



COMP 336 I Natural Language Processing

Lecture 10: Pre-training and large language models (LLMs)

Spring 2024

Announcements

- TA office Hour: Thursday 9 am - 10:15 am. [Book online](#)
- Get started on assignment 2 ASAP!
 - Join [#assignment-2](#) Slack channel for discussion

Lecture plan

- Neural language models: recap
- Traditional to modern NLP
 - Traditional learning paradigm
 - Supervised training/fine-tuning only, NO pre-training
 - Modern learning paradigm
 - Pretrain + fine-tuning, pretrain + prompting/in-context learning
- Pretraining overview
- BERT pretraining

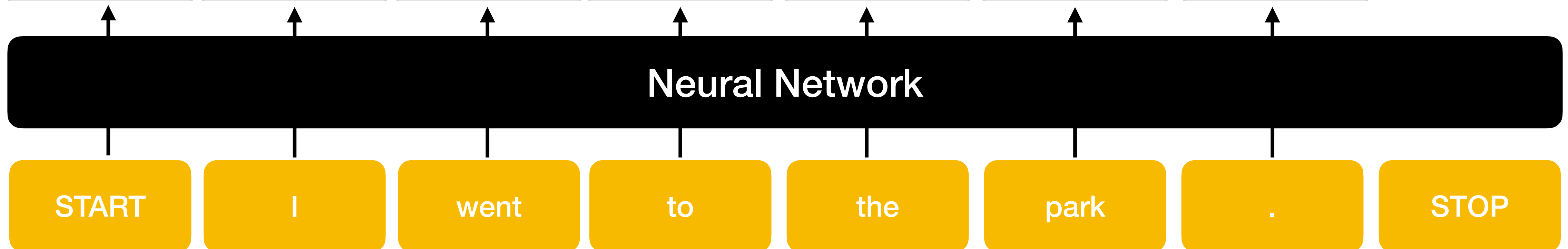
Neural language models: recap

Neural language models: overview

- **Input:** sequences of words (or tokens)
- **Output:** probability distribution over the next word (token)

$p(x|\text{START})$
 $p(x|\text{START I})$
 $p(x|\dots \text{went})$
 $p(x|\dots \text{to})$
 $p(x|\dots \text{the})$
 $p(x|\dots \text{park})$
 $p(x|\text{START I went to the park.})$

The 3	think 11%	to 35%	the 29%	bathroo 3%	and 14%	I 21%
When 2.5%	was 5%	back 8%	a 9%	doctor 2%	with 9	It 6
They 2%	went 2%	into 5%	see 5%	hospita 2%	, 8%	The 3%
... ..	am 1%	through 4%	my 3%	store 1.5%	to 7%	There 3%
I 1%	will 1%	out 3%	bed 2%
... ..	like 0.5%	on 2%	school 1%	park 0.5%	. 6%	STOP 1%
Banana 0.1%%



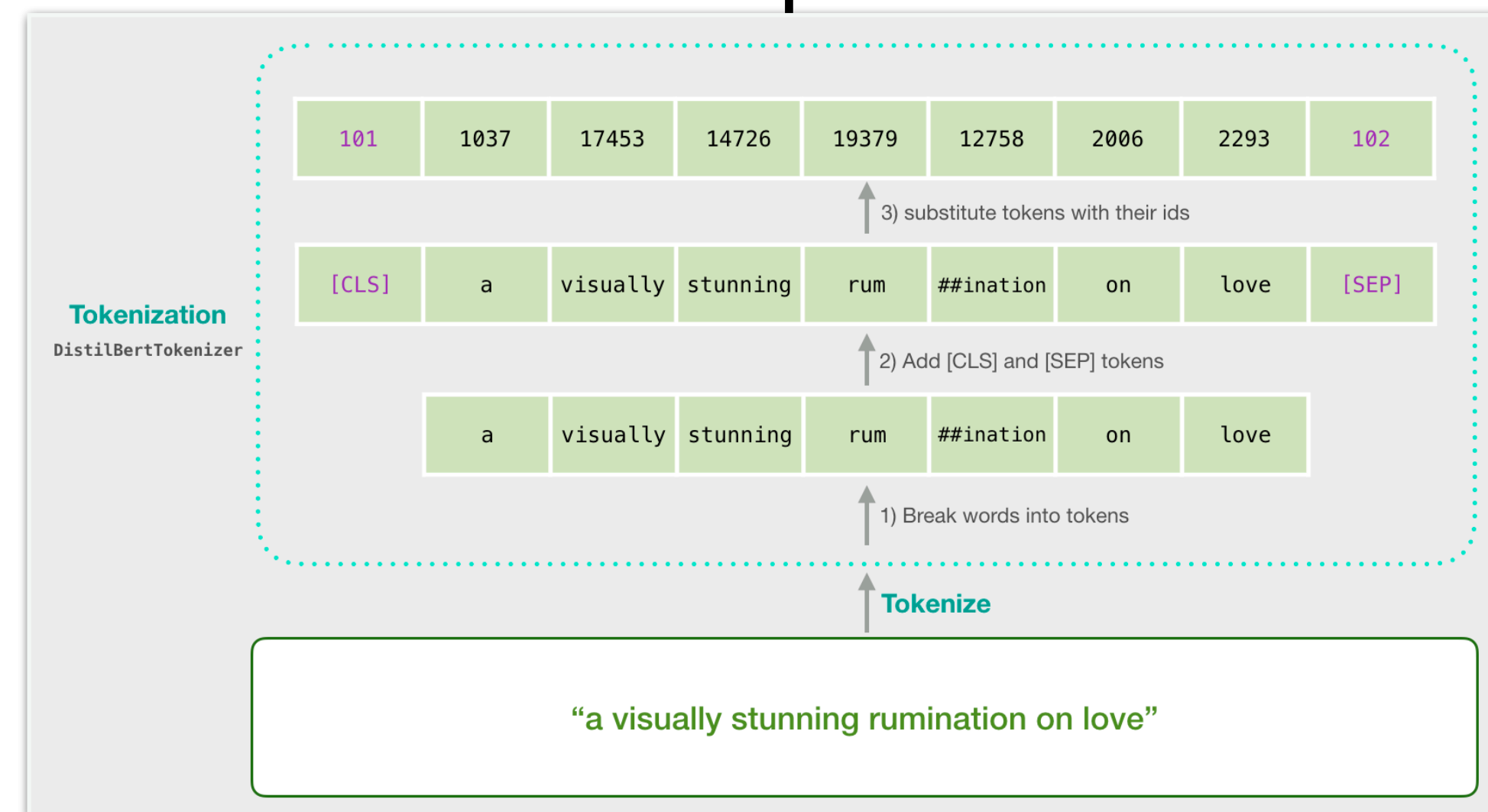
Neural language models: tokenization

$$p(x|\text{START}) p(x|\text{START I}) p(x|\dots \text{went}) p(x|\dots \text{to}) p(x|\dots \text{the}) p(x|\dots \text{park}) p(x|\text{START I went to the park.})$$

Neural Network

Mapping each tokenized id into its corresponding embeddings

Tokenization:



Call me Ishmael. Some years ago—never mind how long precisely—having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and methodically knocking people's hats off—then, I account it high time tozz get to sea as soon as I can. This is my substitute for pistol and ball. With a philosophical flourish Cato throws himself upon his sword; I quietly take to the ship. There is nothing surprising in this. If they but knew it, almost all men in their degree, some time or other, cherish very nearly the same feelings towards the ocean with me.

TEXT TOKEN IDS

START

I

went

to

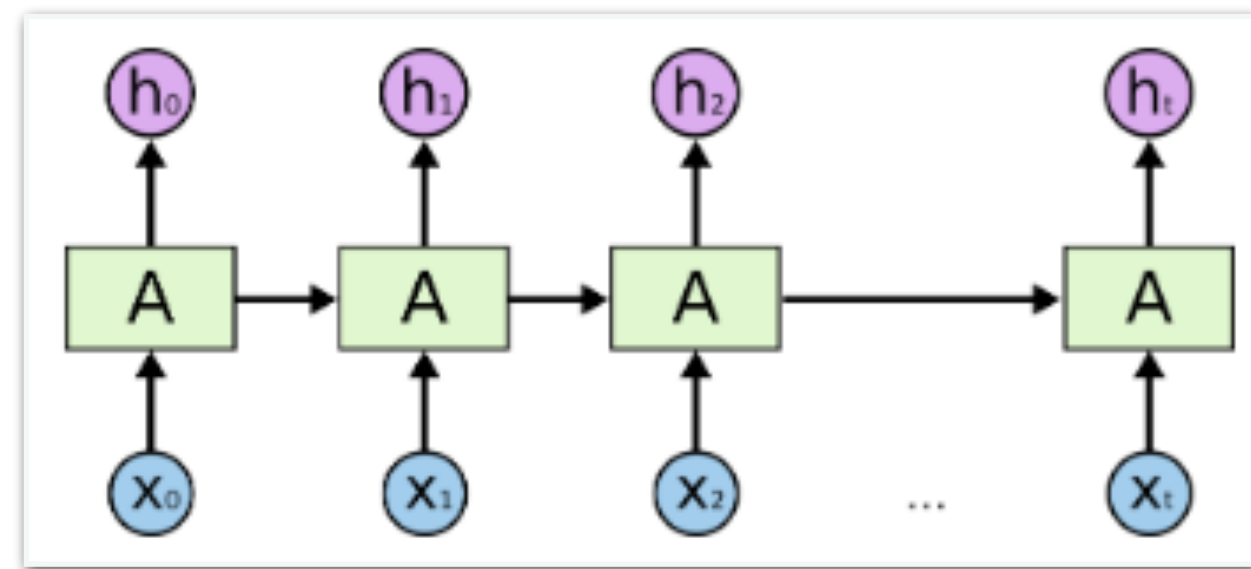
the

park

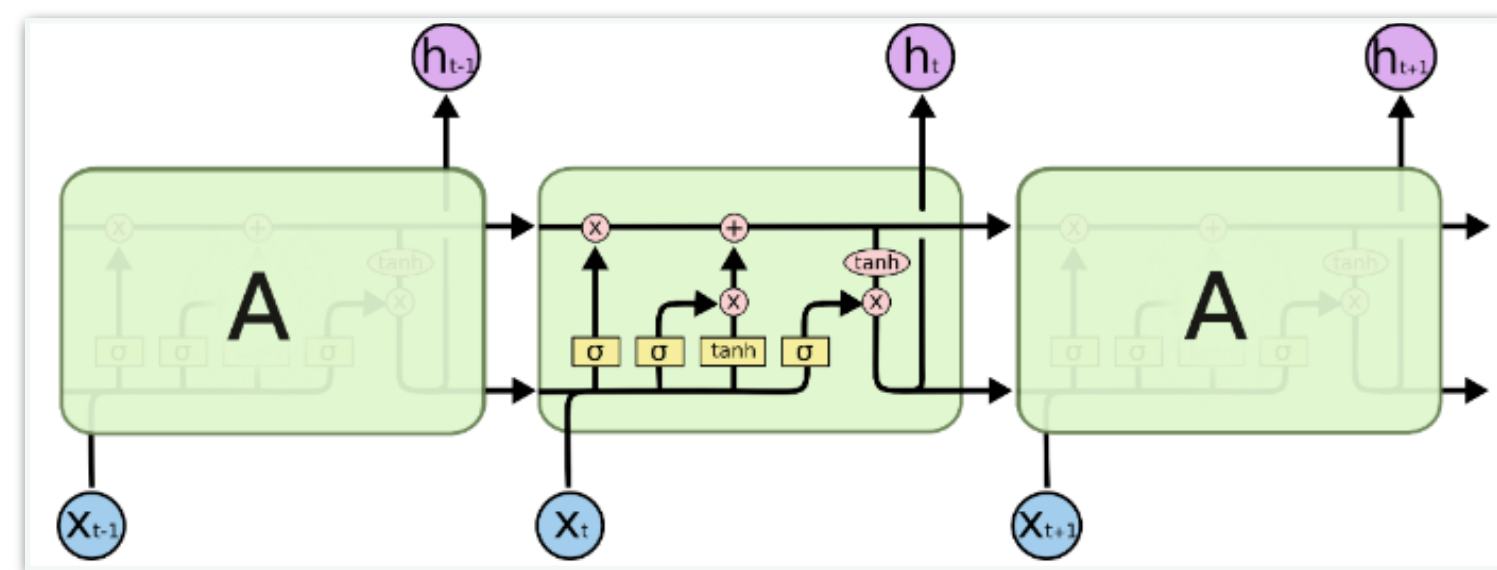
.

STOP

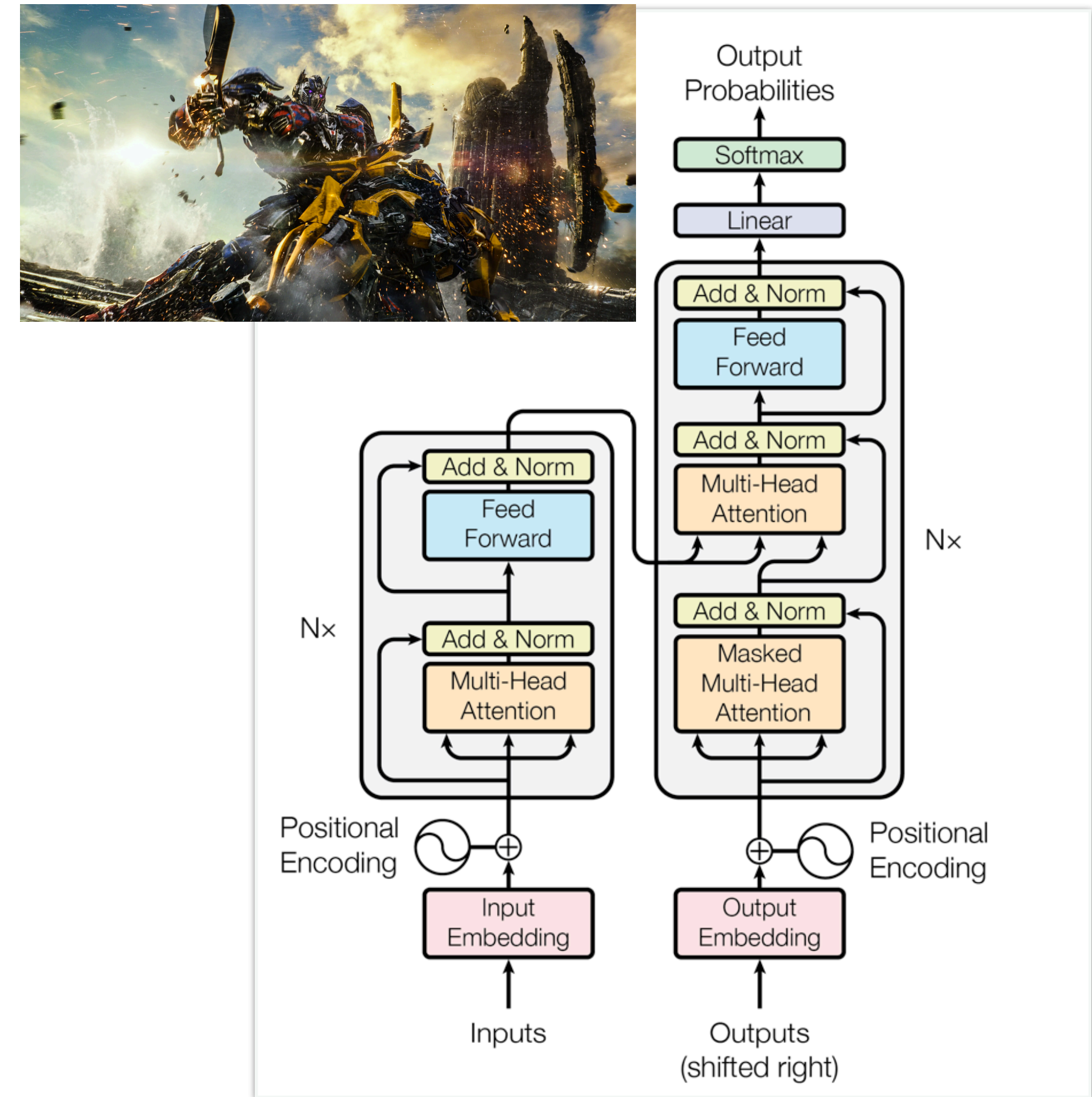
Neural language models: neural networks



RNNs



LSTM



Transformers



Traditional to modern NLP

N-gram language models



Neural language models: BERT, GPT

Traditional models: Naive Bayes

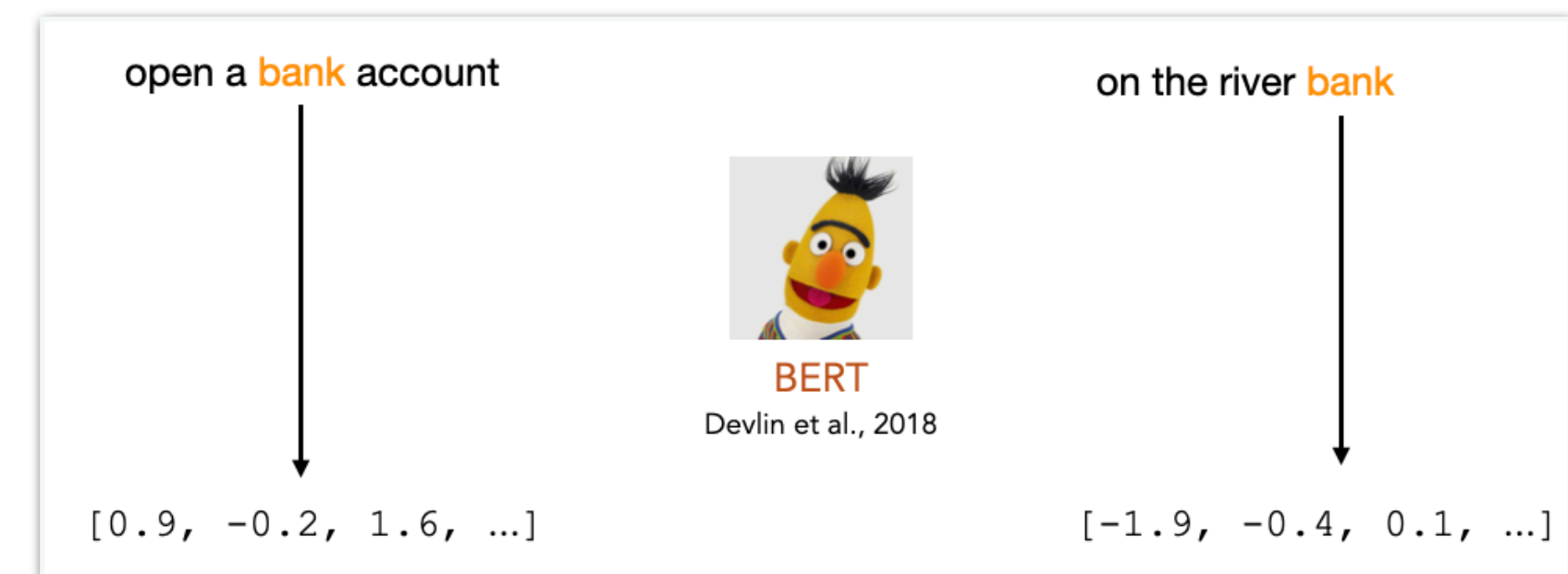
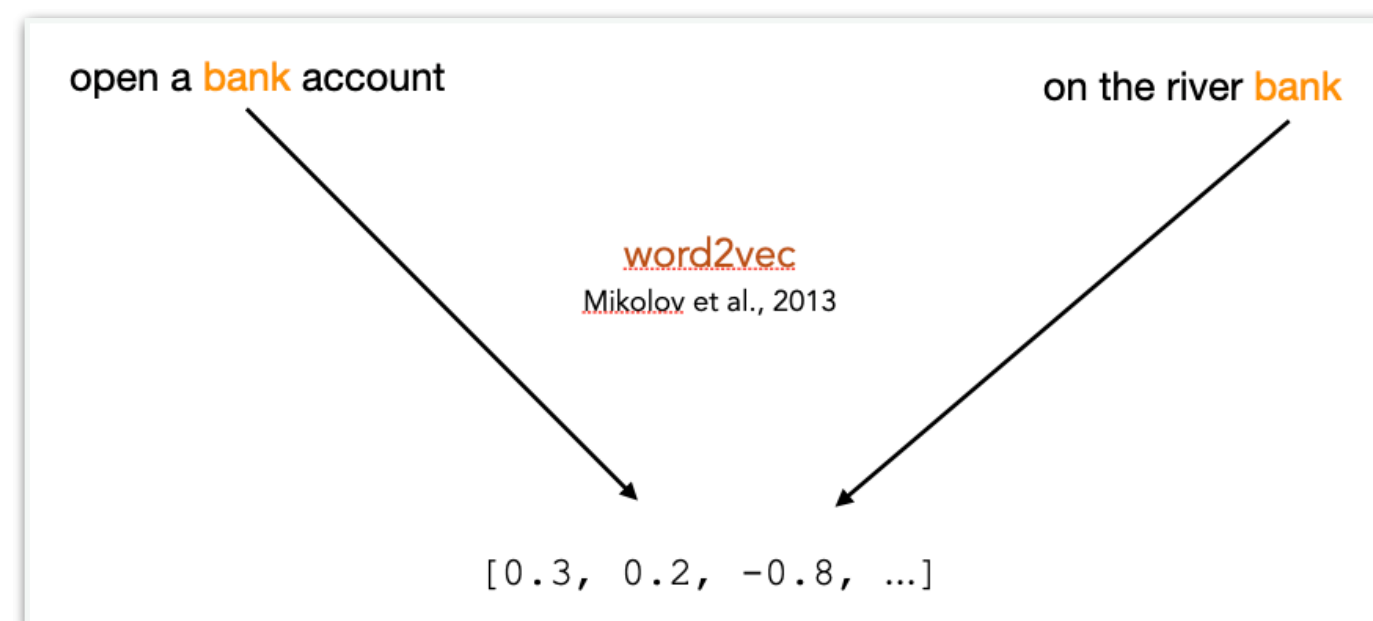


Neural models: Transformers

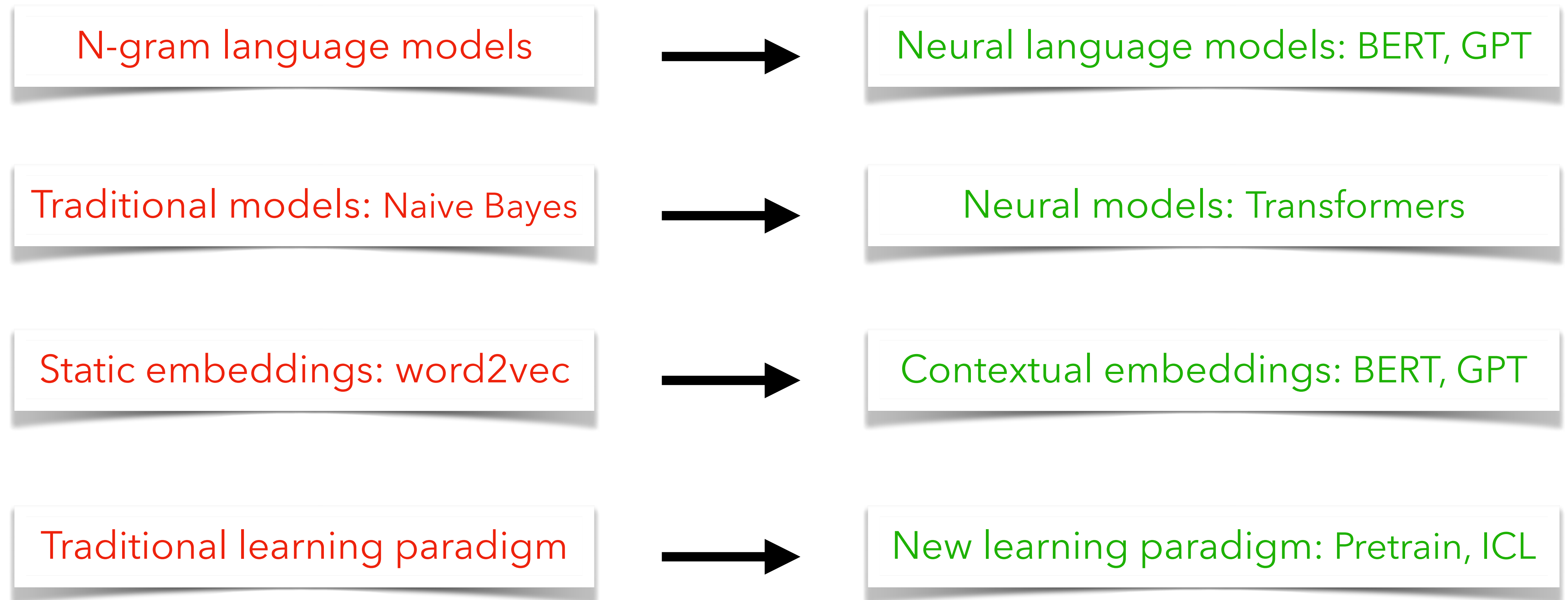
Static embeddings: word2vec



Contextual embeddings: BERT, GPT



Traditional to modern NLP: training paradigm



Question: How to train and use neural language models for different NLP tasks?

Training models for NLP tasks

Foundational Technologies

- Language Modeling
- Part-of-speech Tagging
- Syntactic Parsing
- Dependency Parsing
- Named Entity recognition
- Coreference resolution
- Word Sense Disambiguation
- Semantic Role Labelling
-

High-Level Tasks and Applications

- Sentiment Analysis
- Information Extraction
- Machine Translation
- Question Answering
- Semantic Parsing
- Summarization
- Dialogue systems
- Language and Vision
- Data-to-Text Generation
-

Input X	Output Y	Task
Text	Label	Text Classification (e.g., Sentiment Analysis)
Text	Linguistic Structure	Structured Prediction (e.g., Part-of-Speech Tagging)
Text	Text	Text Generation (e.g., Translation, Summarization)

Example: Training Transformers for sentiment analysis

Task:



Model:



Transformers

Traditional learning paradigm

- **Supervised training/fine-tuning only, NO pre-training**
 - Collect (x, y) task training pairs

Data:

sentence	label
a stirring , funny and finally transporting re imagining of beauty and the beast and 1930s horror films	1
apparently reassembled from the cutting room floor of any given daytime soap	0
they presume their audience won't sit still for a sociology lesson	0
this is a visually stunning rumination on love , memory , history and the war between art and commerce	1
jonathan parker 's bartleby should have been the be all end all of the modern office anomie films	1

Traditional learning paradigm

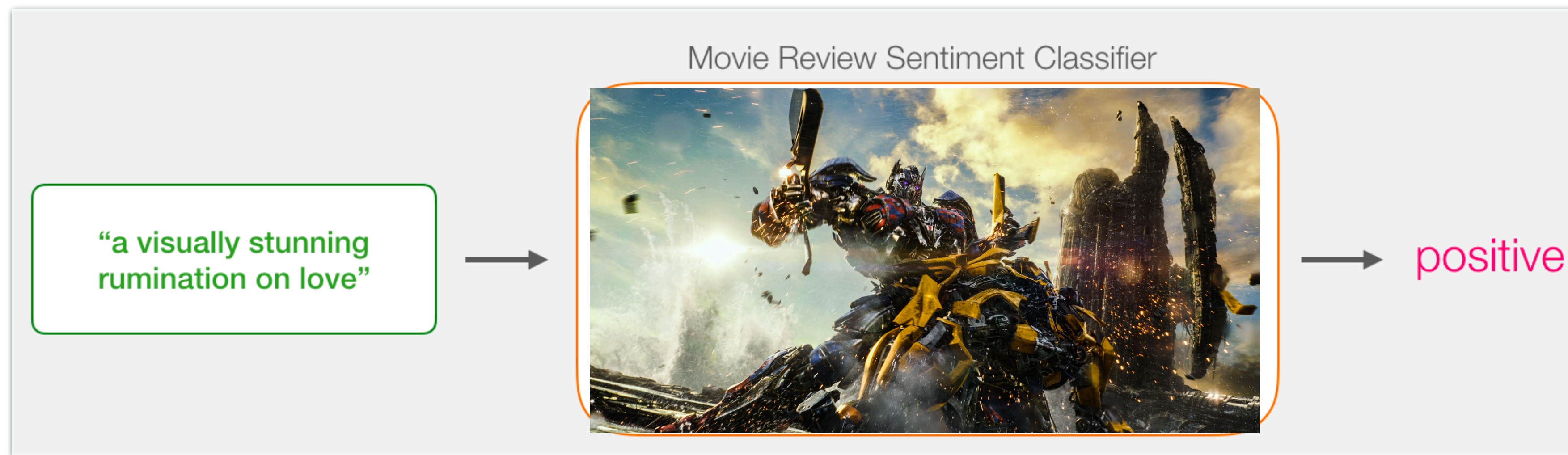
- **Supervised training/fine-tuning only, NO pre-training**
 - Collect (x, y) task training pairs
 - Randomly initialize your models $f(x)$ (e.g., vanilla Transformers)



Randomly initialized Transformers
NO pretrained parameters used

Traditional learning paradigm

- **Supervised training/fine-tuning only, NO pre-training**
 - Collect (x, y) task training pairs
 - Randomly initialize your models $f(x)$ (e.g., vanilla Transformers)
 - Train $f(x)$ on (x, y) pairs



Train Transformers only on task labeled data

Traditional learning paradigm

- **Supervised training/fine-tuning only, NO pre-training**
 - Collect (x, y) task training pairs
 - Randomly initialize your models $f(x)$ (e.g., vanilla Transformers)
 - Train $f(x)$ on (x, y) pairs



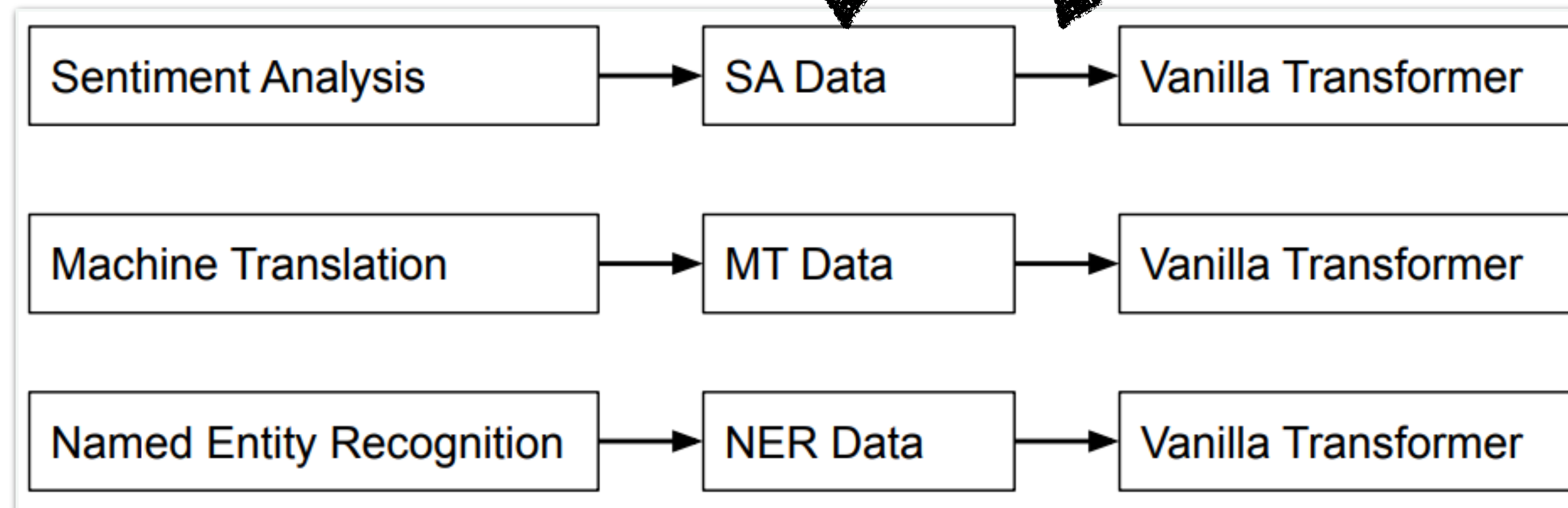
Then you get a trained Transformers **ONLY** for sentiment analysis
The model can be: NB, LR, RNNs, LSTM too

Traditional learning paradigm

- **Supervised training/fine-tuning only, NO pre-training**
 - Train Transformer or other models separately for each task

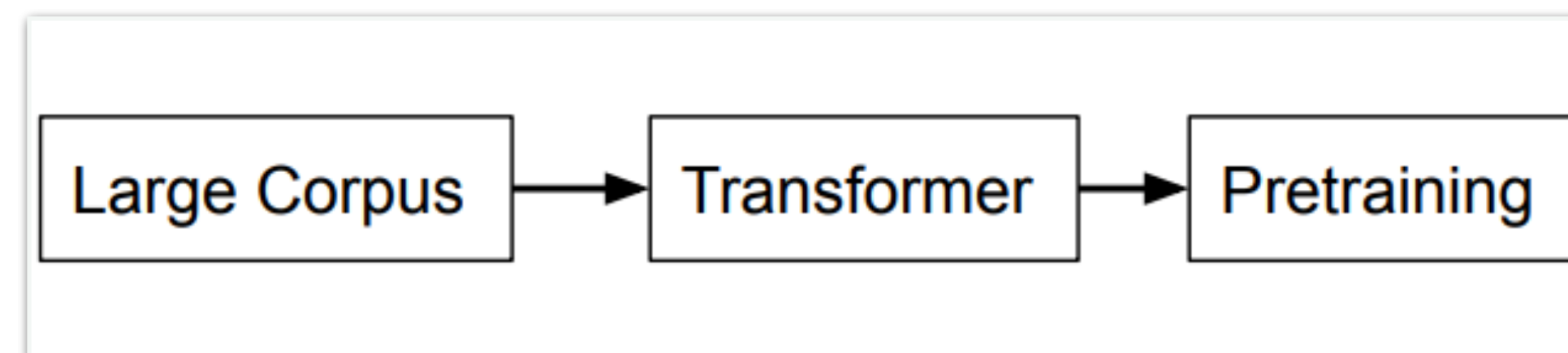
sentence	label
a stirring , funny and finally transporting re imagining of beauty and the beast and 1930s horror films	1
apparently reassembled from the cutting room floor of any given daytime soap	0
they presume their audience won't sit still for a sociology lesson	0
this is a visually stunning rumination on love , memory , history and the war between art and commerce	1
jonathan parker 's bartleby should have been the be all end all of the modern office anomie films	1

Supervised fine-tuning only
NO pre-training



Modern learning paradigm

- **Pre-training + supervised training/fine-tuning**
 - First train Transformer using a lot of general text using unsupervised learning. This is called **pretraining**.

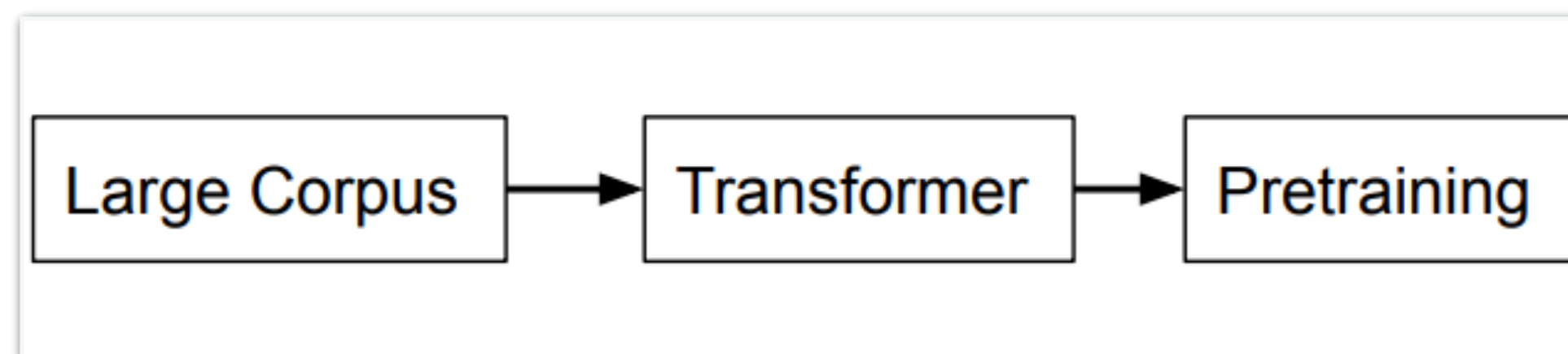


Modern learning paradigm

- **Pre-training + supervised training/fine-tuning**
 - First train Transformer using a lot of general text using unsupervised learning. This is called **pretraining**.



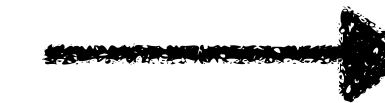
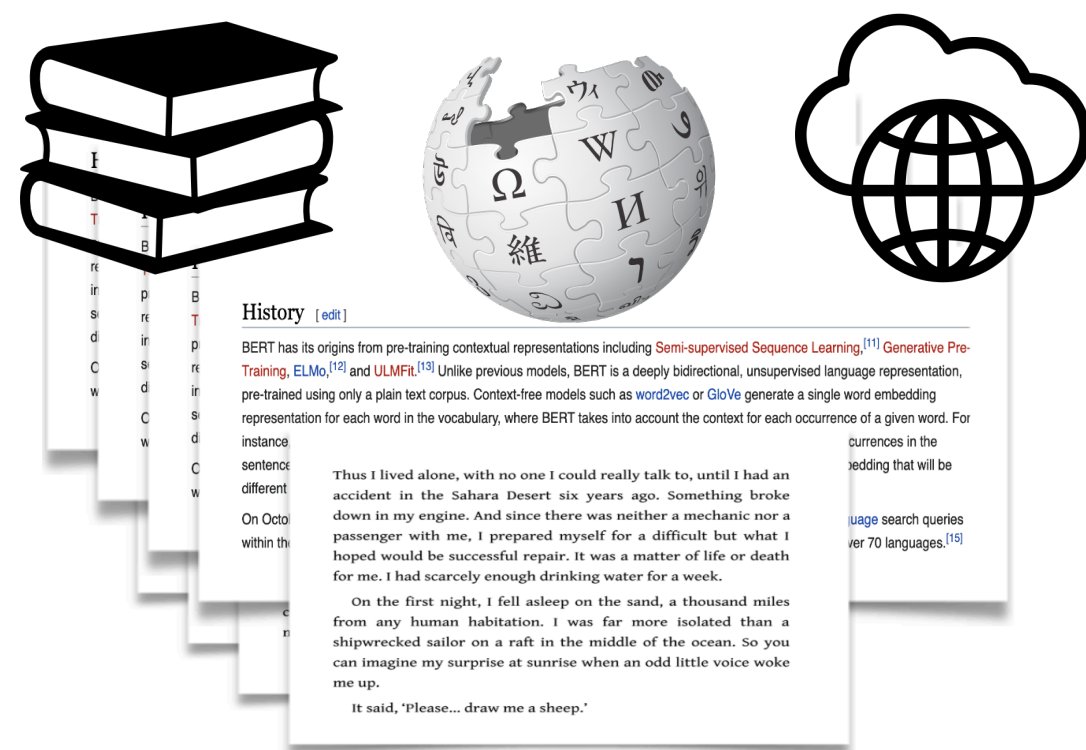
1. Randomly initialized Transformers



Modern learning paradigm

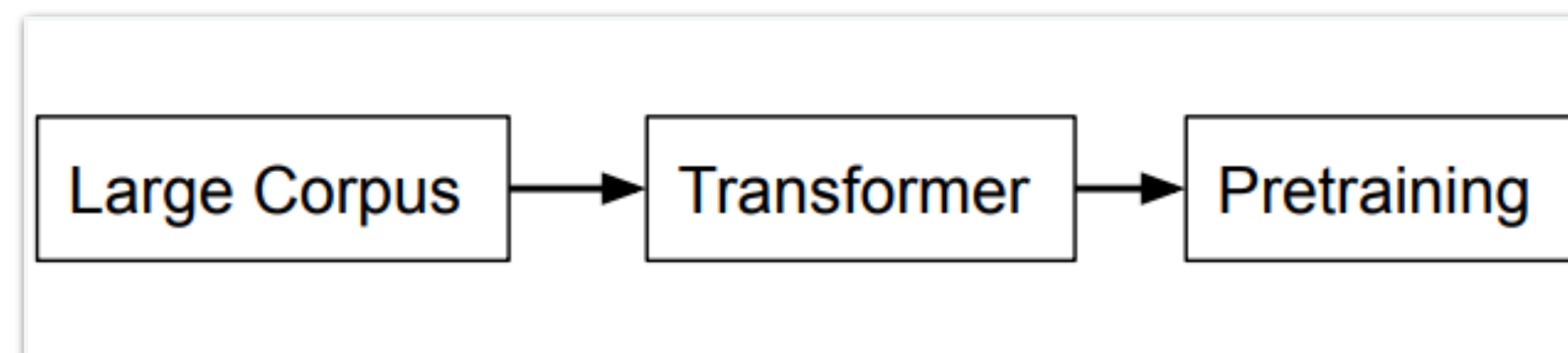
- **Pre-training + supervised training/fine-tuning**

- First train Transformer using a lot of general text using unsupervised learning. This is called **pretraining**.



general training objectives?

2. Pre-train Transformers on large text data with **general objectives**



>2M GPU hours

Modern learning paradigm

- **Pre-training + supervised training/fine-tuning**

- First train Transformer using a lot of general text using unsupervised learning. This is called **pretraining**.
- Then train the pretrained Transformer for a specific task using supervised learning. This is called **finetuning**.

Pretrained Transformers



Sentiment Analysis

Finetuning

Machine Translation

Finetuning

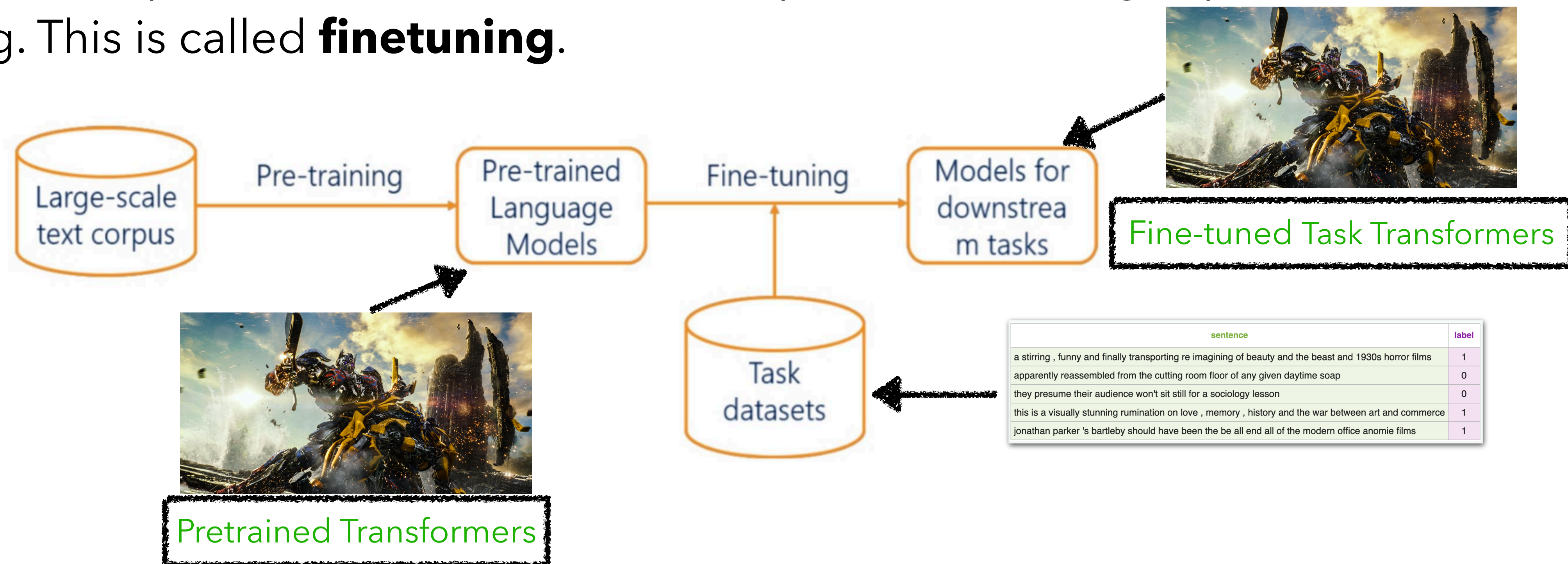
Named Entity Recognition

Finetuning

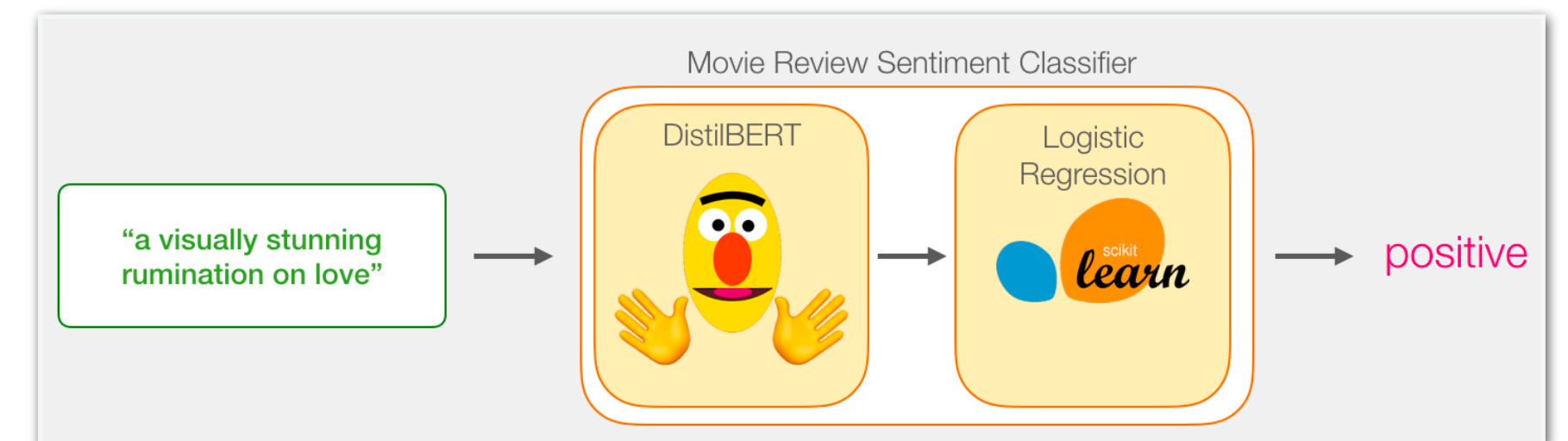
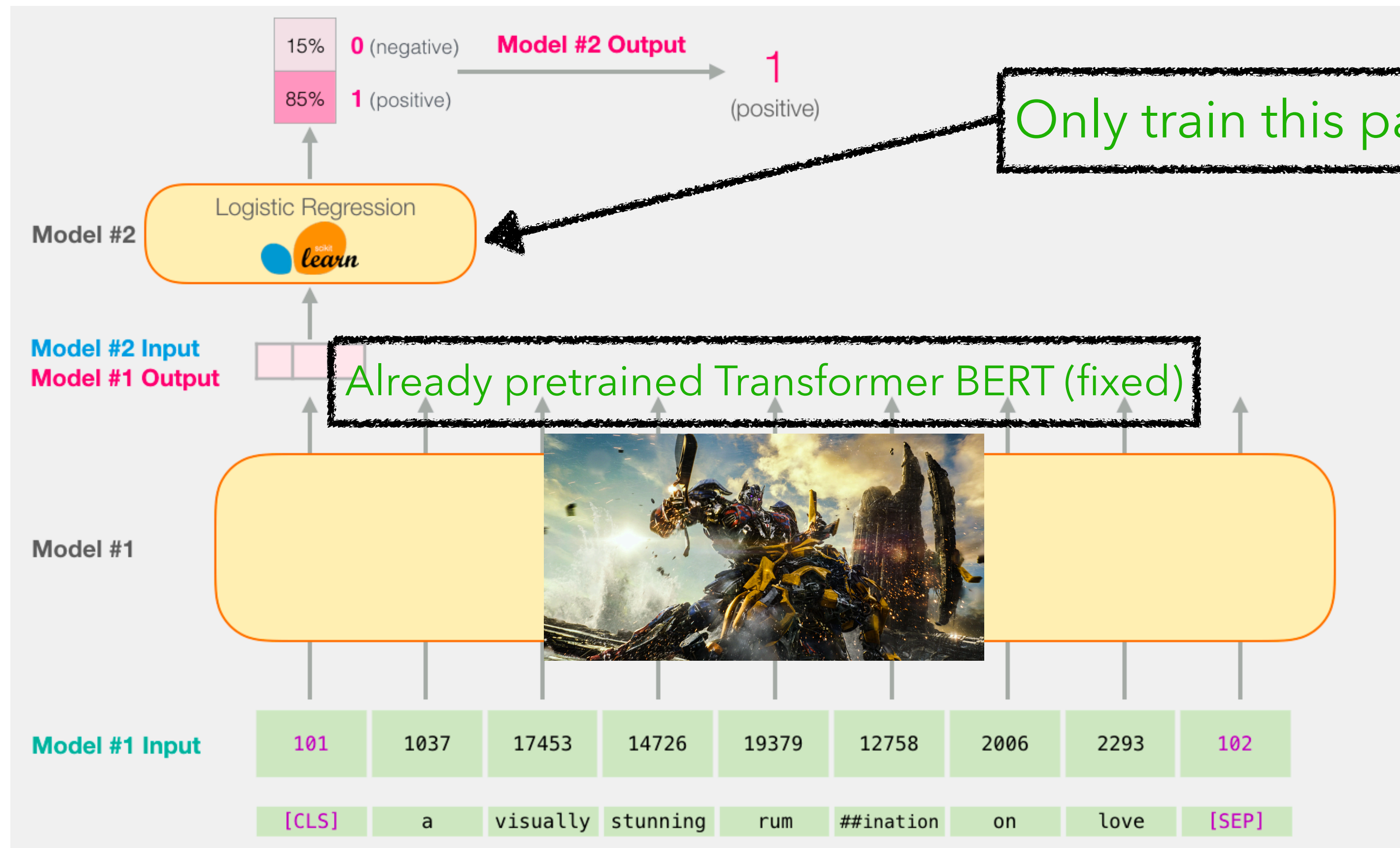
Modern learning paradigm

- **Pre-training + supervised training/fine-tuning**

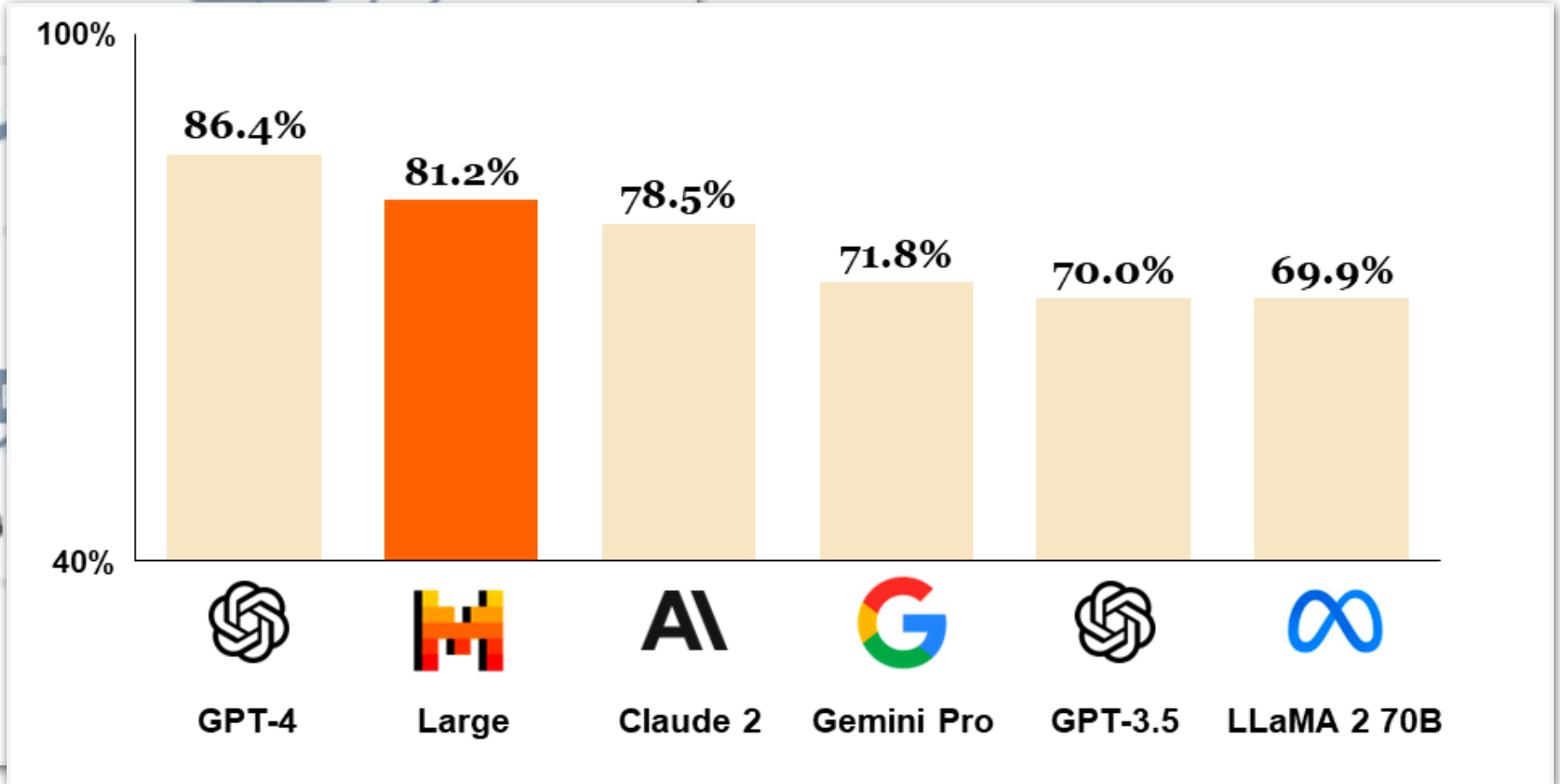
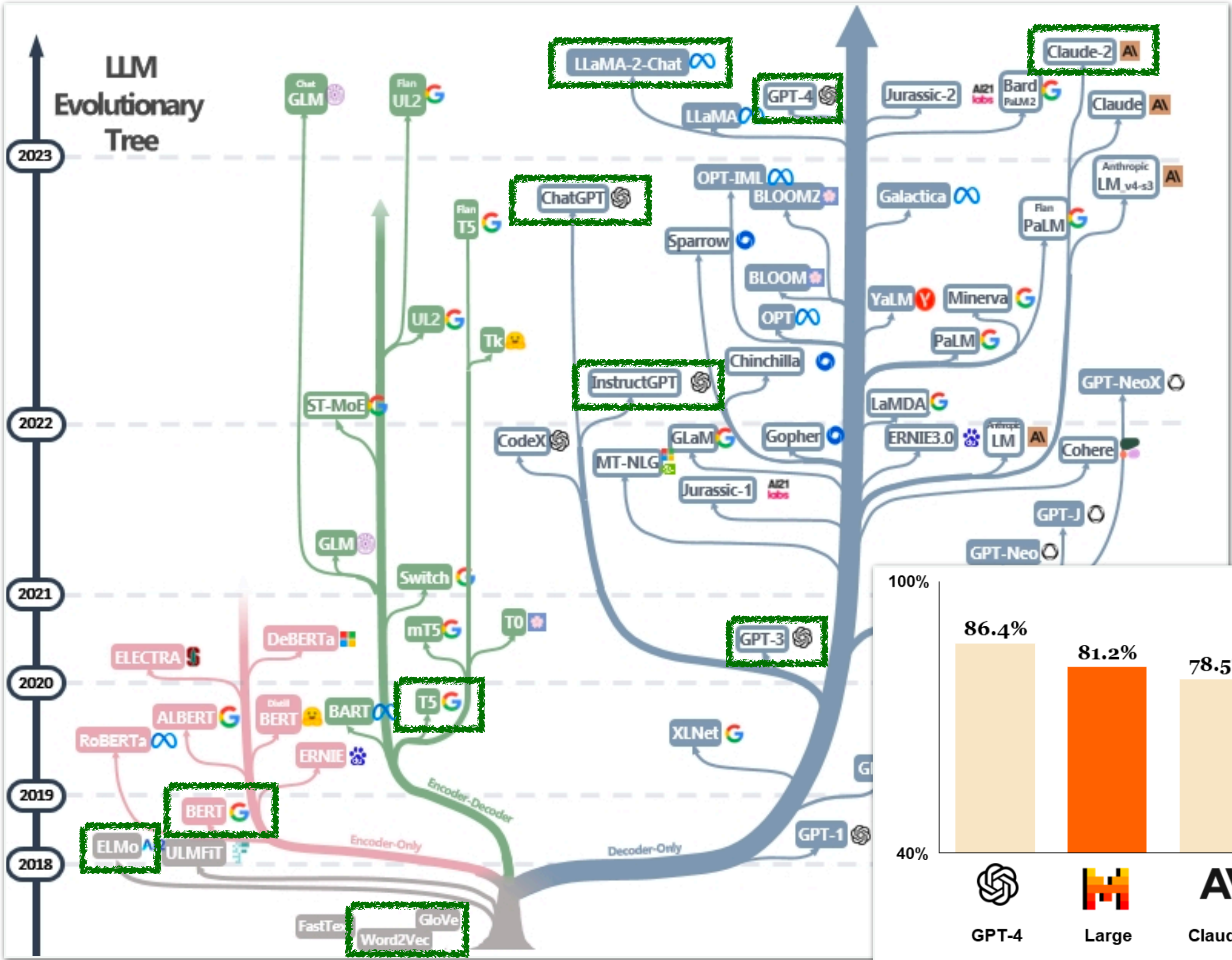
- First train Transformer using a lot of general text using unsupervised learning. This is called **pretraining**.
- Then train the pretrained Transformer for a specific task using supervised learning. This is called **finetuning**.



Example: BERT for sentiment classification



Evolution tree of pretrained LMs



Latest learning paradigm with LLMs

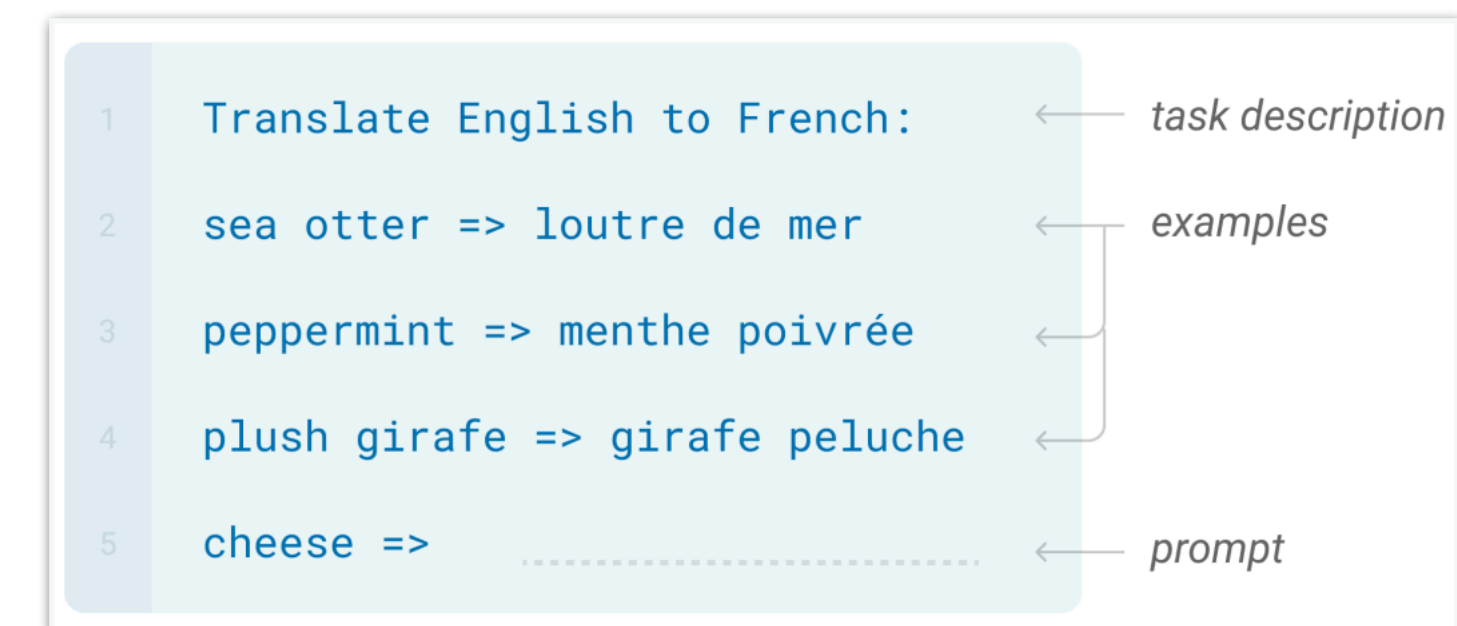
- **Pre-training + prompting/in-context learning (no training this step)**
 - First train a **large (>7~175B)** Transformer using a lot of general text using unsupervised learning. This is called **large** language model **pretraining**.

Latest learning paradigm with LLMs

- **Pre-training + prompting/in-context learning (no training this step)**
 - First train a **large (>7~175B)** Transformer using a lot of general text using unsupervised learning. This is called **large** language model **pretraining**.
 - Then **directly use** the pretrained large Transformer (**no further finetuning/training**) for any different task given only a natural language description of the task or a few task (x, y) examples. This is called **prompting/in-context learning**.



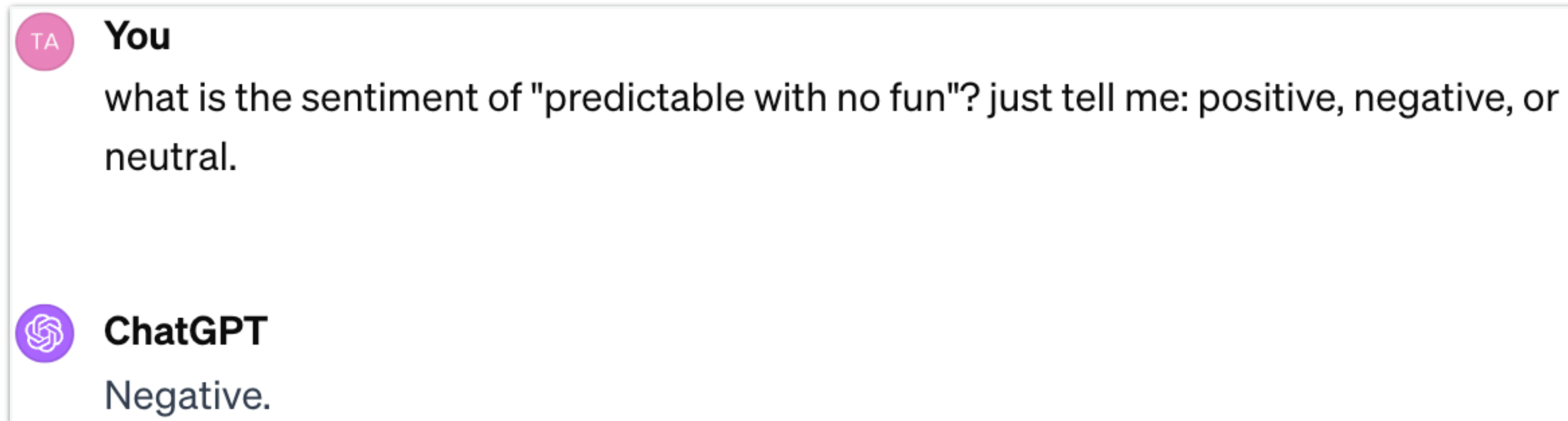
Zero-shot prompting



Few-shot prompting/in-context learning

Example: Prompting ChatGPT for sentiment analysis

- **Pre-training + prompting/in-context learning (no training this step)**

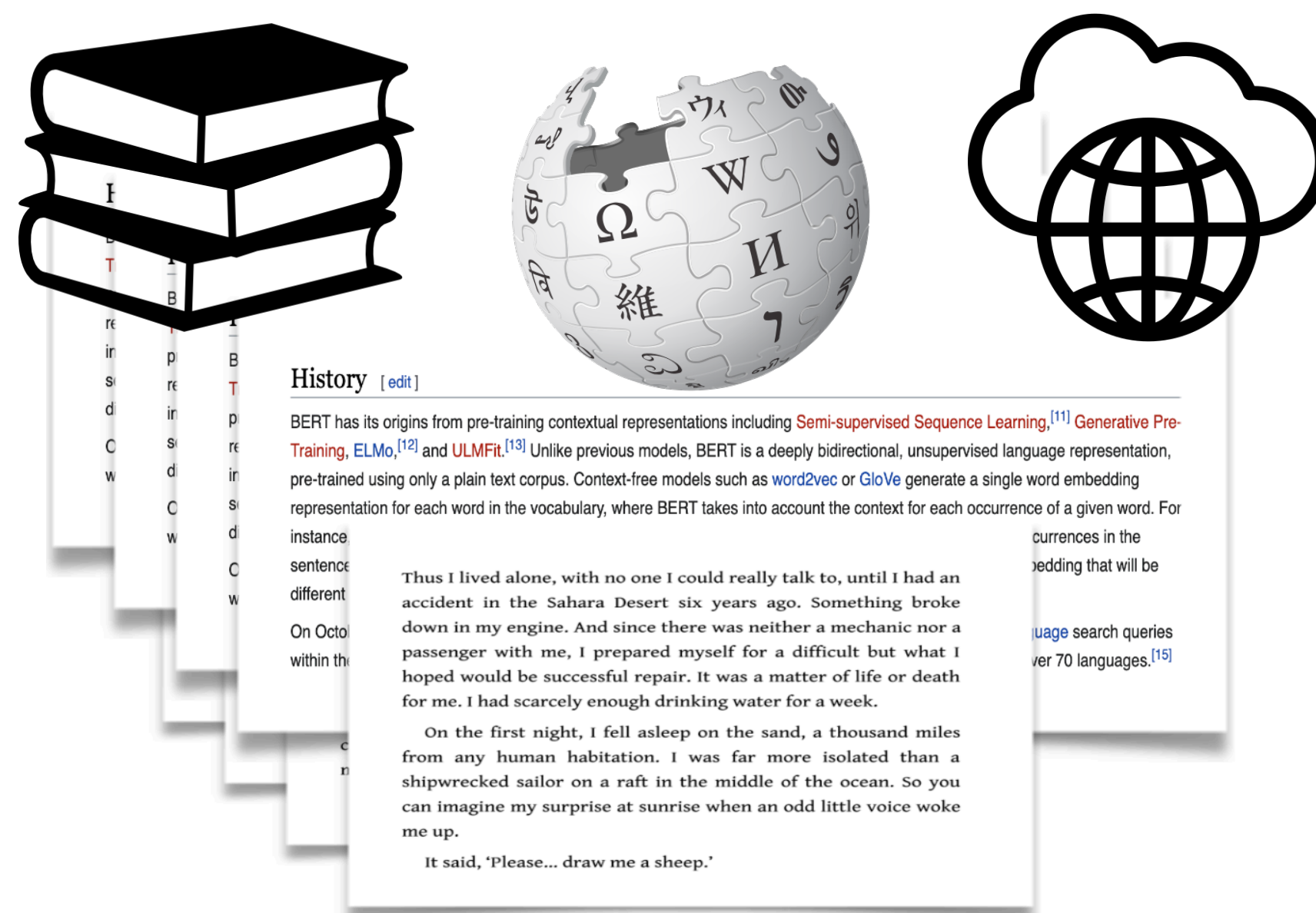


Already pretrained ChatGPT
No further training for sentiment analysis
Just prompting to conduct the task!



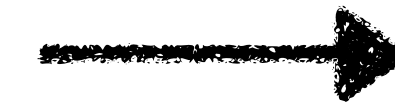
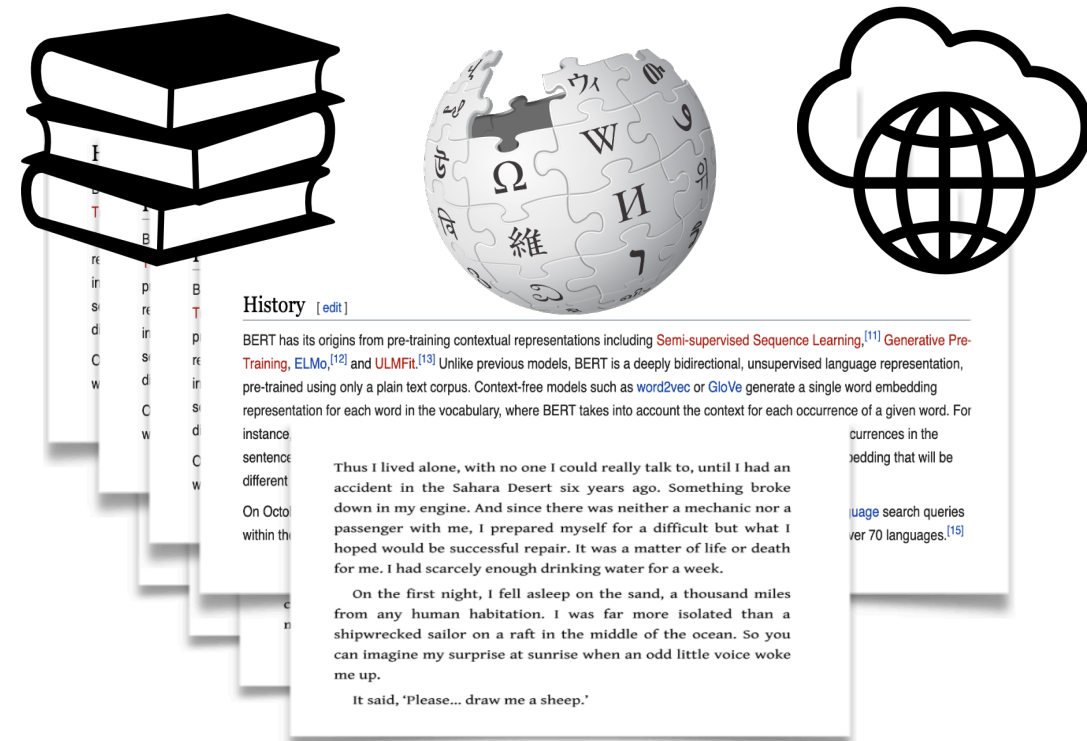
Pretraining: training objectives?

- During pretraining, we have a large text corpus (**no task labels**)
 - **Key question: what labels or objectives used to train the vanilla Transformers?**



Pretraining: training objectives?

- During pretraining, we have a large text corpus (**no task labels**)
 - **Key question: what labels or objectives used to train the vanilla Transformers?**



**Training
labels/objectives?**

Pretraining: training objectives?



BERT

Devlin et al., 2018

The cabs ___ the same rates as those ___ by horse-drawn cabs and were ___ quite popular, ___ the Prince of Wales (the ___ King Edward VII) travelled in ___. The cabs quickly ___ known as "hummingbirds" for ___ noise made by their motors and their distinctive black and ___ livery. Passengers ___ the interior fittings were ___ when compared to ___ cabs but there ___ some complaints ___ the ___ lighting made them too ___ to those outside ___.

charged, used, initially, even, future, became, the, yellow, reported, that, luxurious, horse-drawn, were that, internal, conspicuous, cab

Masked token prediction



T5

Raffel et al., 2019

Original text

Thank you ~~for inviting~~ me to your party ~~last~~ week.

Inputs

Thank you <X> me to your party <Y> week.

Targets

<X> for inviting <Y> last <Z>

Denosing span-mask prediction



GPT - 4

Text: Second Law of Robotics: A robot must obey the orders given it by human beings

Generated training examples

Example #	Input (features)	Correct output (labels)
1	Second law of robotics :	a
2	Second law of robotics : a	robot
3	Second law of robotics : a robot	must
...		

Next token prediction

BERT: Bidirectional Encoder Representations from Transformers

(Released in 2018/10)

- It is a fine-tuning approach based on a deep **bidirectional Transformer encoder** instead of a Transformer decoder
- The key: learn representations based on **bidirectional contexts**

Example #1: we went to the river bank.

Example #2: I need to go to bank to make a deposit.

- Two new pre-training objectives:
 - Masked language modeling (MLM)
 - Next sentence prediction (NSP) - Later work shows that NSP hurts performance though..

