U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

LIACC

M.EIC

Natural Language Processing

Henrique Lopes Cardoso

FEUP / LIACC

hlc@fe.up.pt

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

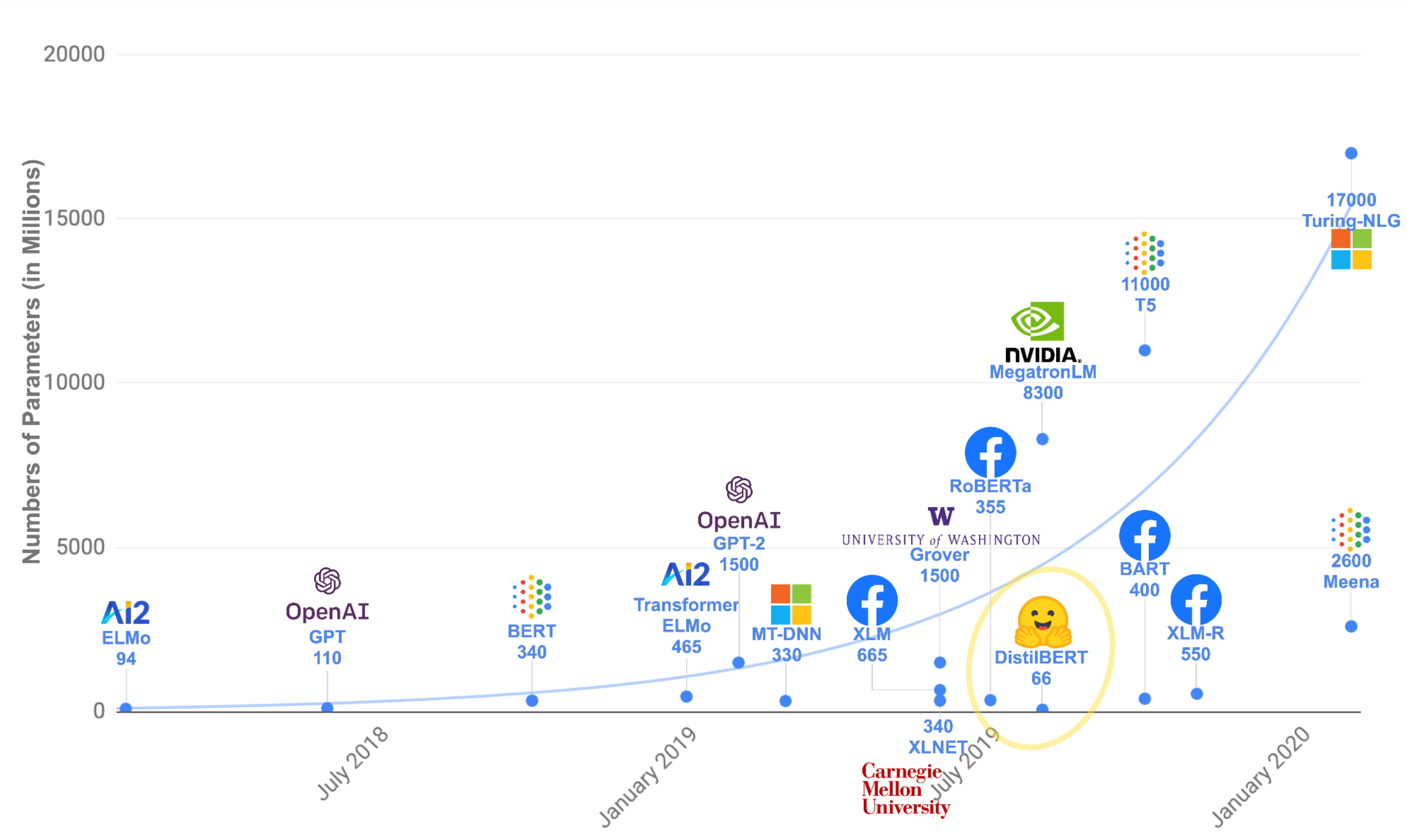
LIACC

Hugging Face Transformers

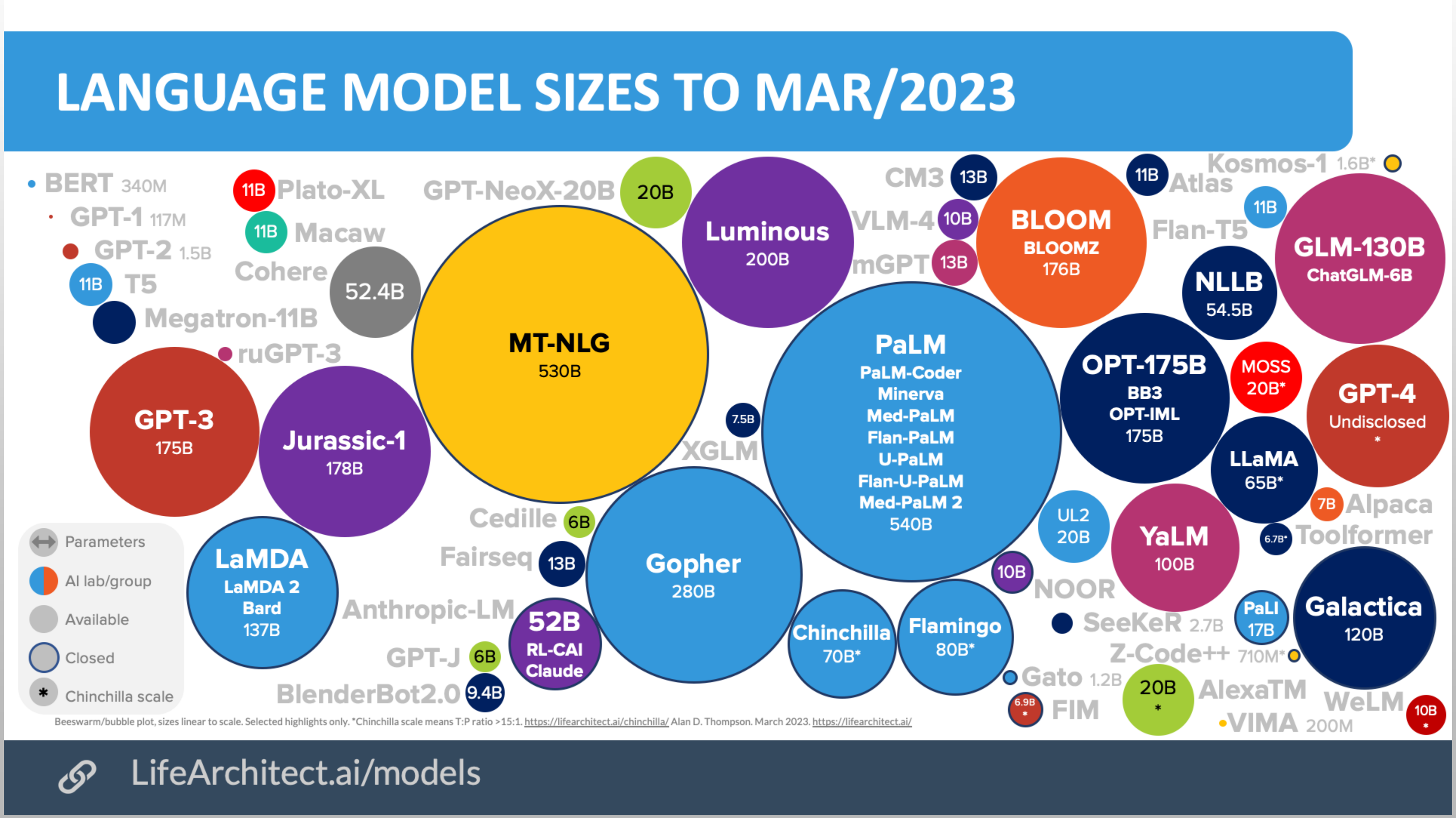
Hugging Face

https://huggingface.co/

Transformers are big models



Transformers are big models

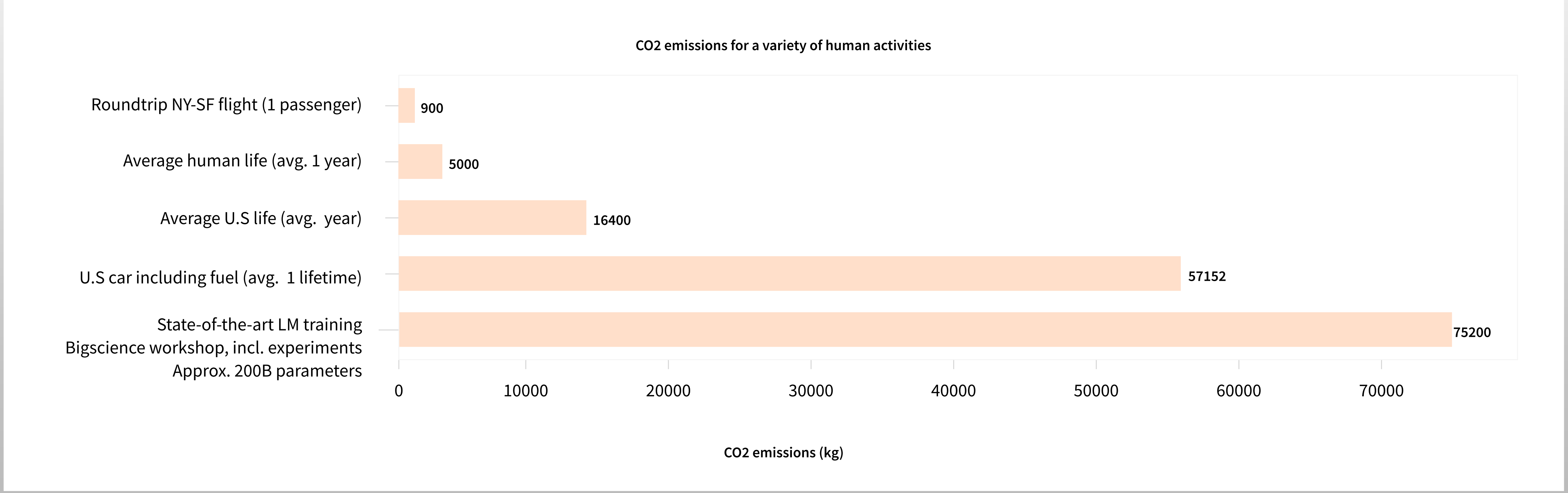


LANGUAGE MODEL SIZES TO MAR/2023

Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. March 2023. https://lifearchitect.ai/

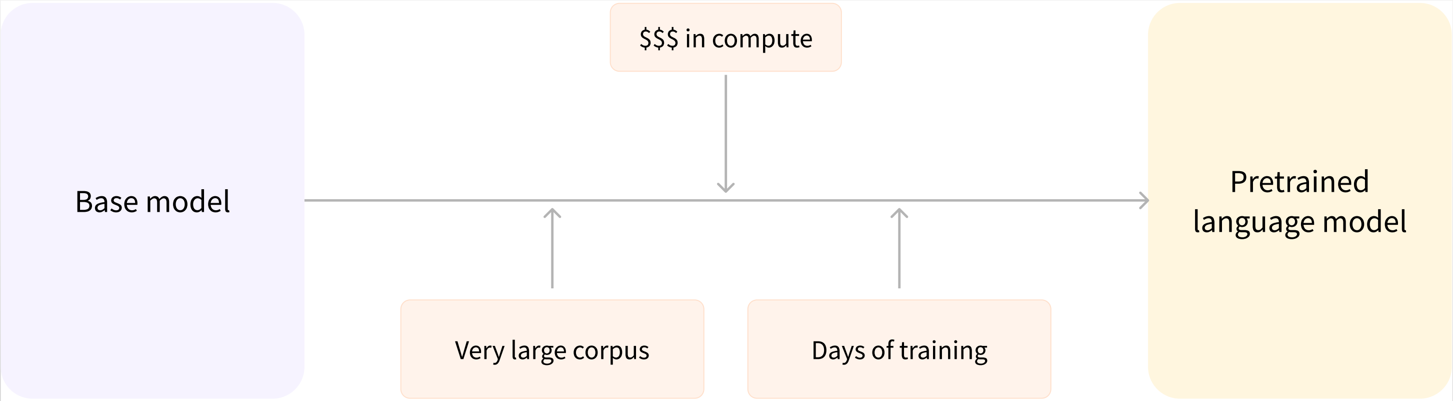
LifeArchitect.ai/models

The carbon footprint



Transfer Learning

- Pretraining



Base model

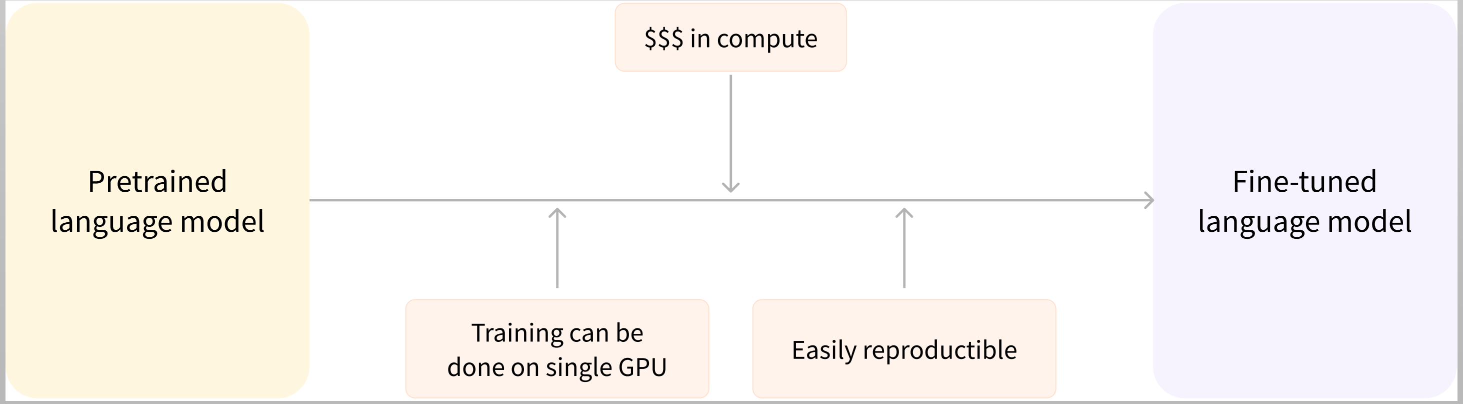
$$$ in compute

Very large corpus

Days of training

Pretrained language model

- Fine-tuning



Pretrained language model

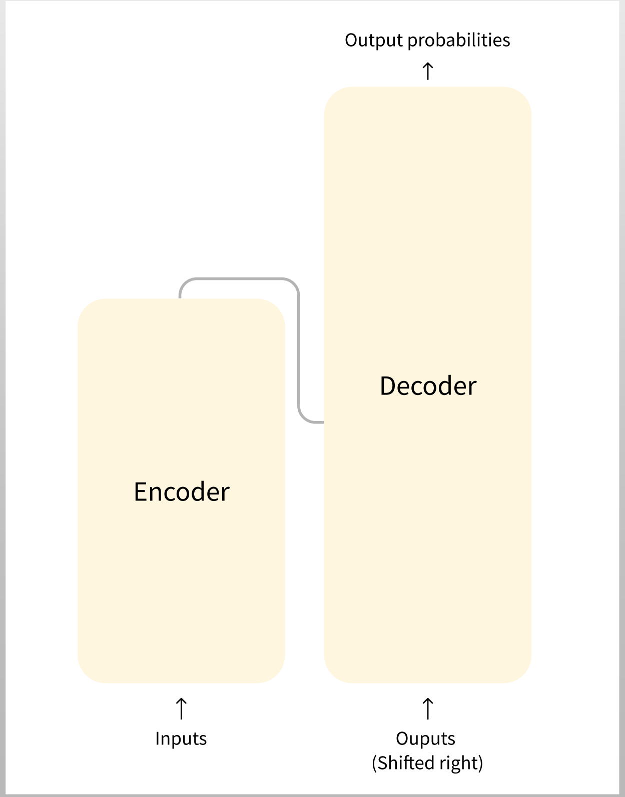
$$$ in compute

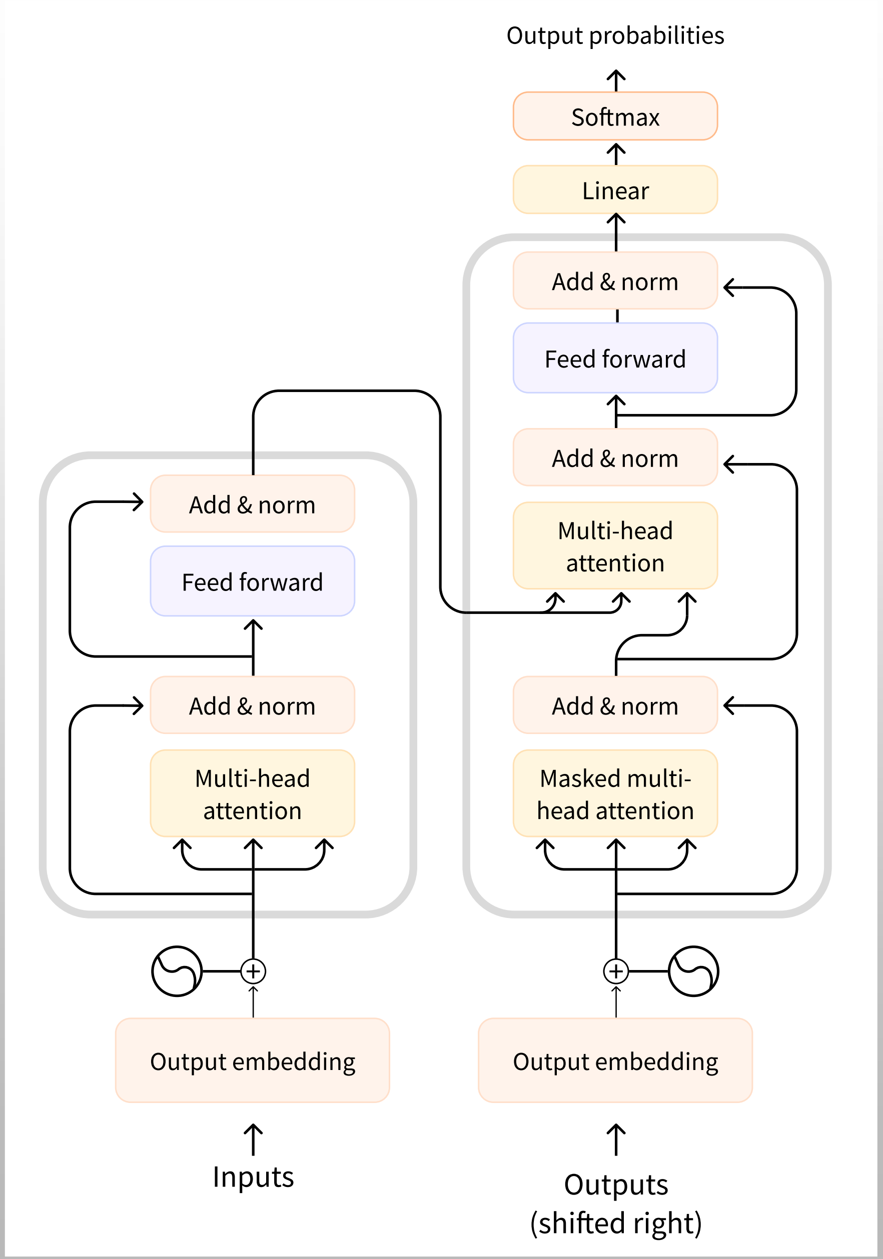
Training can be done on single GPU

Easily reproductible

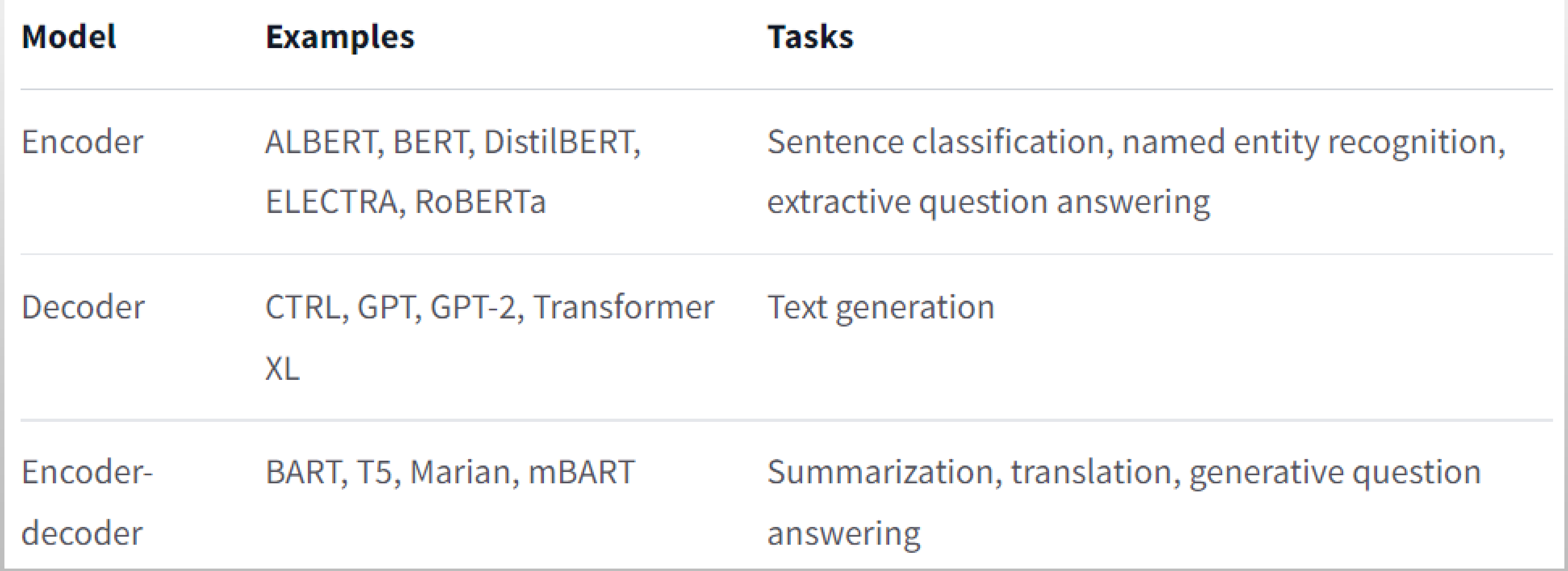
Fine-tuned language model

Transformer Architecture



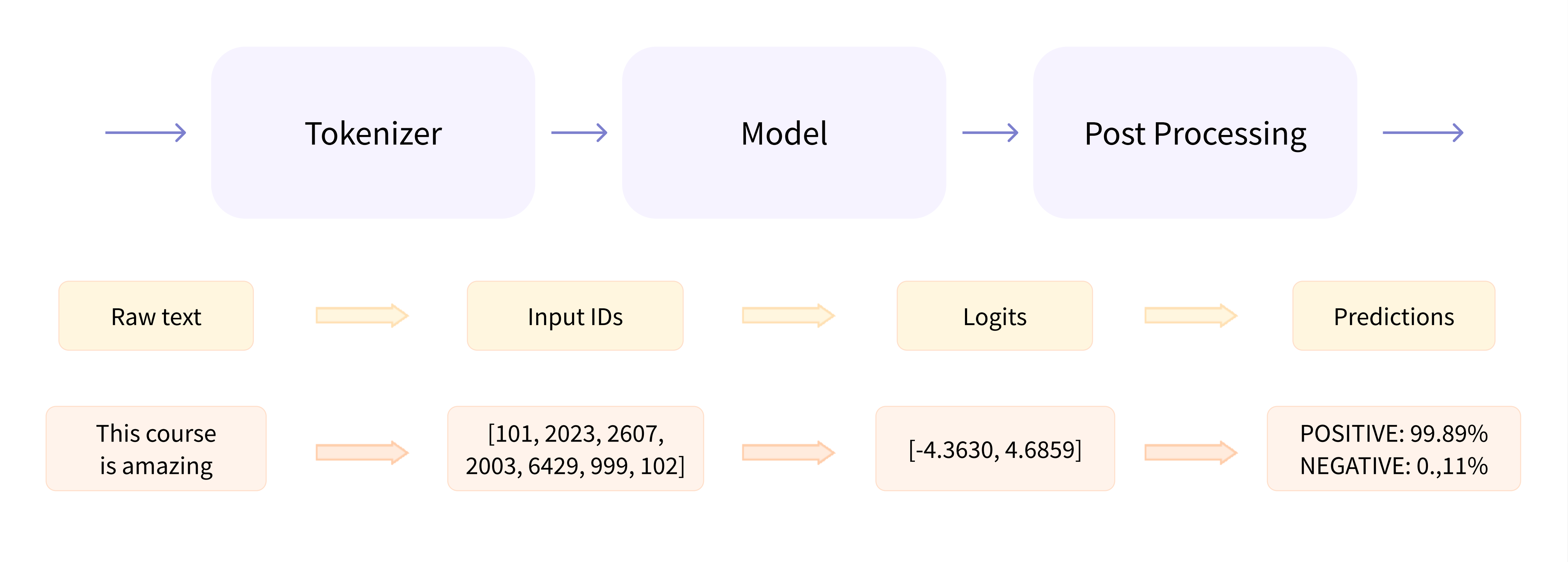


Encoder, Decoder, Encoder-Decoder



Model	Examples	Tasks
Encoder	ALBERT, BERT, DistilBERT, ELECTRA, RoBERTa	Sentence classification, named entity recognition, extractive question answering
Decoder	CTRL, GPT, GPT-2, Transformer XL	Text generation
Encoder-decoder	BART, T5, Marian, mBART	Summarization, translation, generative question answering

Pipelines

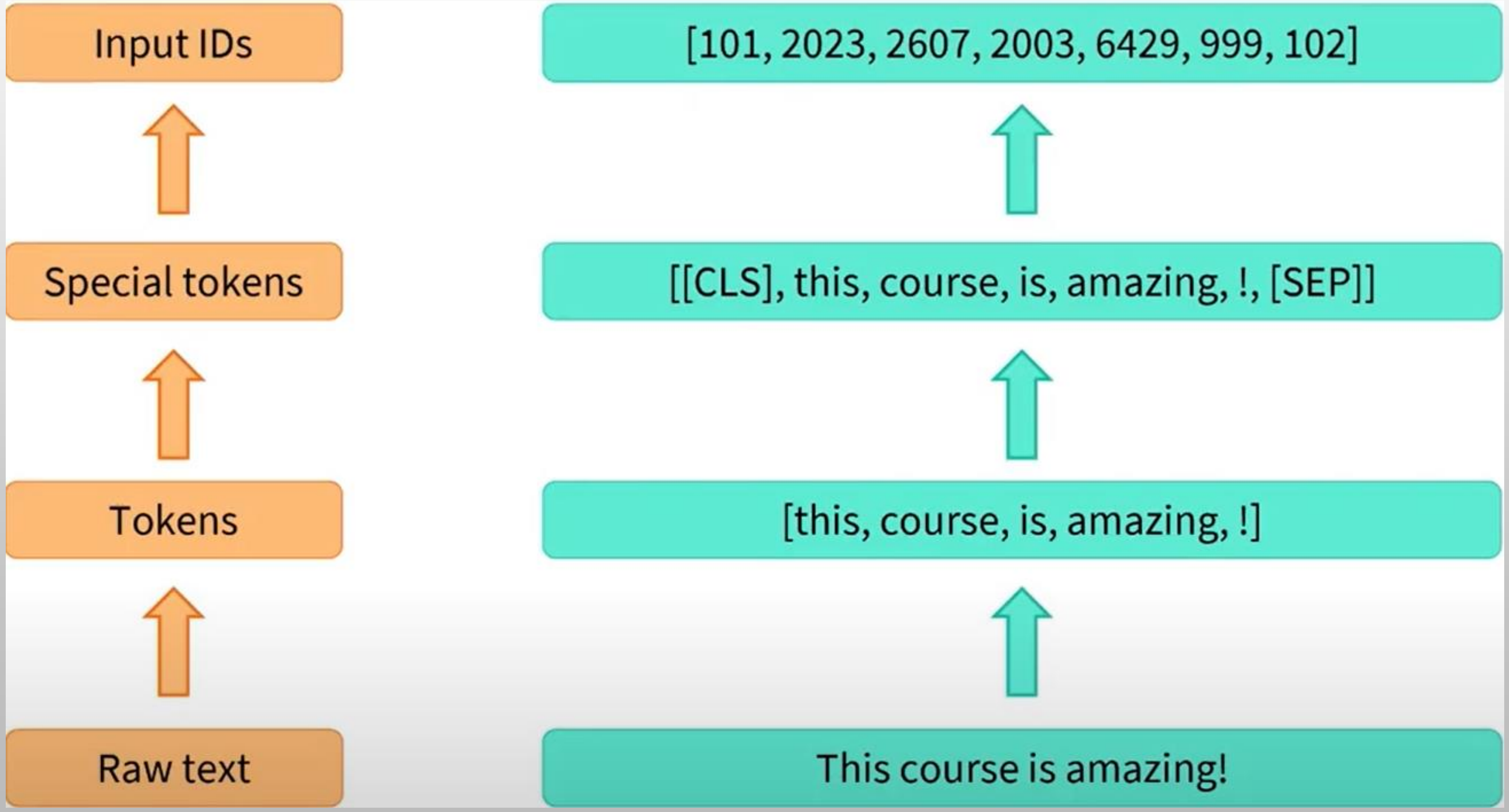


Tokenizer → Model → Post Processing

Raw text → Input IDs → Logits → Predictions

This course is amazing → [101, 2023, 2607, 2003, 6429, 999, 102] → [-4.3630, 4.6859] → POSITIVE: 99.89% NEGATIVE: 0.,11%

Tokenizer



Input IDs

[101, 2023, 2607, 2003, 6429, 999, 102]

Special tokens

[[CLS], this, course, is, amazing, !, [SEP]]

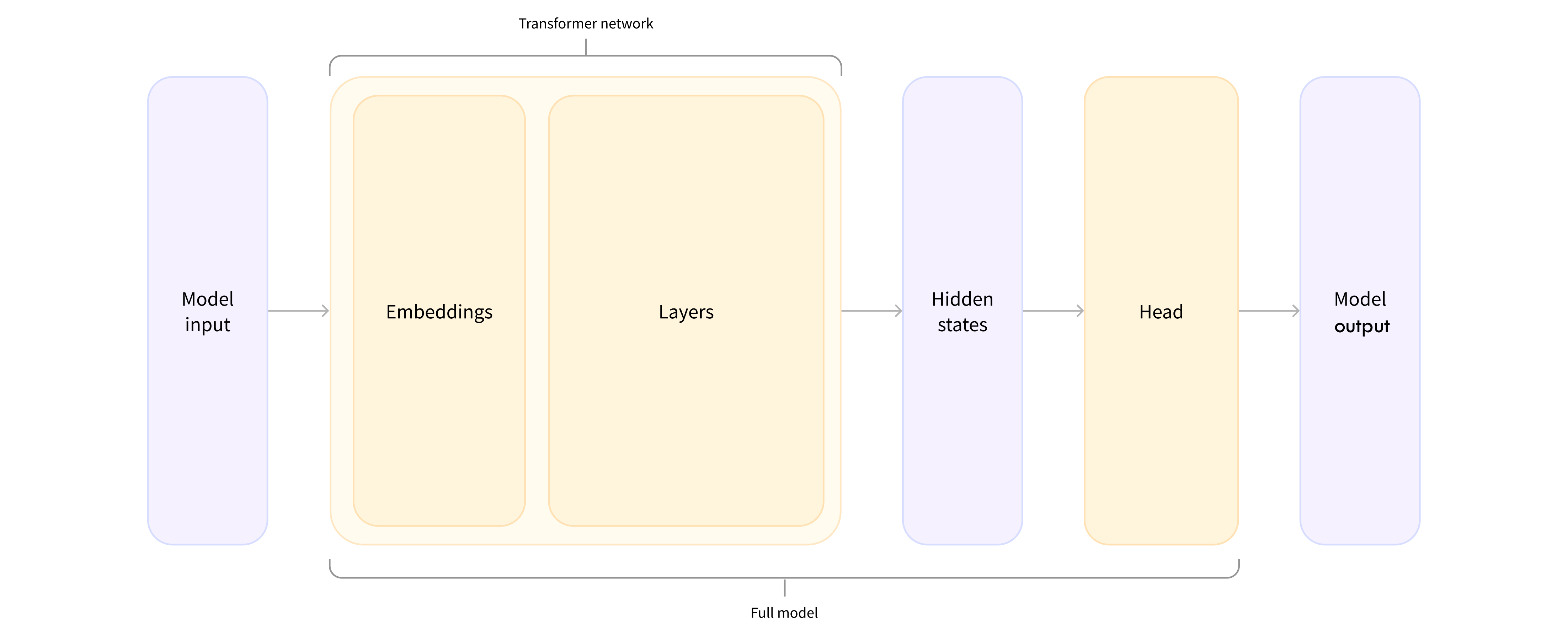
Tokens

[this, course, is, amazing, !]

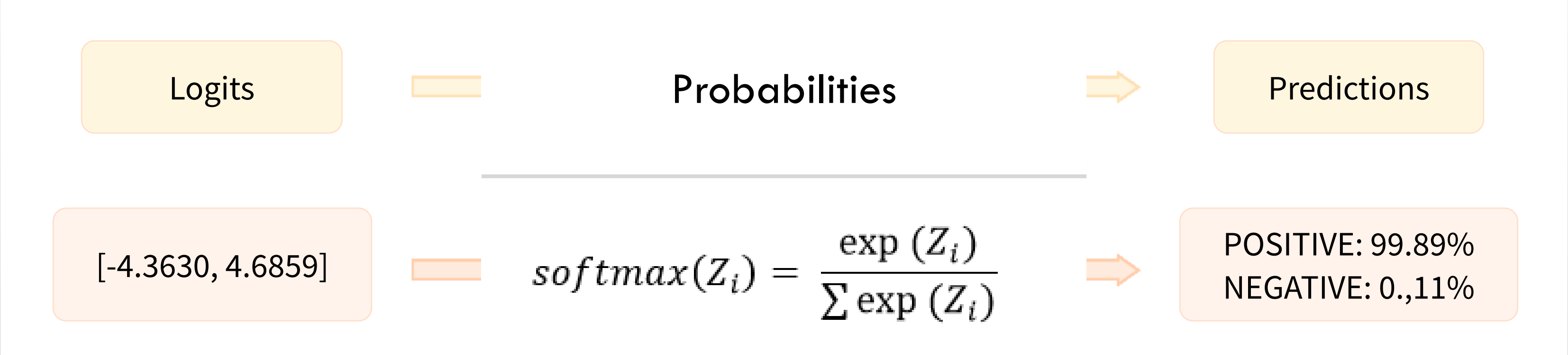
Raw text

This course is amazing!

Model


Transformer network
Model input
Embeddings
Layers
Hidden states
Head
Model output
Full model

Post Processing



Logits

Probabilities

Predictions

[-4.3630, 4.6859]

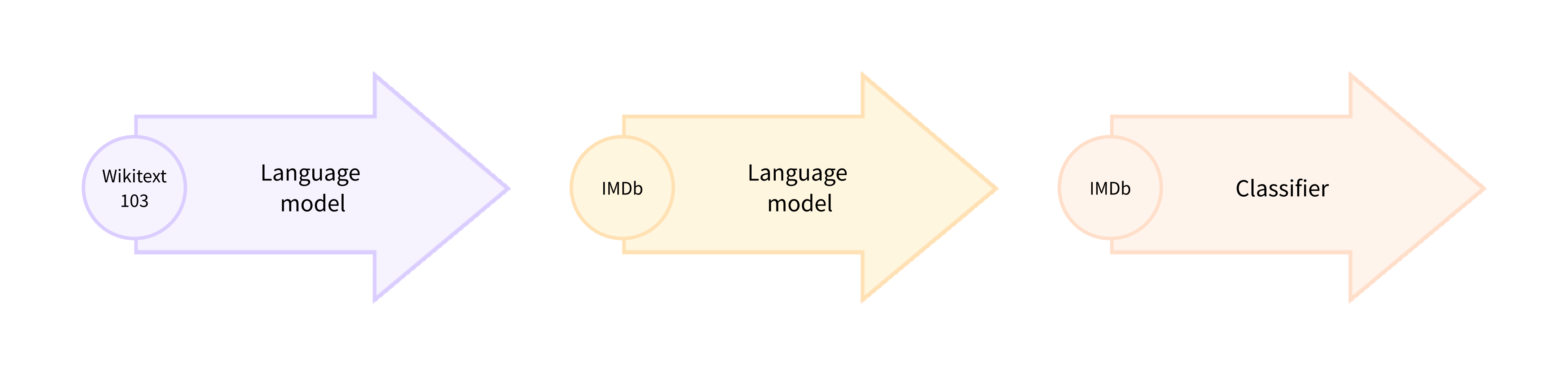
$$softmax(Z_i) = \frac{\exp(Z_i)}{\sum \exp(Z_i)}$$

POSITIVE: 99.89%
NEGATIVE: 0.,11%

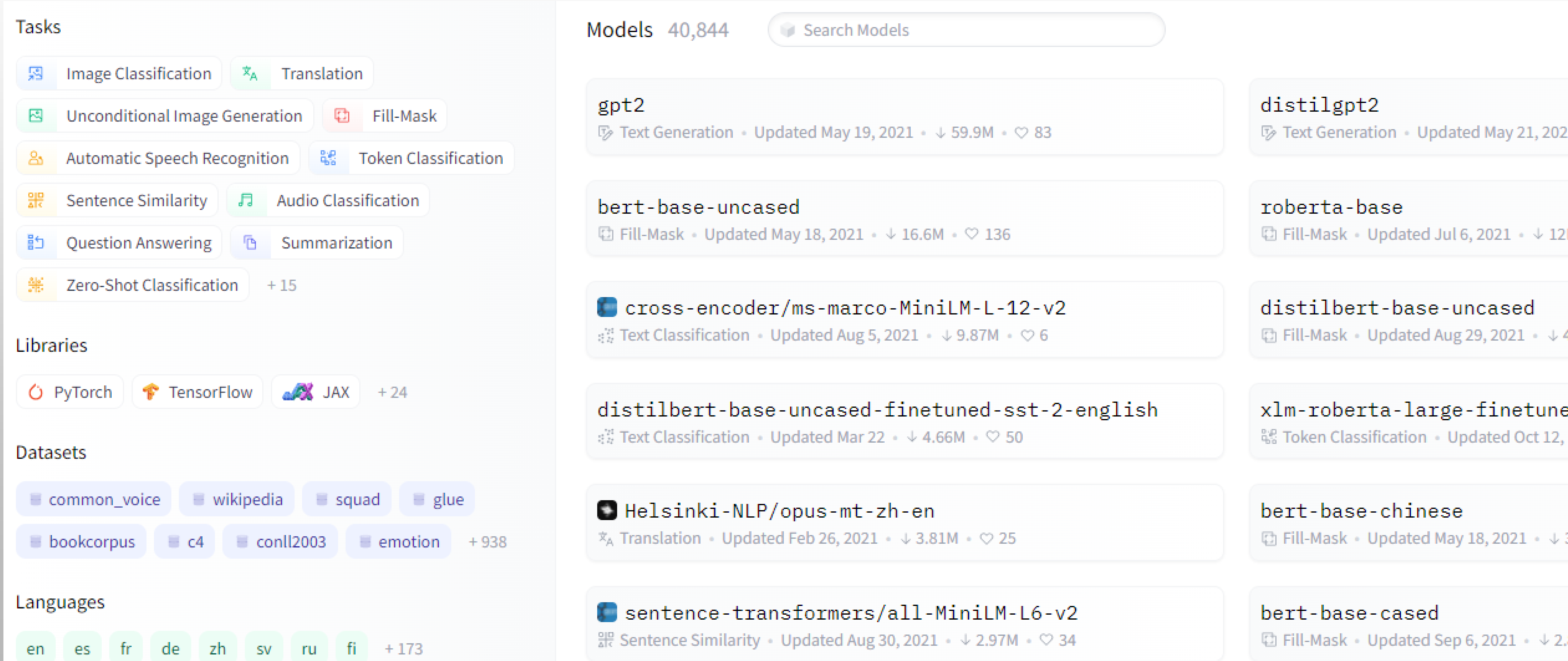
NLP Tasks

- Sequence classification
- Token classification (sequence labelling)
- Masked language modeling (like BERT)
- Summarization
- Translation
- Causal language modeling pretraining (like GPT-2)
- Question answering

Domain Adaptation (through MLM)



Models



Tasks

Image Classification · Translation · Unconditional Image Generation · Fill-Mask · Automatic Speech Recognition · Token Classification · Sentence Similarity · Audio Classification · Question Answering · Summarization · Zero-Shot Classification · + 15

Libraries

PyTorch · TensorFlow · JAX · + 24

Datasets

common_voice · wikipedia · squad · glue · bookcorpus · c4 · conll2003 · emotion · + 938

Languages

en · es · fr · de · zh · sv · ru · fi · + 173

Models 40,844

Search Models

gpt2
Text Generation • Updated May 19, 2021 • ↓ 59.9M • ♡ 83

distilgpt2
Text Generation • Updated May 21, 202

bert-base-uncased
Fill-Mask • Updated May 18, 2021 • ↓ 16.6M • ♡ 136

roberta-base
Fill-Mask • Updated Jul 6, 2021 • ↓ 12

cross-encoder/ms-marco-MiniLM-L-12-v2
Text Classification • Updated Aug 5, 2021 • ↓ 9.87M • ♡ 6

distilbert-base-uncased
Fill-Mask • Updated Aug 29, 2021 • ↓ 4

distilbert-base-uncased-finetuned-sst-2-english
Text Classification • Updated Mar 22 • ↓ 4.66M • ♡ 50

xlm-roberta-large-finetune
Token Classification • Updated Oct 12,

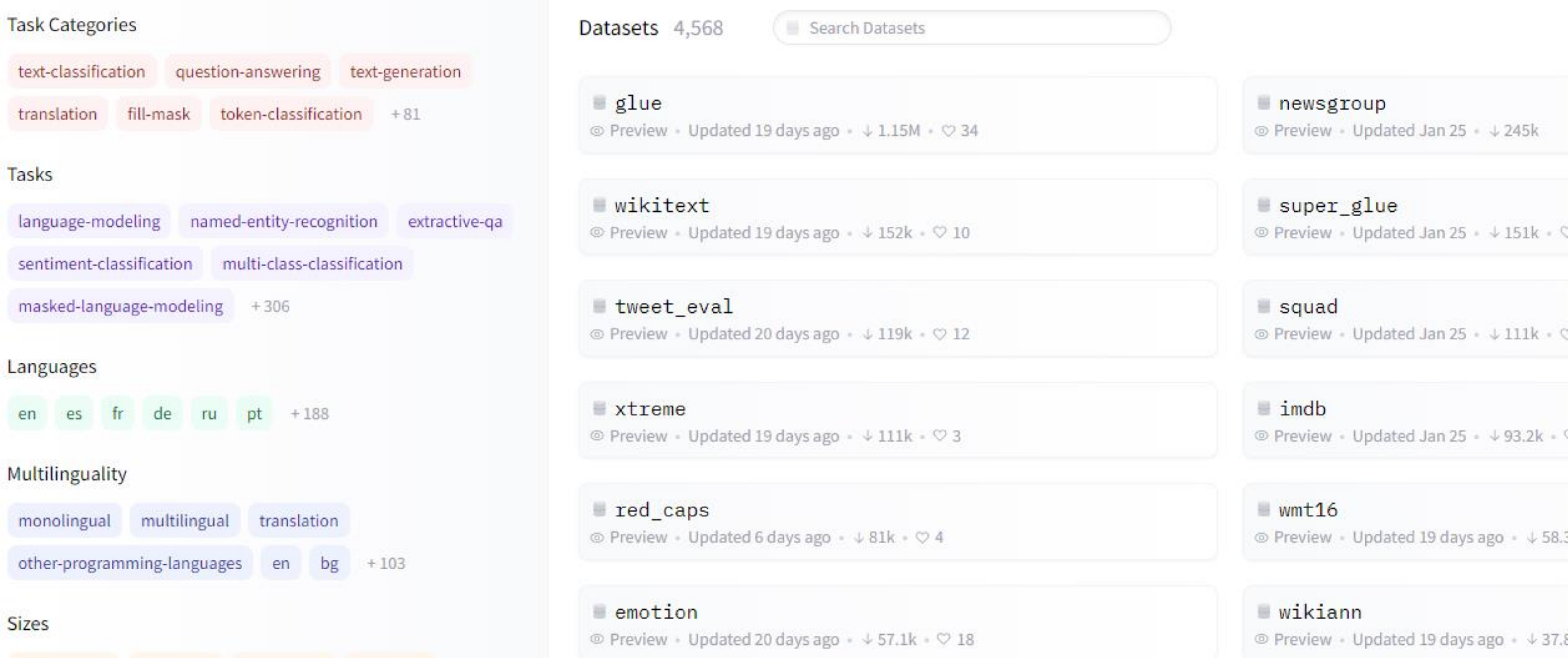
Helsinki-NLP/opus-mt-zh-en
Translation • Updated Feb 26, 2021 • ↓ 3.81M • ♡ 25

bert-base-chinese
Fill-Mask • Updated May 18, 2021 • ↓ 3

sentence-transformers/all-MiniLM-L6-v2
Sentence Similarity • Updated Aug 30, 2021 • ↓ 2.97M • ♡ 34

bert-base-cased
Fill-Mask • Updated Sep 6, 2021 • ↓ 2.

Datasets



Task Categories

text-classification · question-answering · text-generation · translation · fill-mask · token-classification · + 81

Tasks

language-modeling · named-entity-recognition · extractive-qa · sentiment-classification · multi-class-classification · masked-language-modeling · + 306

Languages

en · es · fr · de · ru · pt · + 188

Multilinguality

monolingual · multilingual · translation · other-programming-languages · en · bg · + 103

Sizes

Datasets 4,568 · Search Datasets

glue
Preview • Updated 19 days ago • ↓ 1.15M • ♡ 34

newsgroup
Preview • Updated Jan 25 • ↓ 245k

wikitext
Preview • Updated 19 days ago • ↓ 152k • ♡ 10

super_glue
Preview • Updated Jan 25 • ↓ 151k •

tweet_eval
Preview • Updated 20 days ago • ↓ 119k • ♡ 12

squad
Preview • Updated Jan 25 • ↓ 111k •

xtreme
Preview • Updated 19 days ago • ↓ 111k • ♡ 3

imdb
Preview • Updated Jan 25 • ↓ 93.2k •

red_caps
Preview • Updated 6 days ago • ↓ 81k • ♡ 4

wmt16
Preview • Updated 19 days ago • ↓ 58.3

emotion
Preview • Updated 20 days ago • ↓ 57.1k • ♡ 18

wikiann
Preview • Updated 19 days ago • ↓ 37.8

Explore 🤗

- Hugging Face documentation: https://huggingface.co/docs

- Hugging Face Transformers documentation: https://huggingface.co/docs/transformers/
- Hugging Face course: https://huggingface.co/course/

- Hugging Face models: https://huggingface.co/models
- Hugging Face datasets: https://huggingface.co/datasets