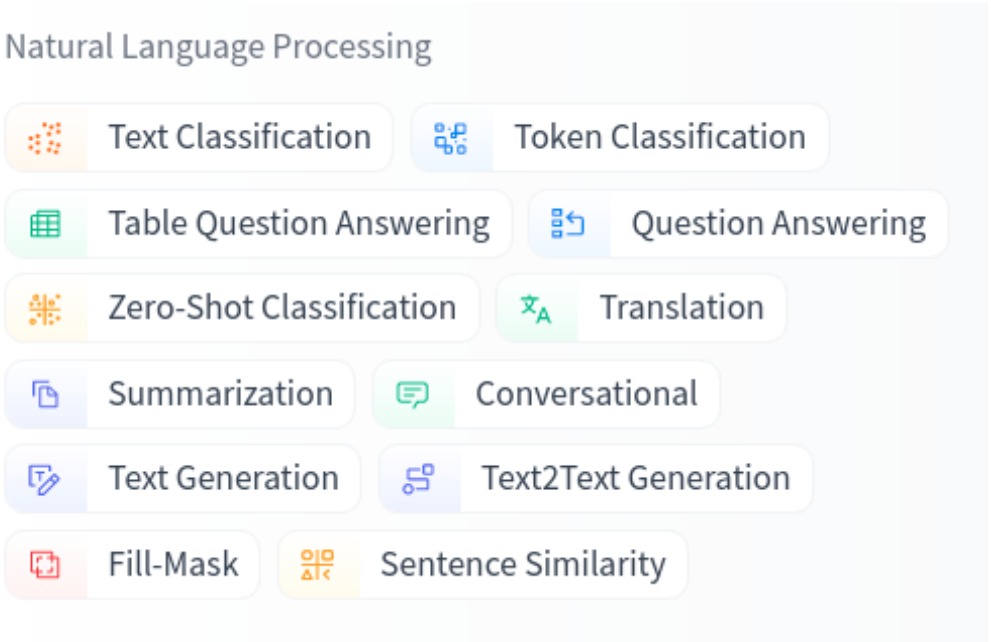
NLP

Natural language processing

NLP tasks

- **Classifying whole sentences**
- **Classifying each word in a sentence**
- **Generating text content**
- **Extracting an answer from a text**
- **Generating a new sentence from an input text**



Go to [hugging face models]

- **Why is id difficult?**
  - The text needs to be processed in a way that enables the model to learn from it.
  - Word meaning

Transformers, what can they do?

- Classify a sentence

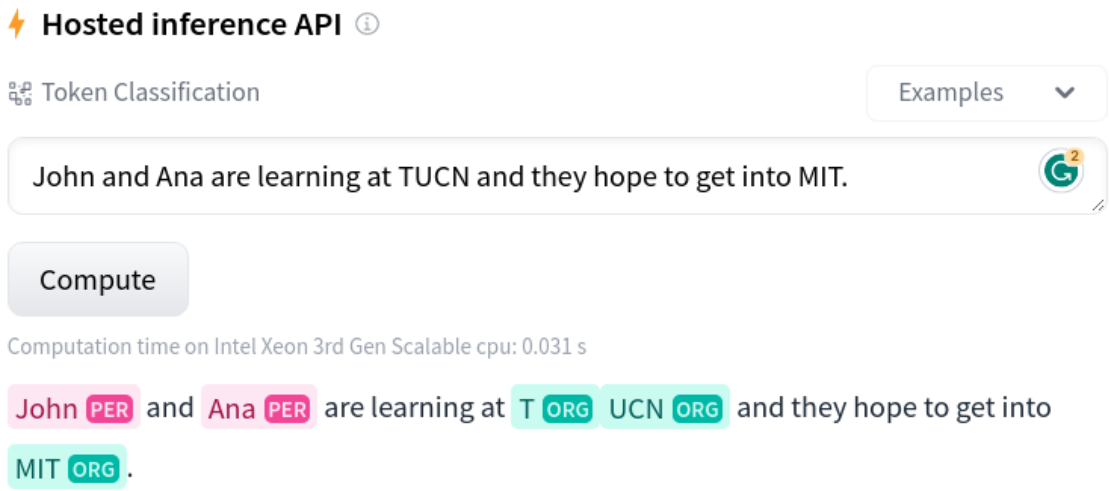

```
from transformers import pipeline

classifier = pipeline("sentiment-analysis")
classifier(
    ["I've been waiting for a HuggingFace course my whole life.",
    "I hate this so much!"])
```

```
[{'label': 'POSITIVE', 'score': 0.9598047137260437},
 {'label': 'NEGATIVE', 'score': 0.9994558095932007}]
```

Transformers, what can th

- Classify a token

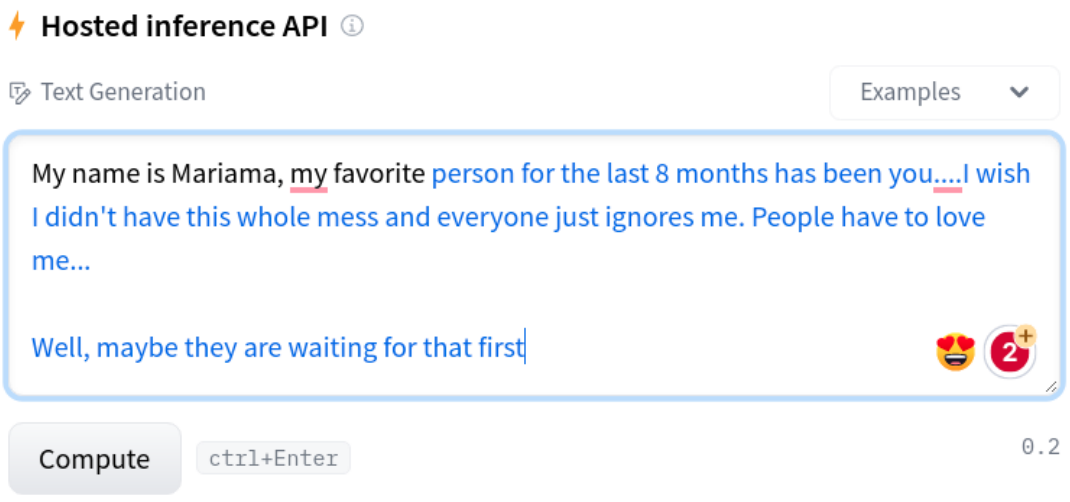



```
from transformers import pipeline

ner = pipeline("ner", grouped_entities=True)
ner("My name is Sylvain and I work at Hugging Face in Brooklyn.")

[{'entity_group': 'PER', 'score': 0.99816, 'word': 'Sylvain', 'start': 11, 'end': 18},
 {'entity_group': 'ORG', 'score': 0.97960, 'word': 'Hugging Face', 'start': 33, 'end': 45},
 {'entity_group': 'LOC', 'score': 0.99321, 'word': 'Brooklyn', 'start': 49, 'end': 57}
]
```

Text generation




```
from transformers import pipeline

generator = pipeline("text-generation")
generator("In this course, we will teach you how to")
```

```
[{'generated_text': 'In this course, we will teach you how to understand and use '
                    'data flow and data interchange when handling user data. We '
                    'will be working with one or more of the most commonly used '
                    'data flows — data flows of various types, as seen by the '
                    'HTTP'}]
```

Mask filling

```
from transformers import pipeline

unmasker = pipeline("fill-mask")
unmasker("This course will teach you all about <mask> models.", top_k=2)
```

```
[{'sequence': 'This course will teach you all about mathematical models.',
  'score': 0.1961983144283294,
  'token': 30412,
  'token_str': ' mathematical'},
 {'sequence': 'This course will teach you all about computational models.',
  'score': 0.04052725434303284,
  'token': 38163,
  'token_str': ' computational'}]
```

Question answering

```
from transformers import pipeline

question_answerer = pipeline("question-answering")
question_answerer(
    question="Where do I work?",
    context="My name is Sylvain and I work at Hugging Face in Brooklyn",
)
```

- Extracts info from the provided context:

```
{'score': 0.6385916471481323, 'start': 33, 'end': 45, 'answer': 'Hugging Face'}
```

Summarization

- Classify a sentence

```
from transformers import pipeline

summarizer = pipeline("summarization")
summarizer(
    """
    America has changed dramatically during recent years. Not only has the number of
    graduates in traditional engineering disciplines such as mechanical, civil,
    electrical, chemical, and aeronautical engineering declined, but in most of
    the premier American universities engineering curricula now concentrate on
    and encourage largely the study of engineering science. As a result, there
```

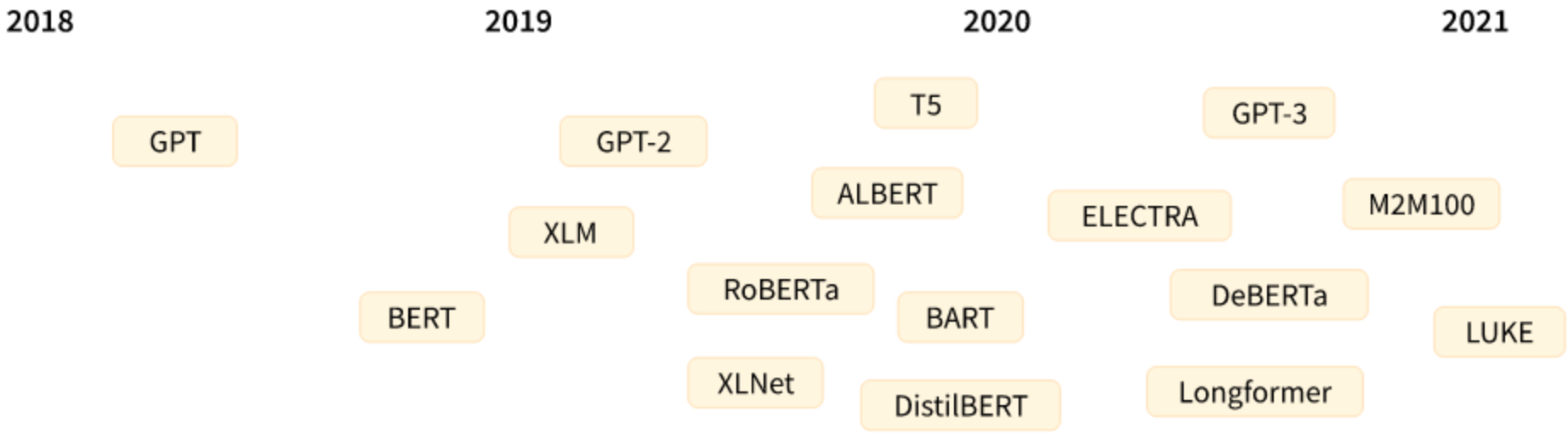
```
[{'summary_text': ' America has changed dramatically during recent years . The  number of engineering graduates
in the U.S. has declined in traditional engineering disciplines such as mechanical, civil, electrical, chemical, and
aeronautical engineering . Rapidly  developing economies such as China and India, as well as other  industrial
countries in Europe and Asia, continue to encourage  and advance engineering .'}]
```

Translation

- from transformers import pipeline
-
- translator = pipeline("translation", model="Helsinki-NLP/opus-mt-fr-en")
- translator("Ce cours est produit par Hugging Face.")

- [{'translation_text': 'This course is produced by Hugging Face.'}]

Era transformers



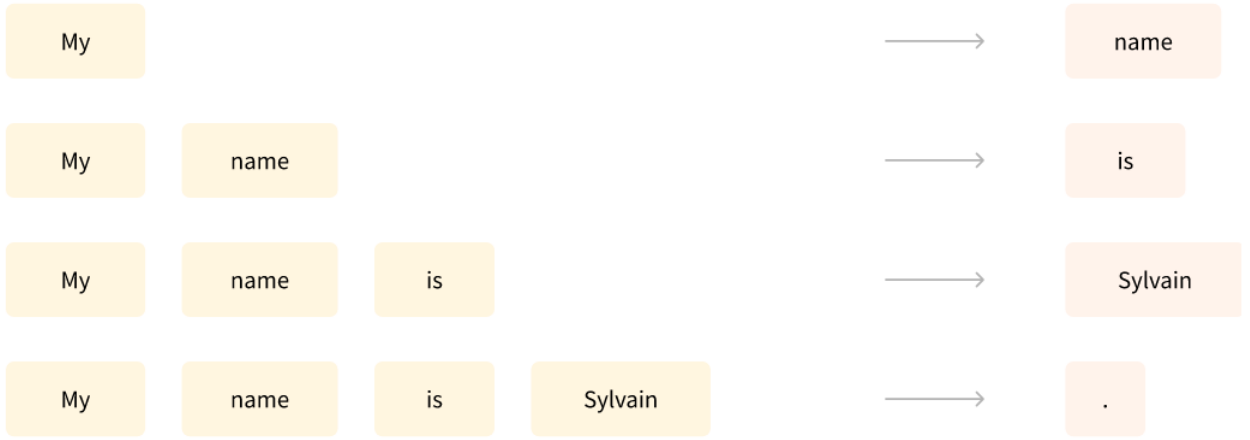
3 mari categorii:

- GPT-like- (also called *auto-regressive* Transformer models
- BERT-like - also called *auto-encoding* Transformer models
- BART/T5-like - also called *sequence-to-sequence* Transformer models

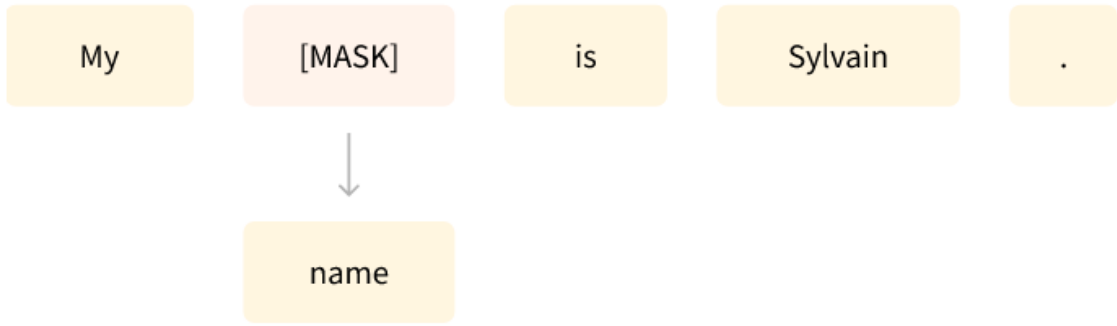
Language models

- Statistical understanding of the text
- Trained on large amounts of raw text in **self-supervised learning**
  - **Pretrain – no adaptation to a specifical practical task**
  - **Fine tune – adapted to a specifical practical task (supervised)**
- **Pretraining tasks:**
  - **Causal language model – predict the next word**
  - **Masked language model – mask some words and predict them**
  - **Next sentence prediction – introduce 2 sentences and predict whether the second is a next sentence for the first one**

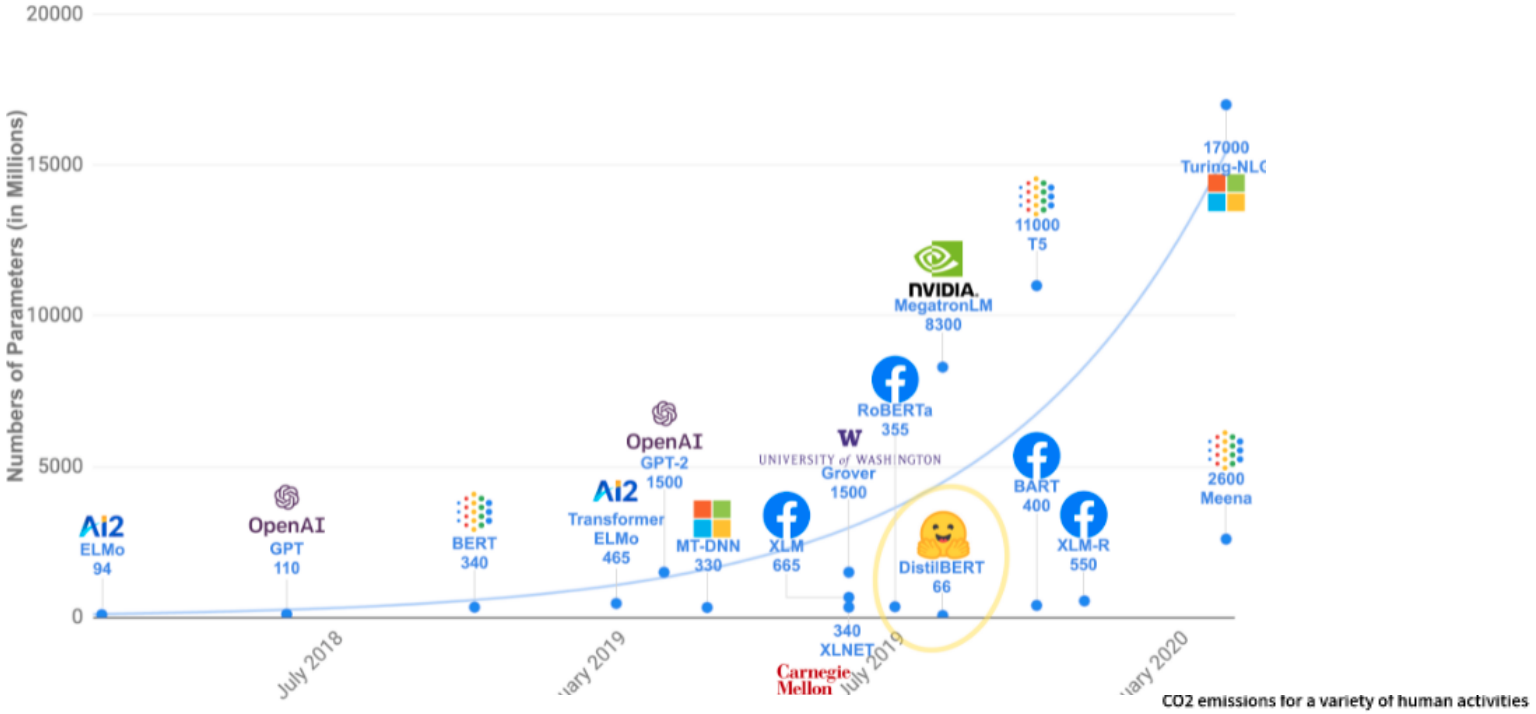
- Causal language model

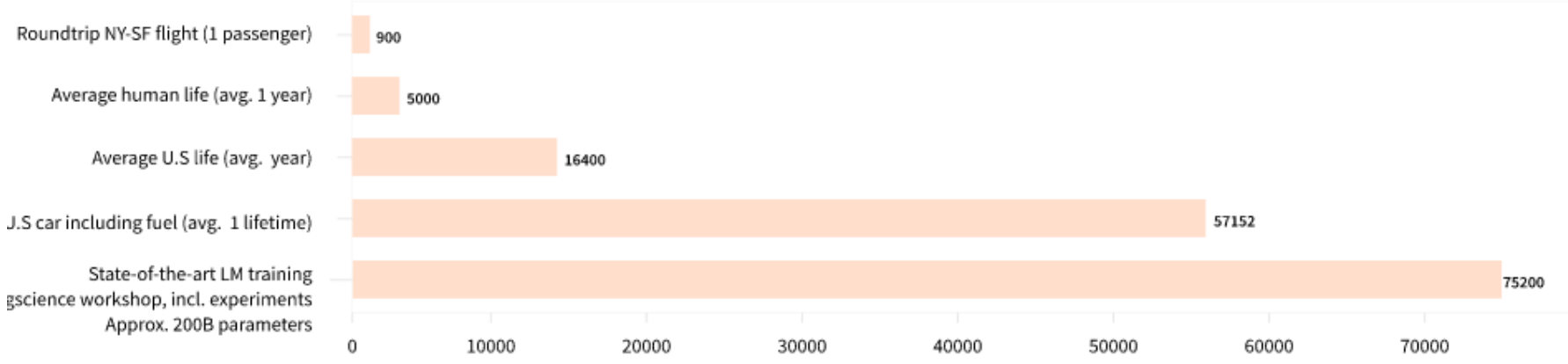


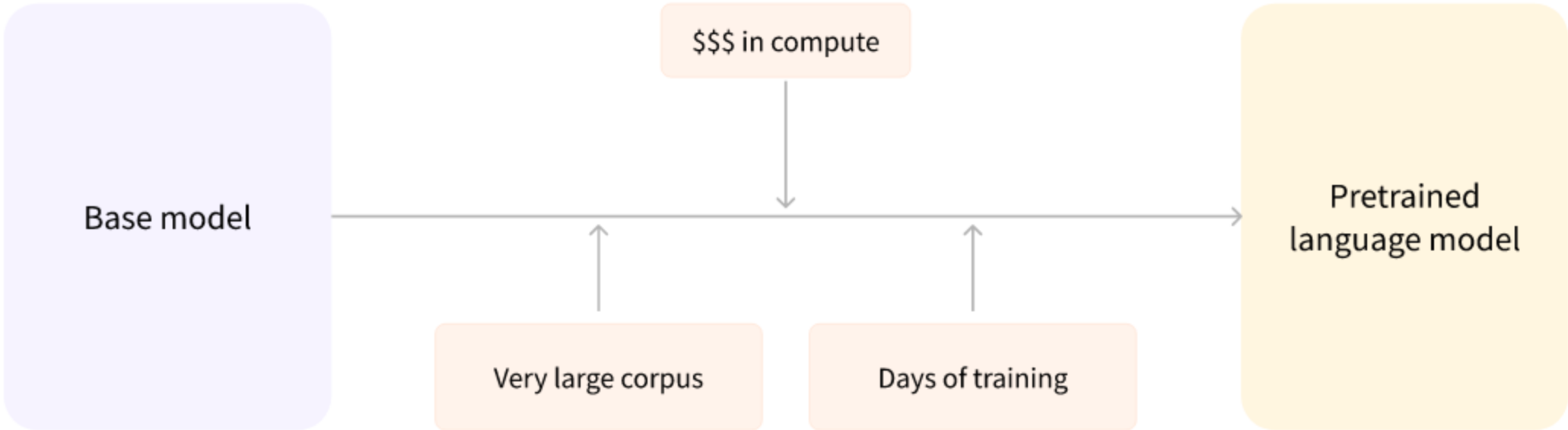
- Masked language model

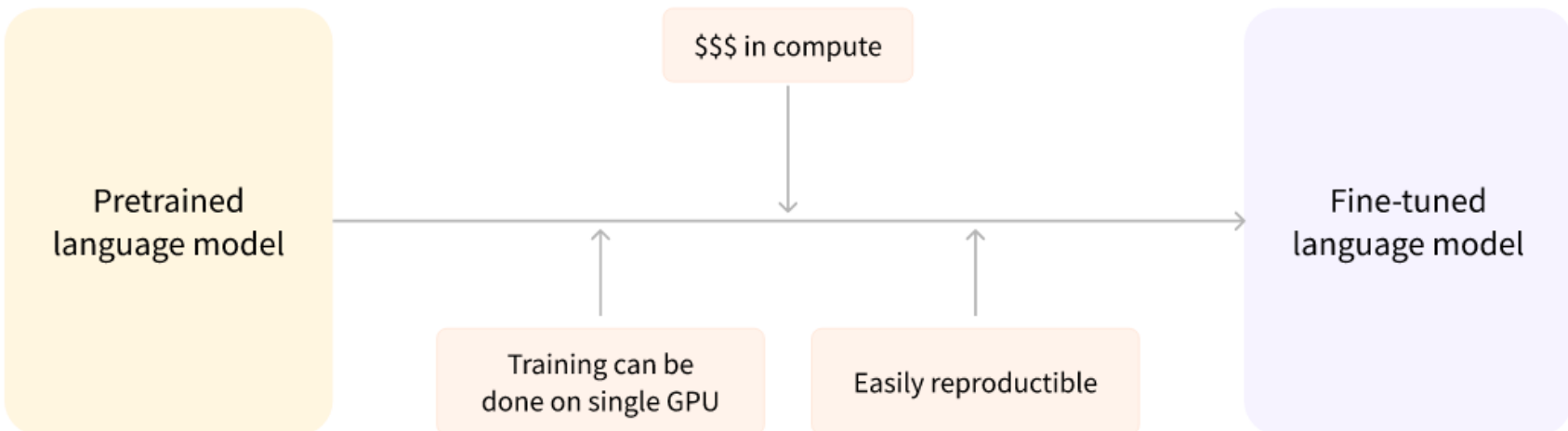


Language models are large models






Base model
$$$ in compute
Very large corpus
Days of training
Pretrained language model


Pretrained language model
$$$ in compute
Training can be done on single GPU
Easily reproductible
Fine-tuned language model