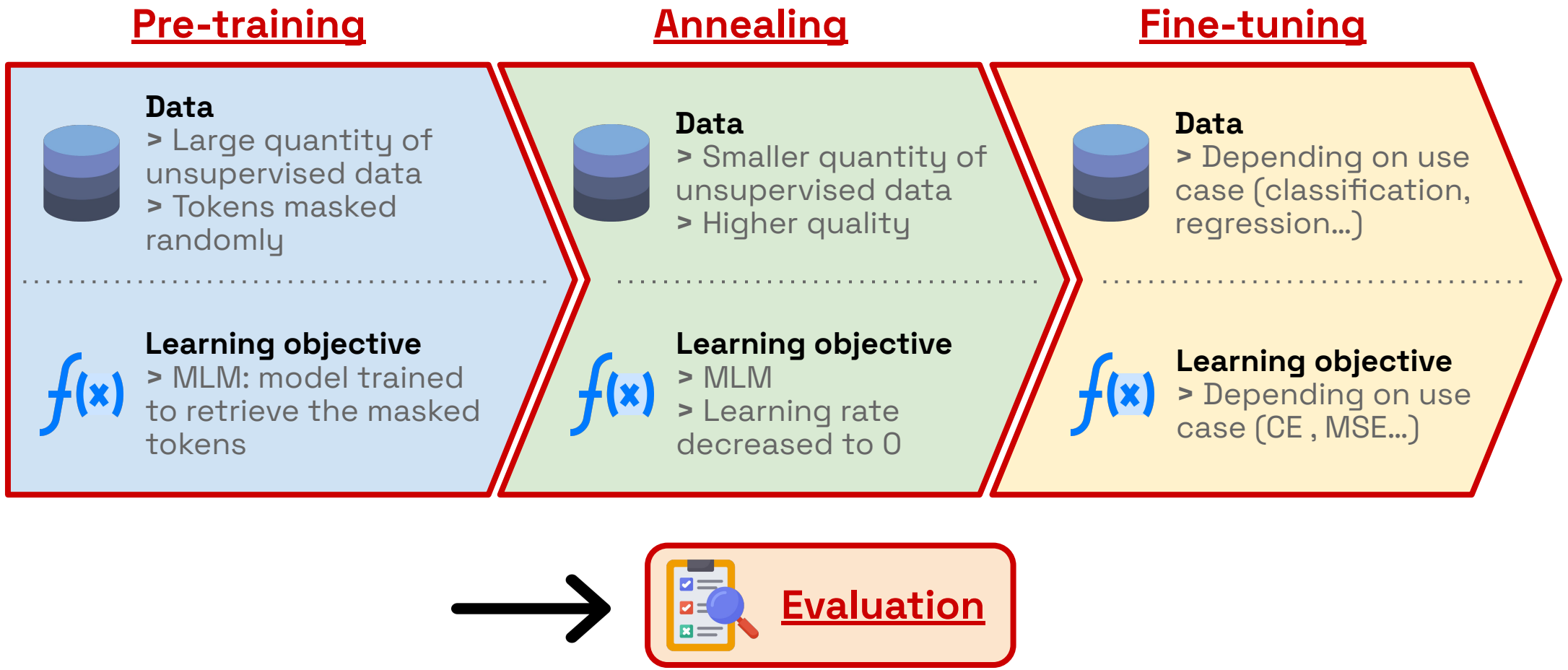


Long context performance

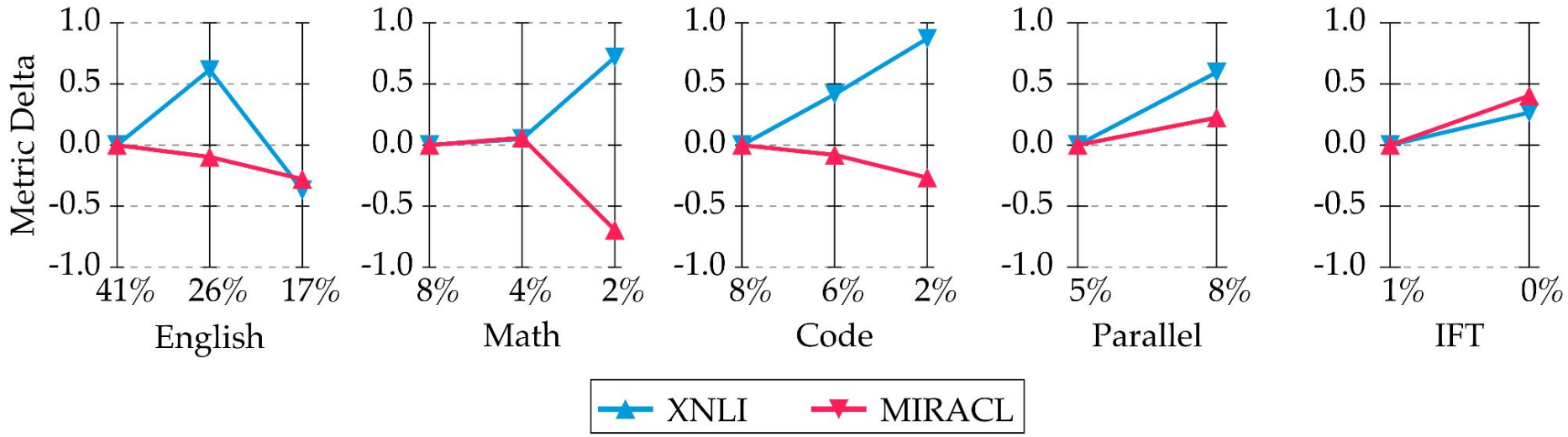
> EuroBERT models **outperform** all other encoder models of similar size on long-context tasks

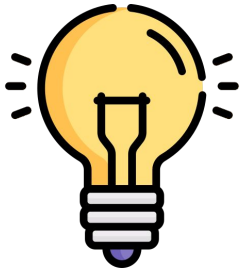
> In long-context retrieval, they demonstrate **particularly strong performance on very long documents**

Ablation Studies



Data Mix



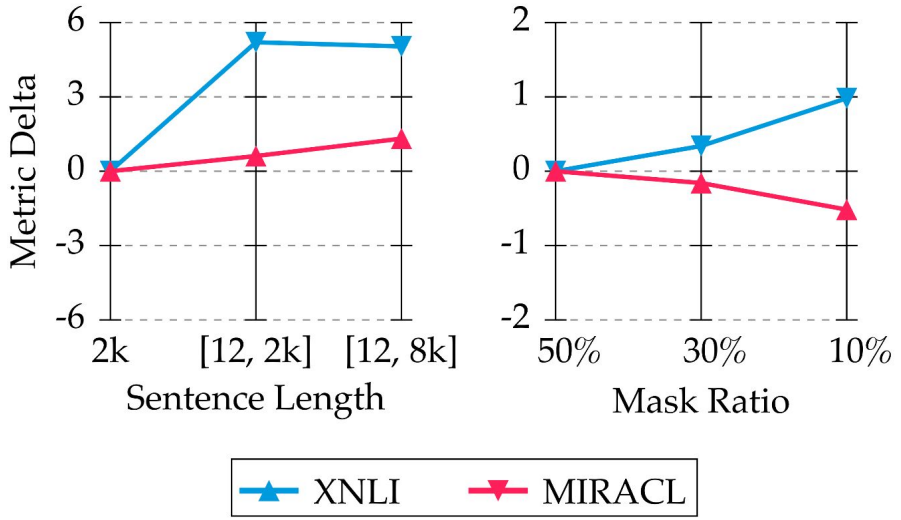


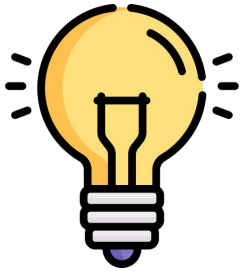
Impact of data composition during annealing on downstream performance

> **Positive performance factors**: balanced language representation, inclusion of parallel data, exclusion of IFT data

> **Inter-task competition**: presence of math and code data introduces trade-offs across tasks

Training Hyperparameters



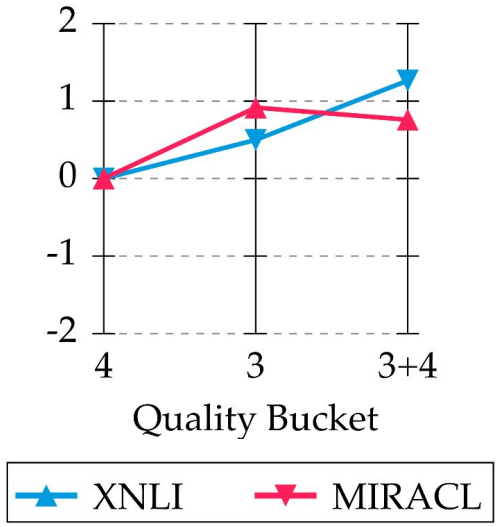


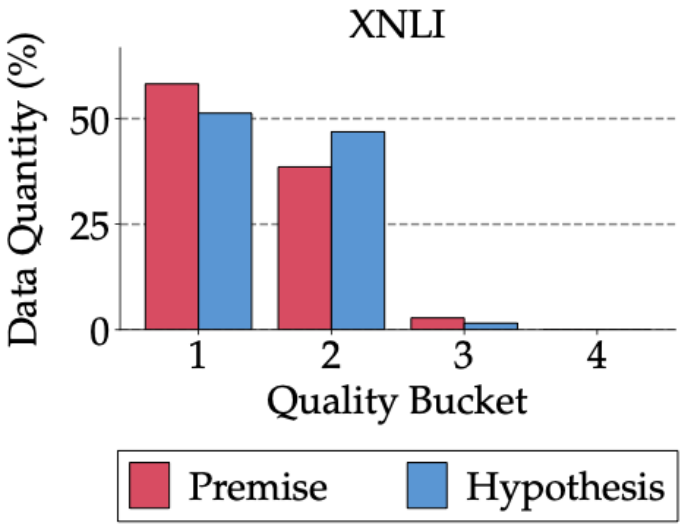
Impact of major training hyperparameters on downstream performance

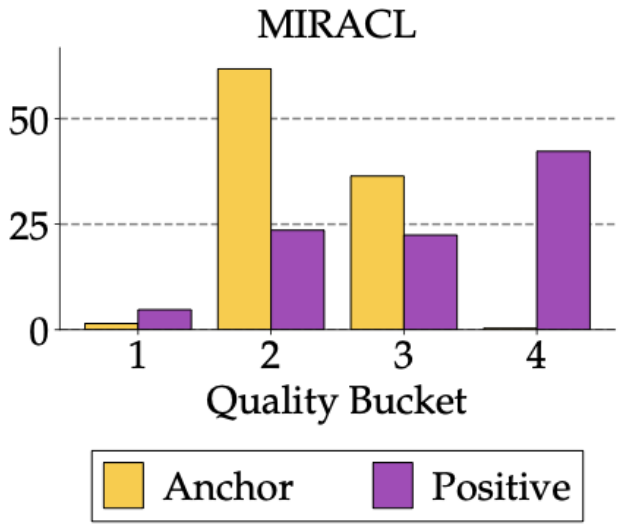
> **Positive performance factors**: longer sequence lengths, variability in sequence length

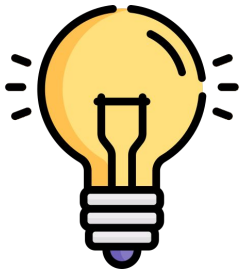
> **Inter-task competition**: masking ratio introduces trade-offs (higher masking ratios generally favor retrieval, while lower ratios tend to support classification tasks)

Quality Filtering







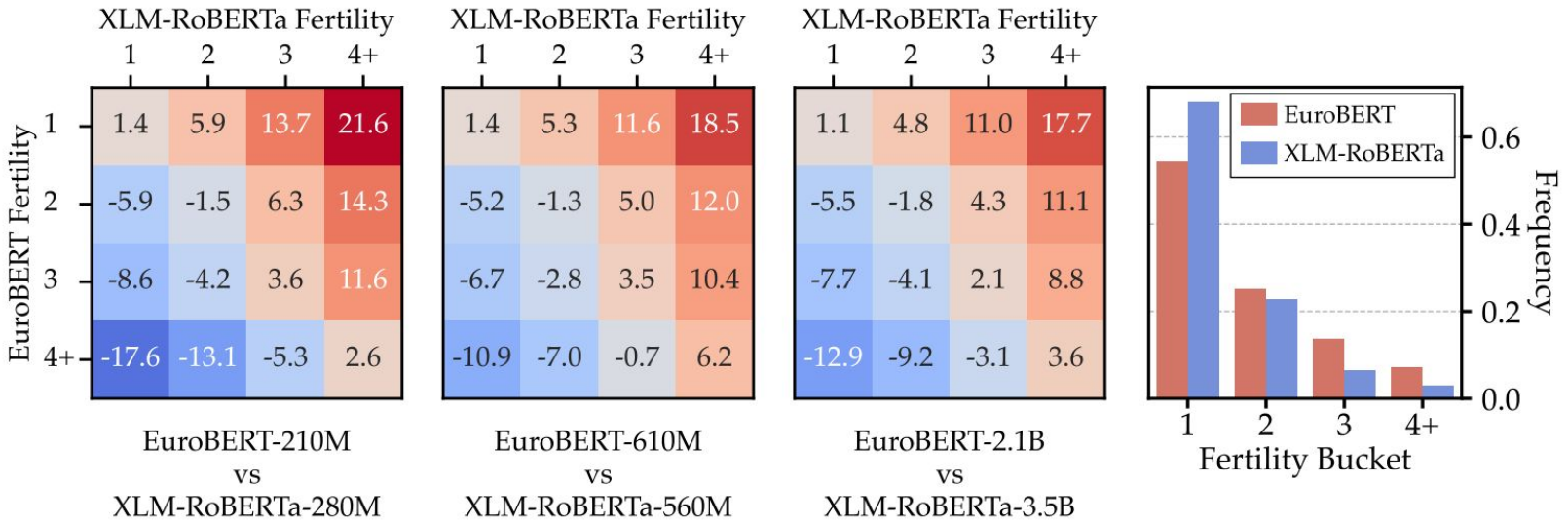


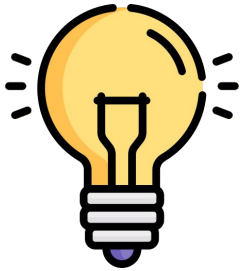
Impact of data quality on downstream performance

> **Counterintuitive finding**: lowering the quality threshold can actually improve downstream performance

> **Explanation**: this effect may stem from the nature of evaluation datasets and the biases of quality classifiers, which tend to favor longer contexts, that are inherently more informative

Token Classification





Impact of tokenizer fertility on NER performance

> **Observation**: EuroBERT models underperform compared to similarly sized counterparts on NER, despite maintaining reasonable overall performance

> **Explanation**: higher fertility implies splitting entities into more subtokens, diluting their semantic coherence and thus complicating the supervision

Future Work



Supporting more languages
> **Strategy**: expand language coverage to include all official EU languages, followed by broader inclusion of languages spoken across the entire European continent



Contrastive training for enhanced sequence-level representations
> **Approach**: apply contrastive training to the pre-trained model to strengthen its ability to learn meaningful sequence-level representations
> **Goal**: achieve competitive or superior performance on embedding benchmarks such as MTEB, relative to models of similar size

Thank you for your attention!

Paper:

HF page:



ARTEFACT, TECHDREDI
APRIL 25, 2025