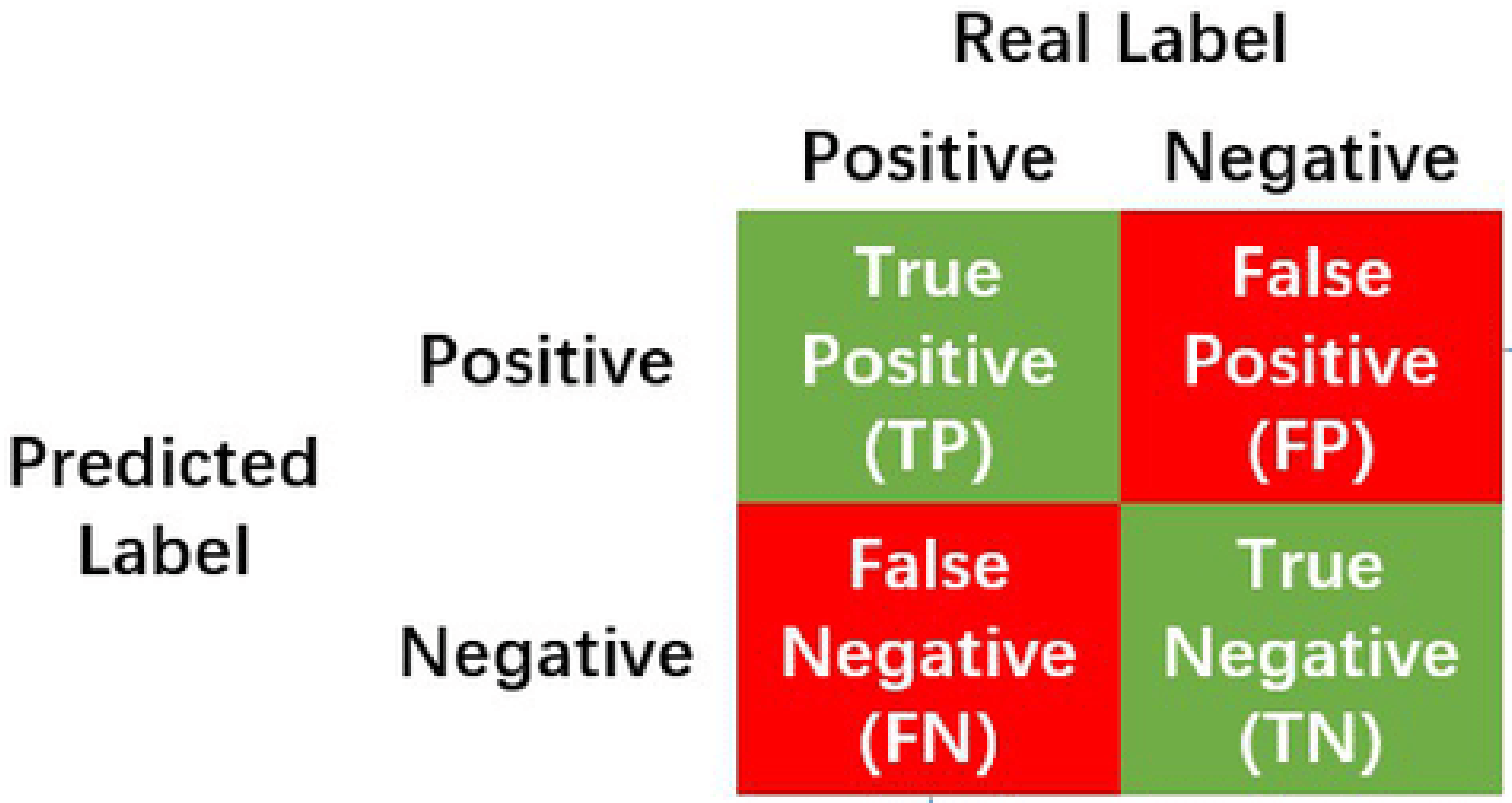


$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Precision/Recall

- When computing one of the two measures, it is important (if possible) to report the other
- The reason is that it is easy to increase one at the expense of the other
  - Predicting more positive would increase recall at the expense of precision
  - Predicting more negative would increase precision at the expense of recall
- Often their harmonic mean is used to report one is known as F-score:

$$\text{F-score} = \frac{2 \cdot \ \text{Precision} \cdot \ \text{Recall}}{\text{Precision} + \text{Recall}}$$

Pretraining and Contextualized Word Embeddings

Contextualized Word Embeddings

- Pre-neural NLP: manually crafted features, most supervised learning
- Labeled data is scarce, unlabeled data is abundant
  - And there is much we can learn from unlabeled data
  - Yet, supervised learning requires labeled data…
- Word embeddings liked *word2vec* partially address this:
  - Using unlabeled data we can get embeddings, which in turn can serve as features for supervised learning
- But bag-of-words embeddings give up on a lot of useful information:
  - Word order
  - Different senses/uses of a word are conflated

Contextualized Word Embeddings

- Neural language models implicitly represent much information about words
  - To predict the next word, we need to represent not only the neighbors of a word, but also their senses, their order etc.
  - They therefore implicitly induce embeddings from unlabeled data that contain richer information than, say, *word2vec*
- The same can potentially hold for sentences:
  - A sentence can be represented with a representation that can help predict the preceding and following sentence

What does an LM need to know implicitly?

- Much like with Machine Translation, Language Modeling requires very diverse capabilities
- Grammar:

  He grew up to be taller ______ **(than)**

  I ate breakfast and ______ **(then)**

- Disambiguation:
  - The odd one out of the words crane, pelican, excavator, hoist and upraise is ________ **(pelican)**

https://aclanthology.org/D18-1182.pdf

What does an LM need to know implicitly?

- General factual knowledge:

  The capital of France is ______ **(Paris)**

- Paraphrasing:

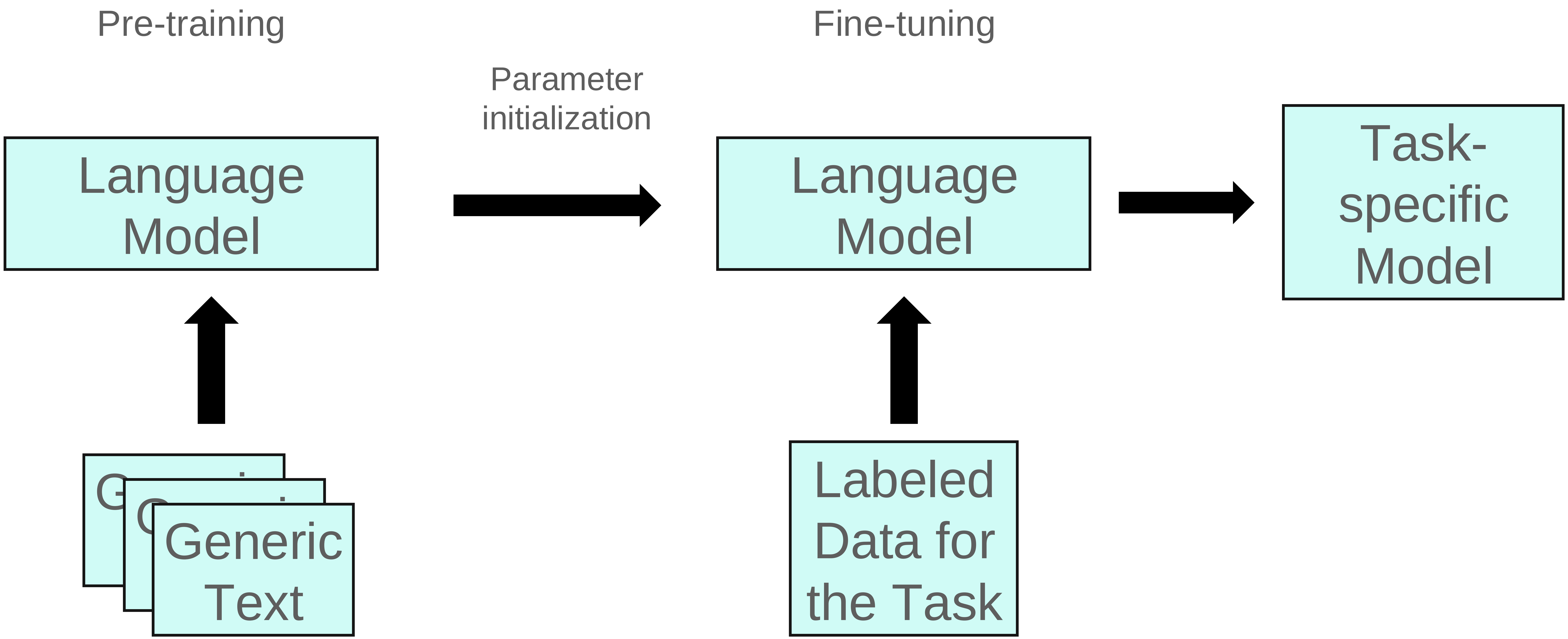
  Instead of saying that you could not disagree more, you can say ___________ **(you strongly disagree)**

Among many others…

Pre-training

- But for language models all we need is <u>unlabeled</u> text
- We can therefore train language models to perform next word prediction and use the obtained model as initialization for a downstream task
- This stage is known as *pre-training* and the supervised part is called *fine-tuning*
- Pre-training tends to be very demanding computationally, but many fine-tuned models can be used with it

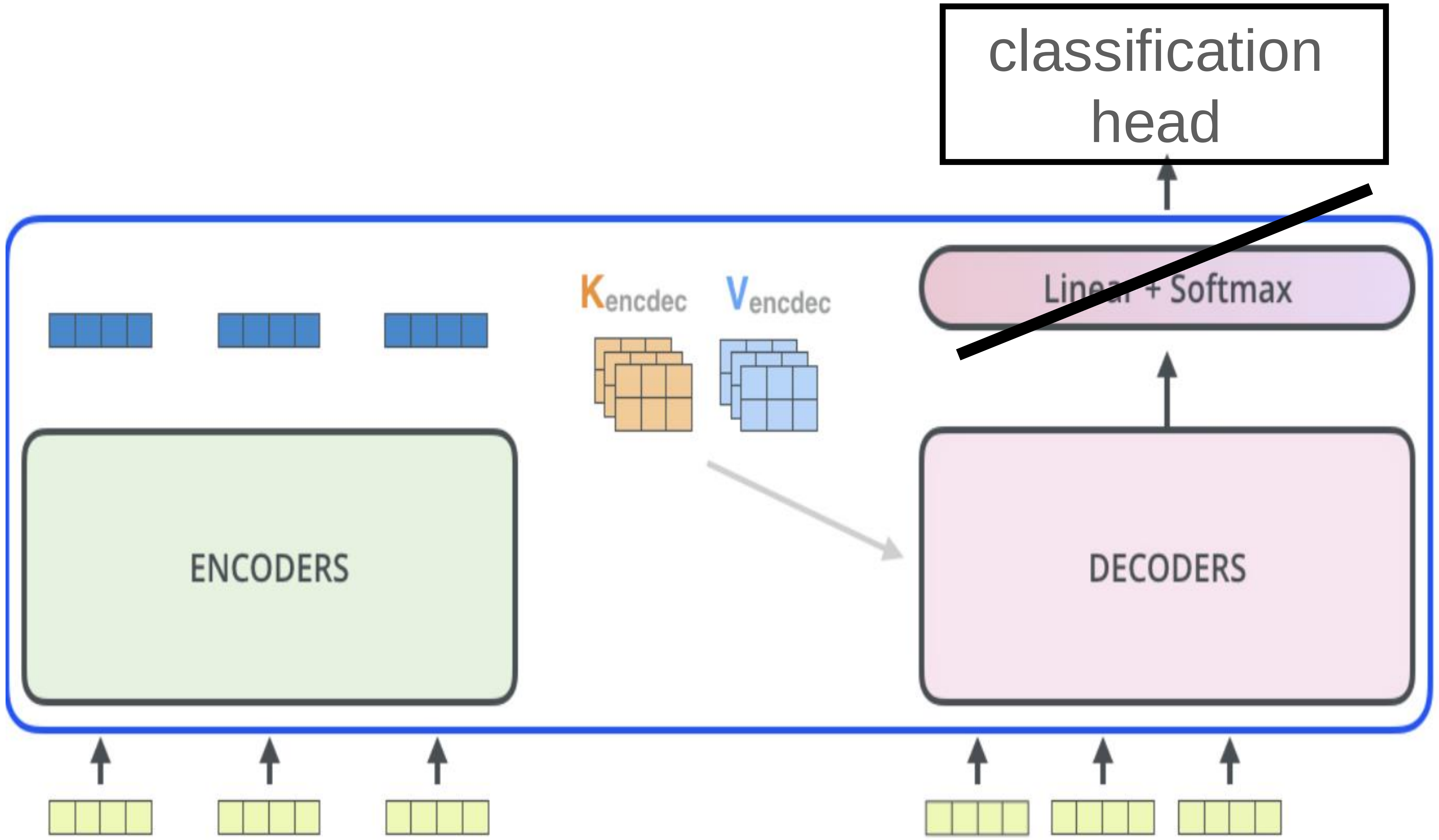
Pre-training and Fine-tuning



Pre-training and Fine-tuning

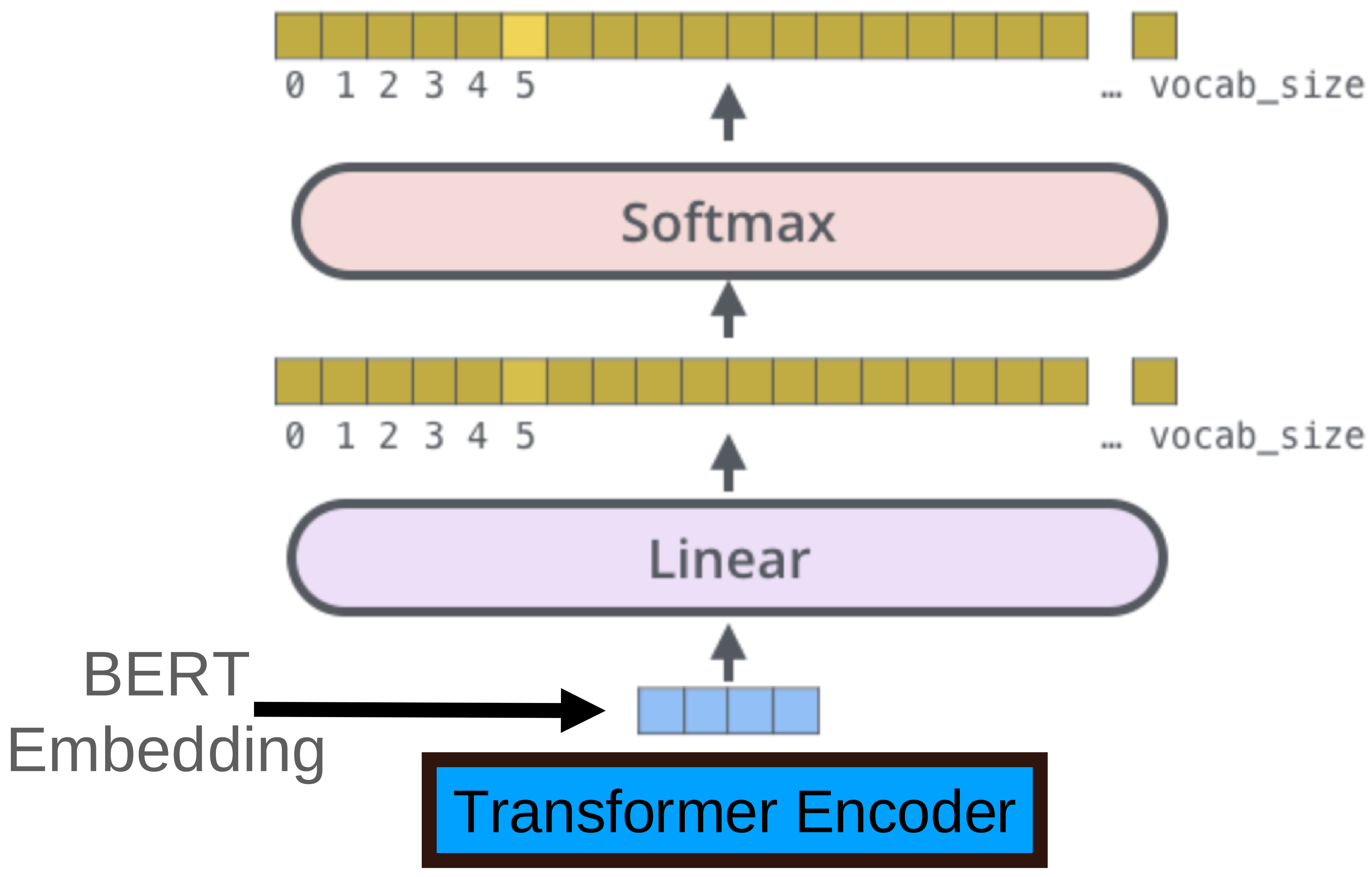
- For the fine-tuning task, we often replace the linear+soft-max layer we used in the LM for decoding, with a task-specific sub-network

- Most common: a classification head – usually a feed forward network and a soft-max over the labels at the end



The BERT Model

- The first widely popular model to use pre-training
- The BERT model is a standard model for obtaining contextualized word embeddings
- The model is built like a *Transformer* **encoder**, except that instead of generating a list of vector embeddings, it feeds these vectors to a linear layer, followed by a soft-max layer, followed by a soft-max
- BERT provides embeddings for each token in the input sequence – the computed vector that enters the linear layer



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
https://arxiv.org/abs/1810.04805

Masked Language Models

- BERT is trained as a **masked language model** (MLM)
- MLMs are statistical models that, given a sentence where part of the tokens are replaced with masks, predict the identity of the masked tokens

My dog is [MASK] and likes to [MASK] the entire [MASK].

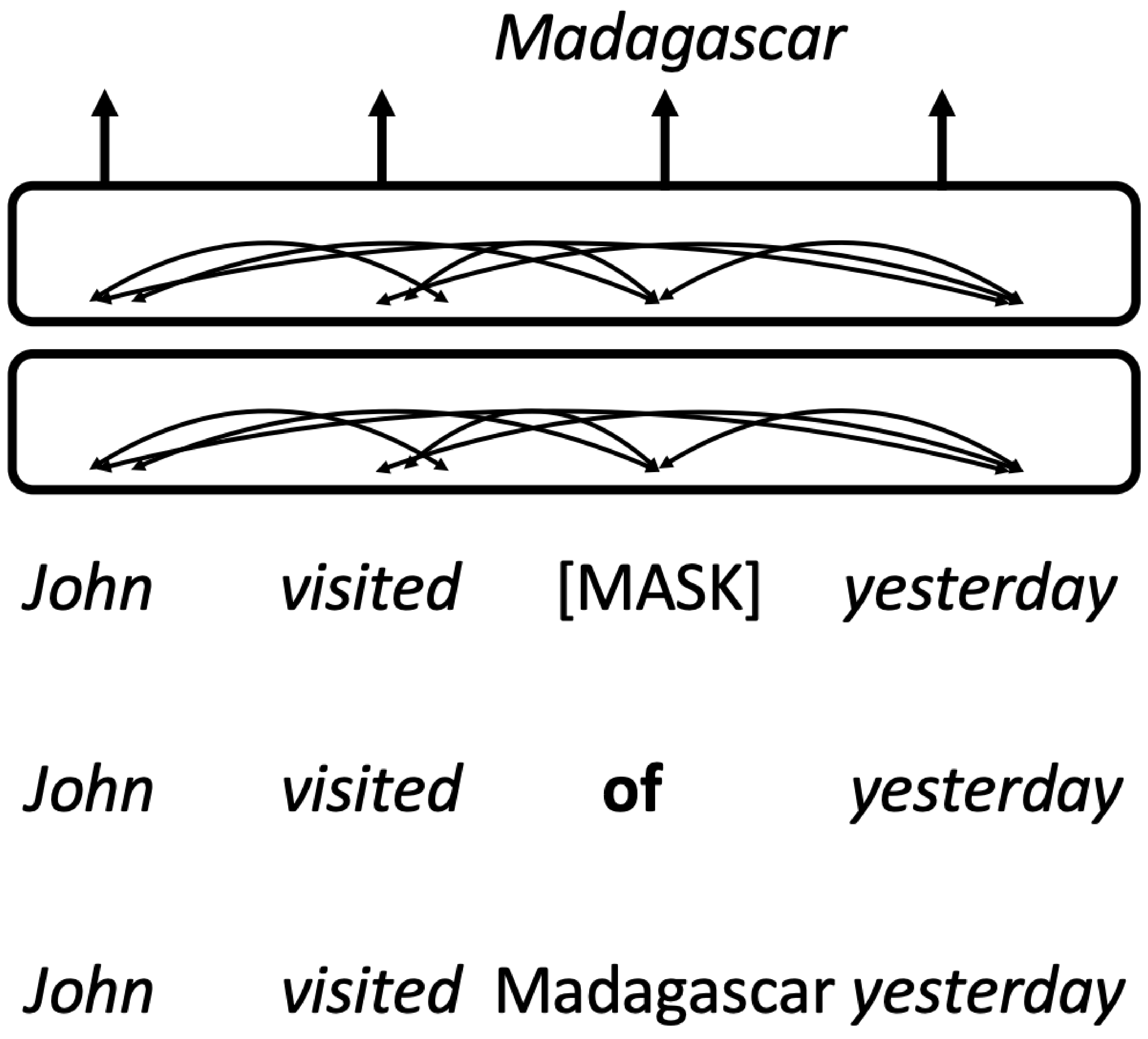
Mask 1

Prediction	Score
My dog is **hungry** and likes to *[MASK2]* the entire *[MASK3]* .	5.6%
My dog is **friendly** and likes to *[MASK2]* the entire *[MASK3]* .	4.8%
My dog is **cute** and likes to *[MASK2]* the entire *[MASK3]* .	3.7%
My dog is **nice** and likes to *[MASK2]* the entire *[MASK3]* .	2.9%
My dog is **smart** and likes to *[MASK2]* the entire *[MASK3]* .	2.4%

https://demo.allennlp.org/masked-lm

BERT’s Training

- BERT’s training is carried out by segmenting texts into windows in the size of the model’s input window, and selecting 15% of the tokens (denote the set with *S*)
- The tokens in *S* are altered such that:
    - 80% of them are replaced with the special token <MASK>
    - 10% of them are replaced with a random token
    - 10% of them remain the same
- Denote the modified input sequence with ***x’***

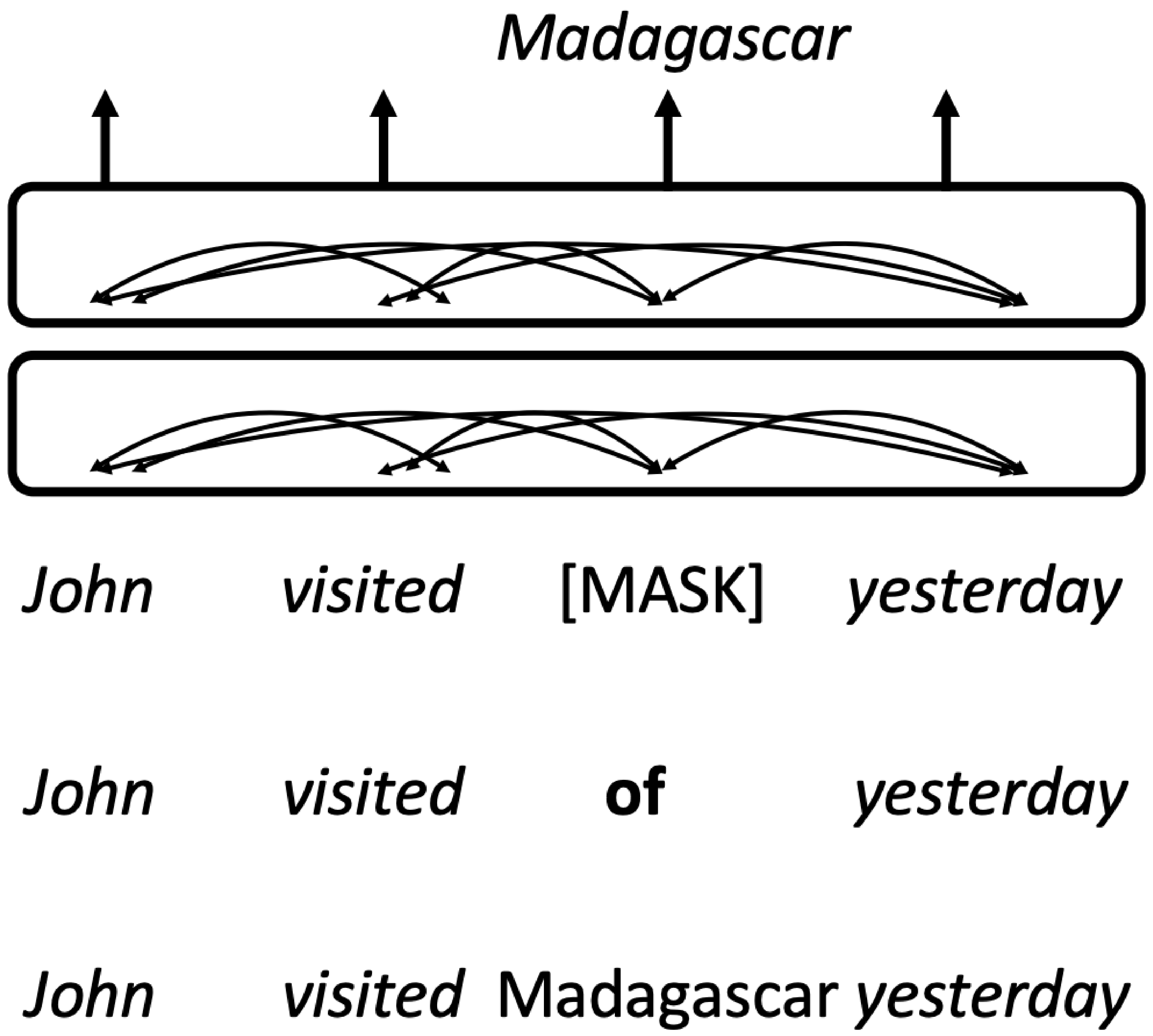


BERT's Training

- The loss function is then the cross entropy loss <u>over the tokens in S</u>:

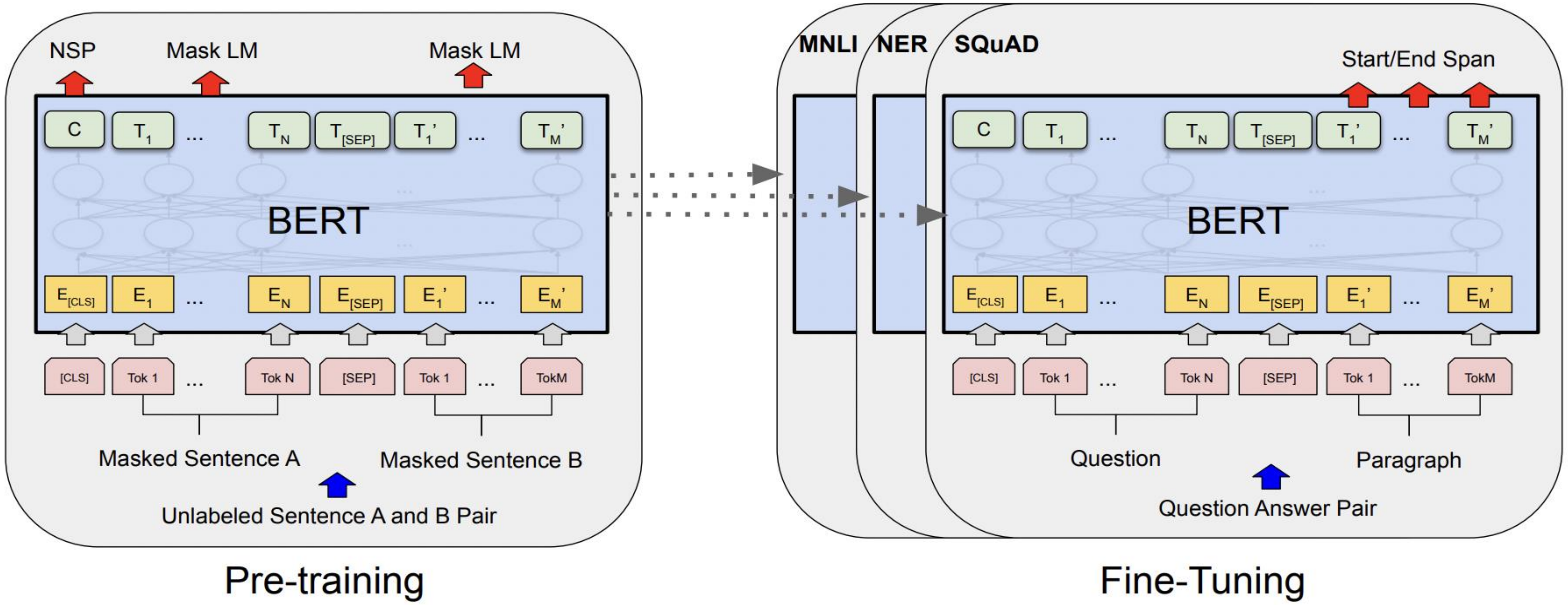
$$L(\theta) = -\frac{1}{|S|}\sum_{i\in\ S} log(p_\theta(x_i|x'))$$

- During inference, no masking is employed



Fine Tuning BERT

- The standard paradigm for using BERT in NLP includes two phases:
  - Pre-training
  - Fine-tuning



Examples for fine-tuning BERT

- POS tagging:
  - Often viewed as an independent prediction task (not structured prediction)
  - A classification head for each word
    - Its input is the BERT embedding and its output is a soft-max in the dimension of the number of POS tags.
  - As usual, the *j*-th coordinate of the soft-max is interpreted as $P(y_j|x_1,\ldots,x_n)$
  - Training set:
    - Each sample consists of the sentence $(x_1,\ldots,x_n)$, index i, and gold POS tag for i-th token y_i^*
    - Cross entropy loss over $P(y_i^*|x_{1,\ldots,n})$

Examples for fine-tuning BERT

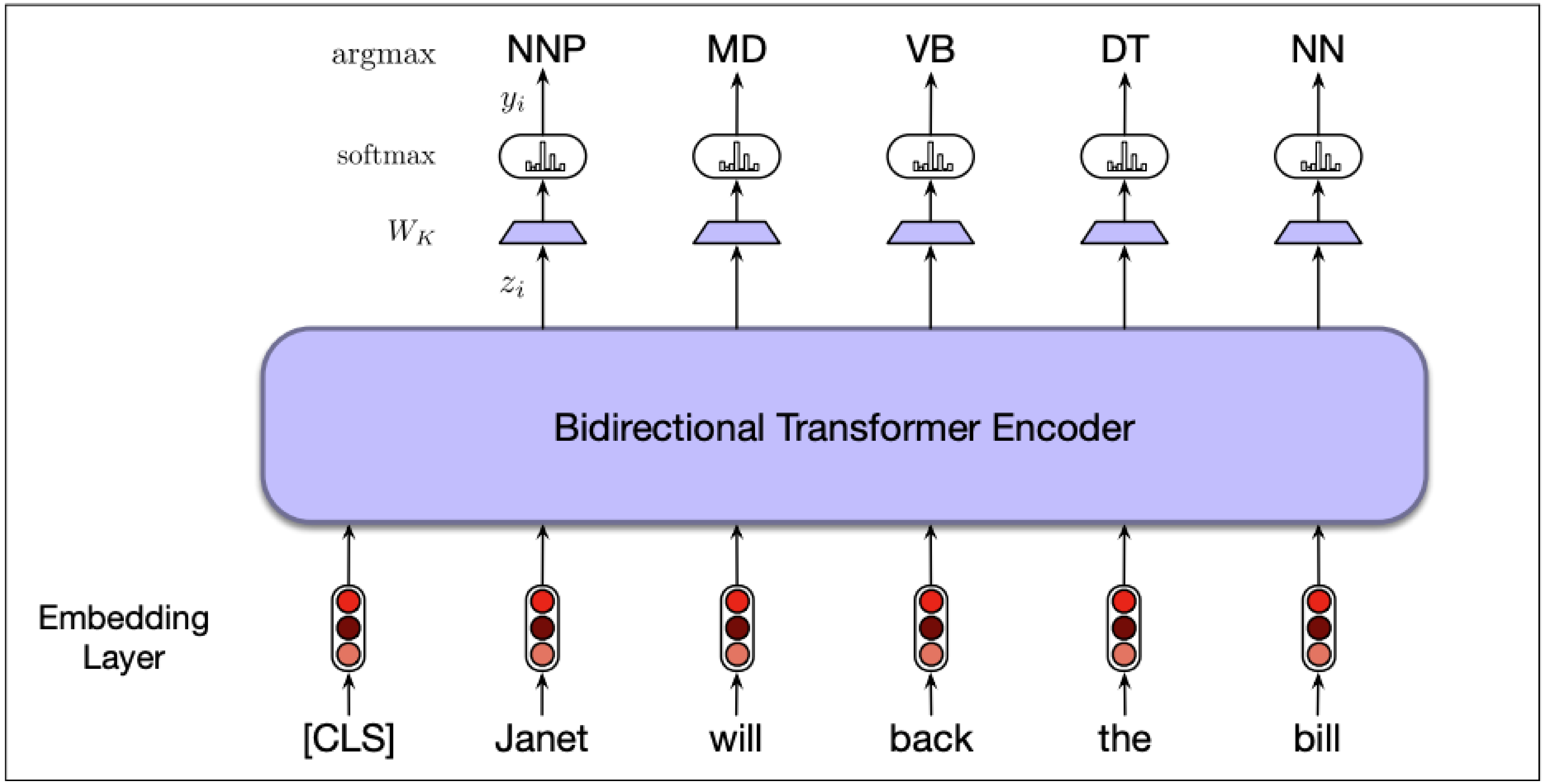


Figure 11.9 Sequence labeling for part-of-speech tagging with a bidirectional transformer encoder. The output vector for each input token is passed to a simple k-way classifier.

Next Sentence Prediction

- BERT is in fact pretrained not only as a masked language, but also as a model for Next Sentence Prediction (NSP)
- NSP is the task that requires, given two sentences, to determine whether the two sentences form a sensible consecutive pair or is just two random sentences
  - NSP training data: positive - pairs of adjacent sentences, negative - pairs of random sentences

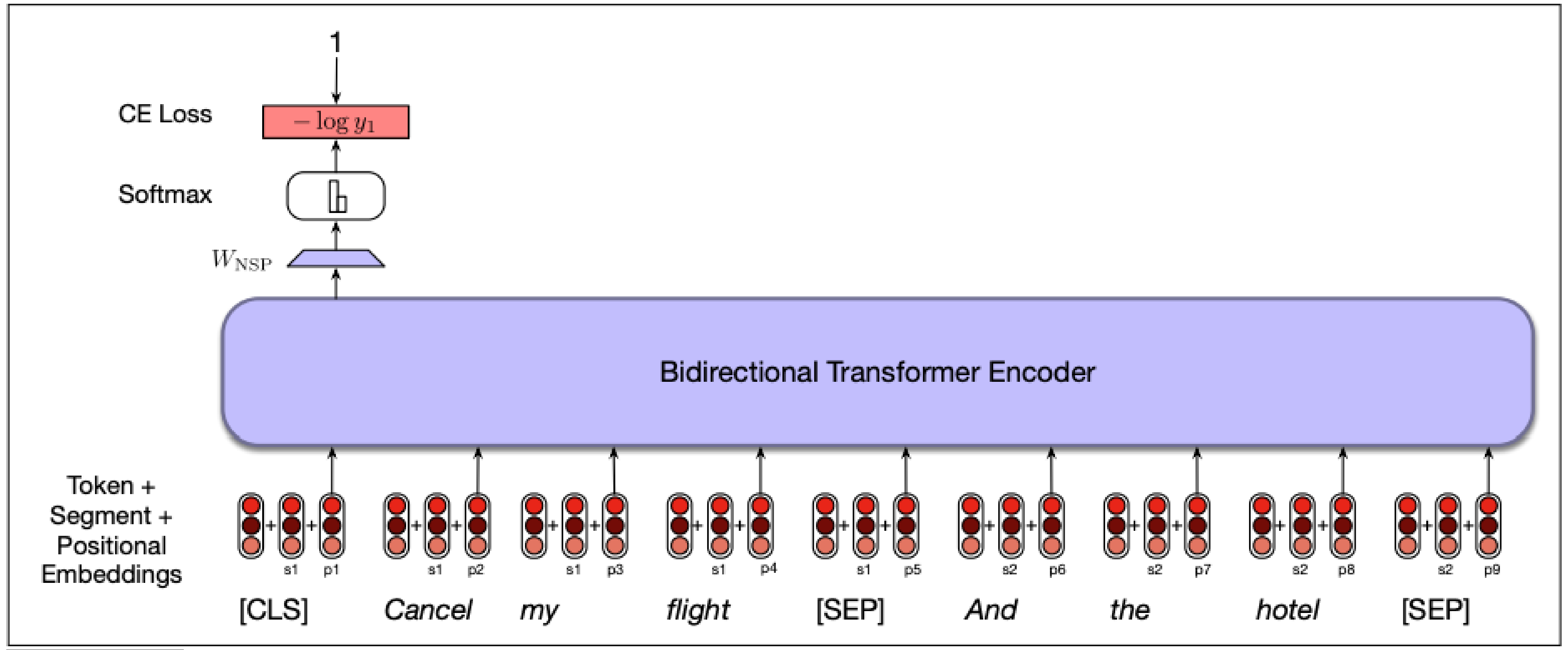


Figure 11.7 An example of the NSP loss calculation.

Next Sentence Prediction

- Each training sample is prepended with a special [CLS] token. A special [SEP] token is added between the sentences
- The loss function is over the [CLS] where the labels are 1/0
- [CLS] is often used for fine-tuning on sentence classification tasks

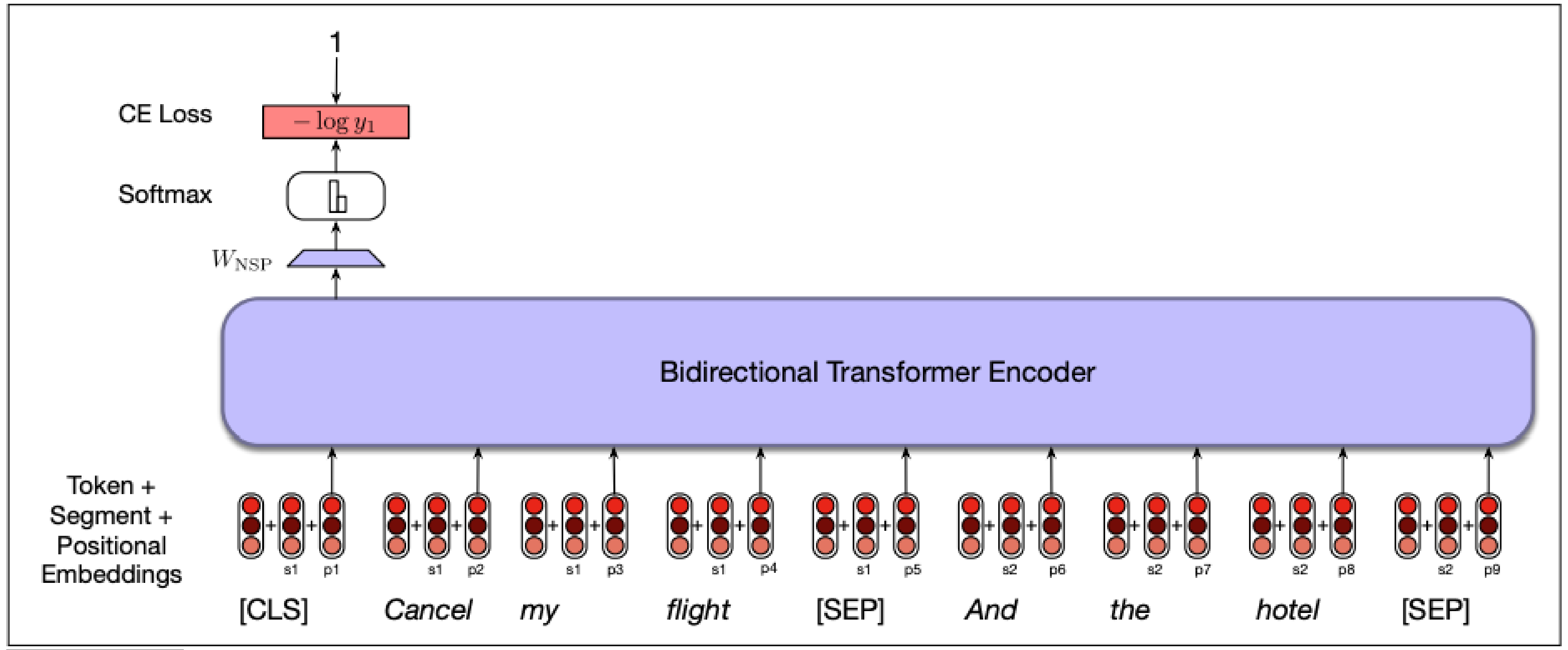


Figure 11.7 An example of the NSP loss calculation.

Examples for fine-tuning BERT

- Sentiment analysis:
  - Viewed as a sentence classification task
  - A classification head is placed over a special token ([CLS])
    - As before, usually a linear layer followed by a soft-max
  - The soft-max is interpreted as $P(class|x_1,\ldots,x_n)$
    - *class* takes values in the set of potential sentiment classes
  - Training set:
    - Each sample consists of the sentence $(x_1,\ldots,x_n)$, gold POS tag for the sentence y^*
    - The loss function is the expectation over $\log P(y^* |x_1,\ldots,x_n)$

Results of Fine-tuning BERT

- BERT achieved very impressive results across a range of token and sentence classification tasks, sequence prediction, as well as other baselines
- A very broad experimental setup in terms of task types and baselines
- It was widely adopted, and is still often used (about 5 years after its inception)
- For detailed results, see the original BERT paper

Zero-shot Prediction: Alternative to using MLM for embeddings

- (Masked) language models can also be used to directly decode the answer as a token (without fine-tuning). This is called zero-shot prediction.
- For example, this can be used to answer the question
  "what movies did Brad Pitt appear in?"

On November 22, 2001, Pitt made a guest appearance in the eighth season of the television series *Friends*, playing a man with a grudge against Rachel Green, played by Jennifer Aniston, to whom Pitt was married at the time.[86] For this performance he was nominated for an Emmy Award in the category of Outstanding Guest Actor in a Comedy Series.[87] In December 2001, Pitt played Rusty Ryan in the heist film *Ocean's Eleven*, a remake of the 1960 Rat Pack original. He joined an ensemble cast including George Clooney, Matt Damon, Andy García, and Julia Roberts.[88] Well received by critics, *Ocean's Eleven* was highly successful at the box office, earning $450 million worldwide.[32] Pitt appeared in two episodes of MTV's reality series *Jackass* in February 2002, first running through the streets of Los Angeles with several cast members in gorilla suits,[89] and in a subsequent episode participating in his own staged abduction.[90] In the same year, Pitt had a cameo role in George Clooney's directorial debut *Confessions of a Dangerous Mind*.[91] He took on his first voice-acting roles in 2003, speaking as the titular character of the DreamWorks animated film *Sinbad: Legend of the Seven Seas*[92] and playing Boomhauer's brother, Patch, in an episode of the animated television series *King of the Hill*.[93]

+

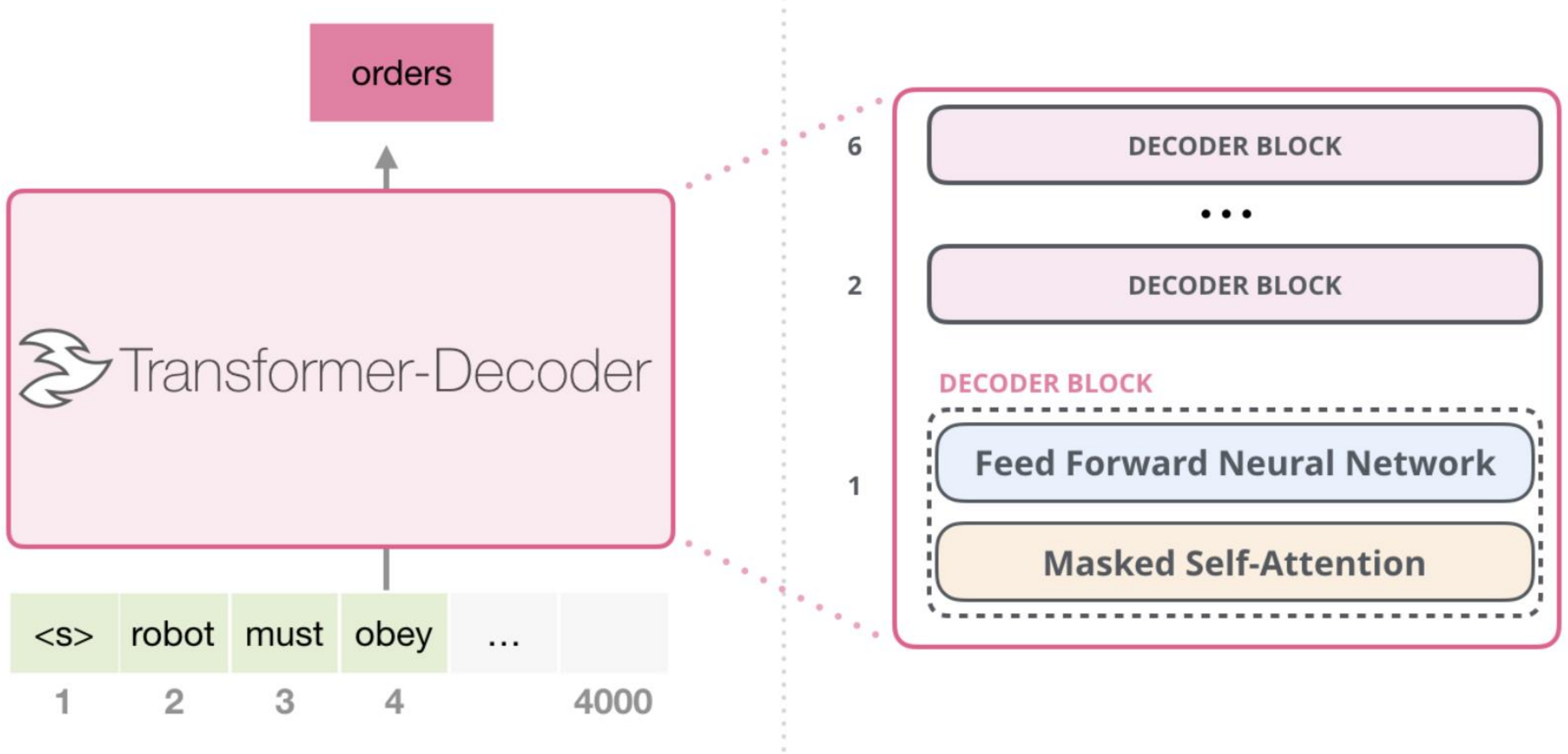
Pitt appeared in the movie [MASK]

0.3	Ocean's Eleven
0.15	Confessions of a Dangerous Mind
0.01	Friends
0.01	Jackass
…	

decoder output

Decoder-only Models

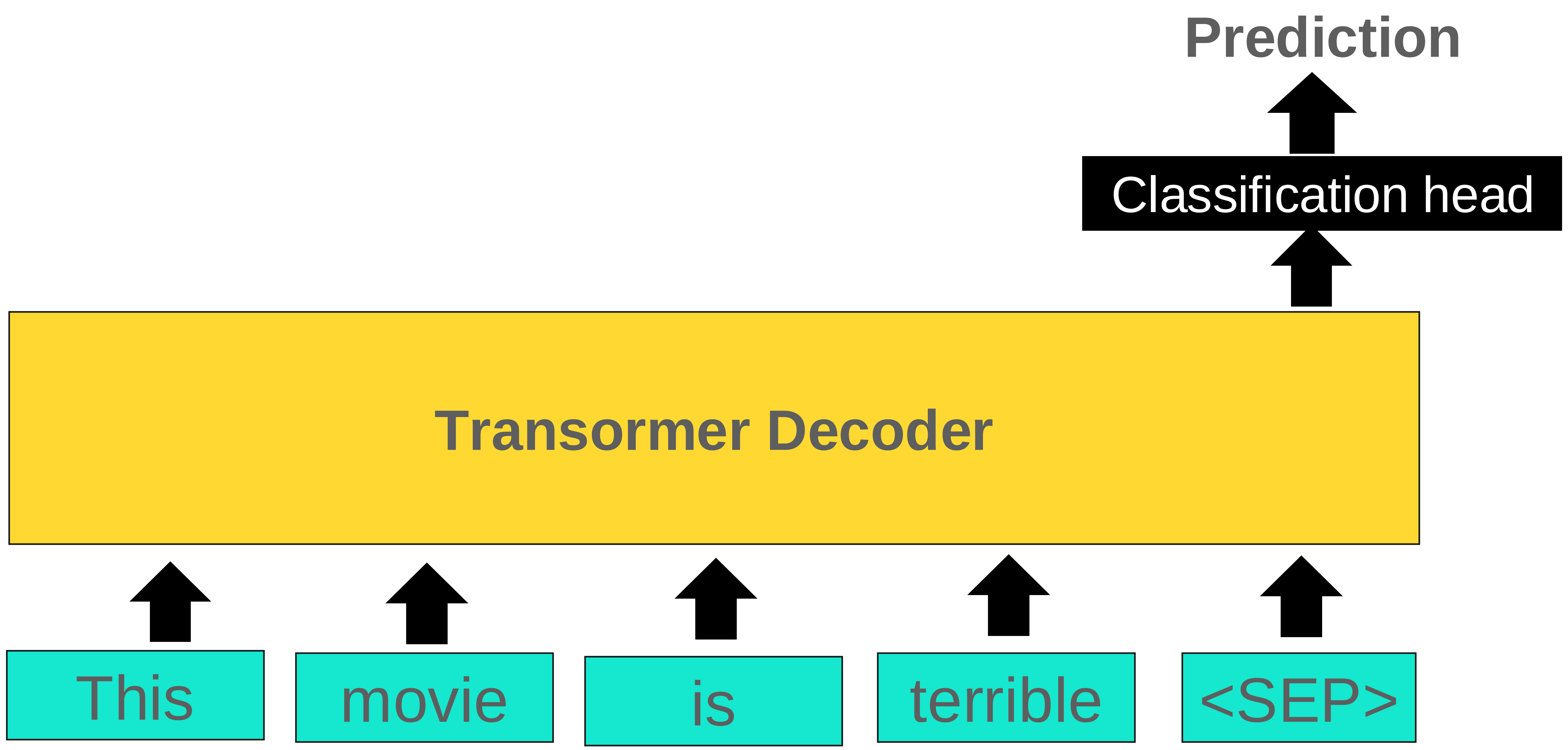
- Another prominent alternative to Transformer-based language models are decoder-only models
  - Sometimes referred to as GPT (generative pretraining) models
- These models are left-to-right language models, consisting only of a Transformer decoder



http://jalammar.github.io/illustrated-gpt2/

Fine-tuning Decoder-only Models

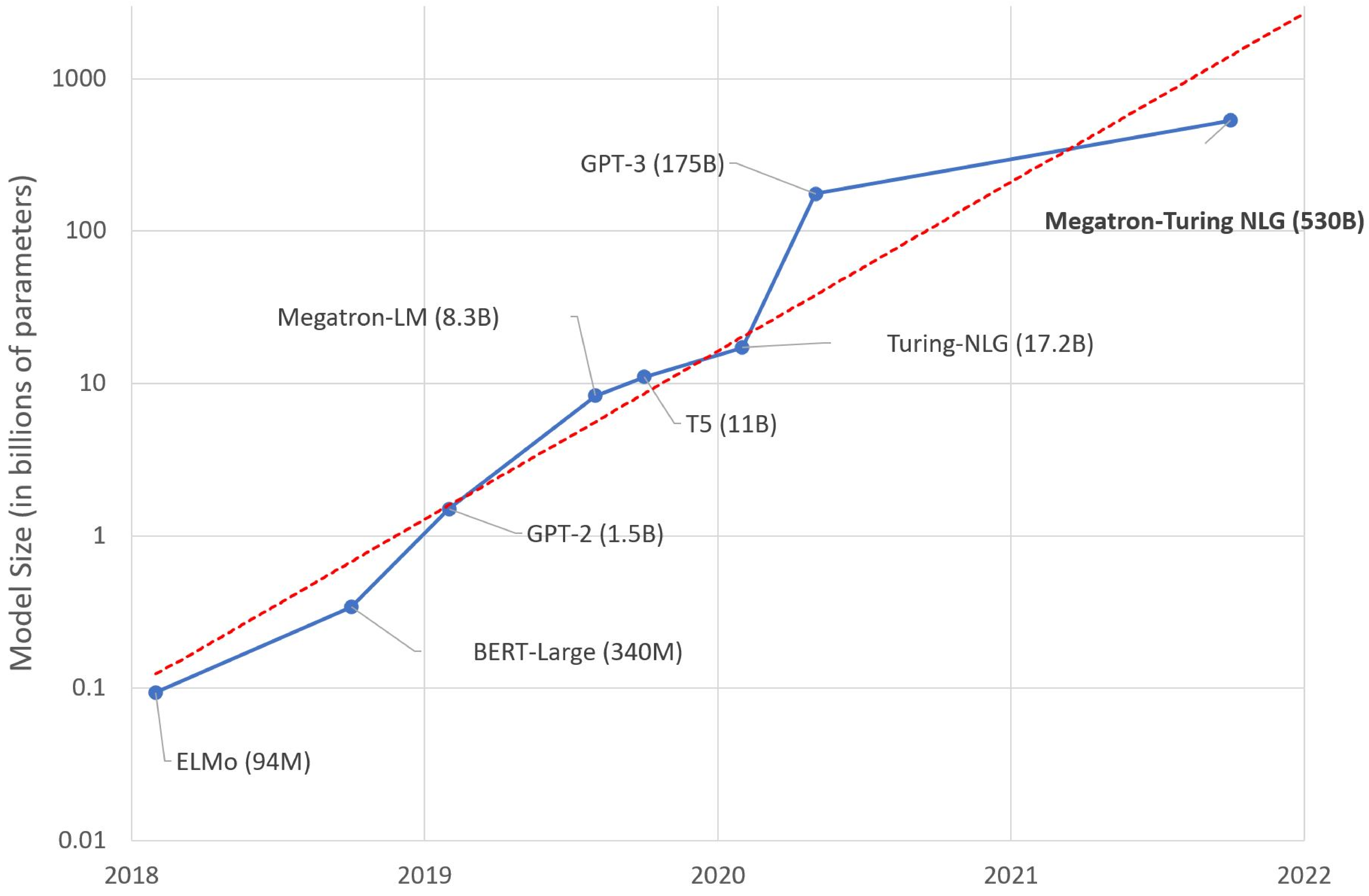
- Decoder-only models can also be fine-tuned
  - Implementation details vary but the crux is the same – using the parameters of the pretrained model as the initialization for a prediction model built on top of it



Left-to-right vs. Masked Language Models

- There has been much debate over recent years as to the relative merits of (ltr) LMs and MLMs
  - LMs define a probability distribution over strings, while MLMs do not
  - MLMs take both sides of the context into account in their representations (LMs only have representations of prefixes)
  - LMs can be easily fine-tuned not only for prediction, but also for text generation (this property turns out to be crucial)

Very Large Language Models

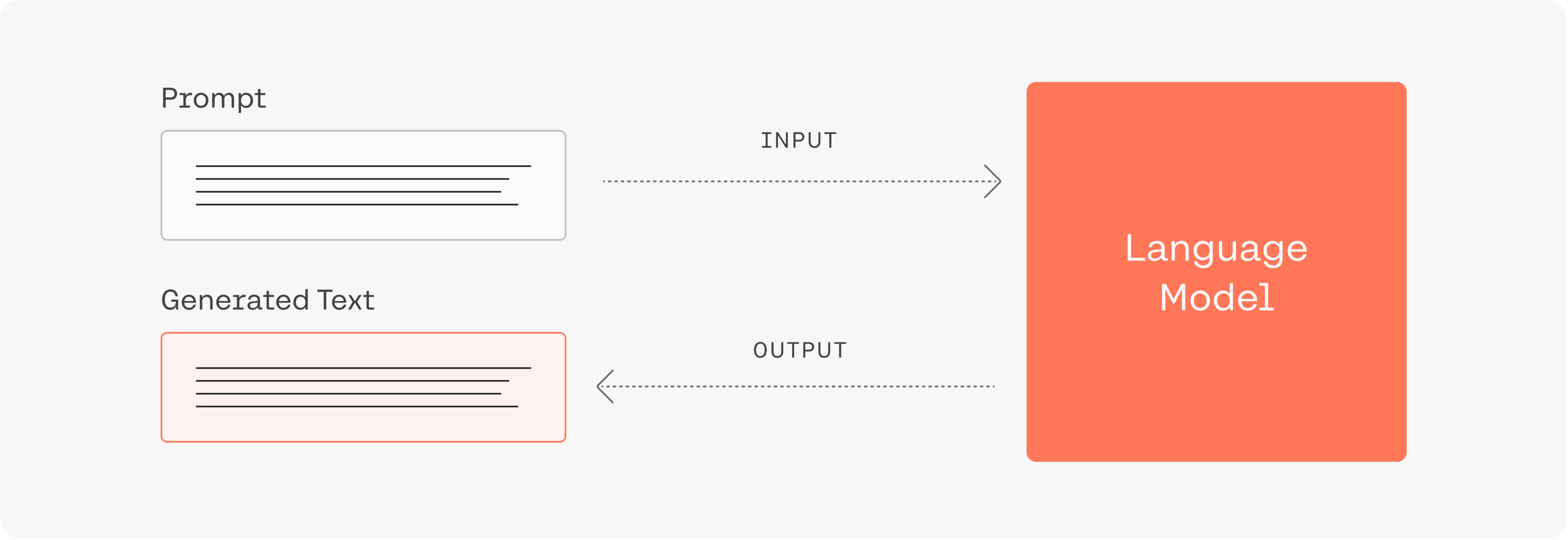


and they are continuing to get bigger since...

https://huggingface.co/blog/large-language-models

Prompting

- Large language models (to give a ballpark: > 1B parameters) have become good enough at some point, that instead of training them, people began to query them directly
  - “Prompt” is the technical name of the prefix of the text given

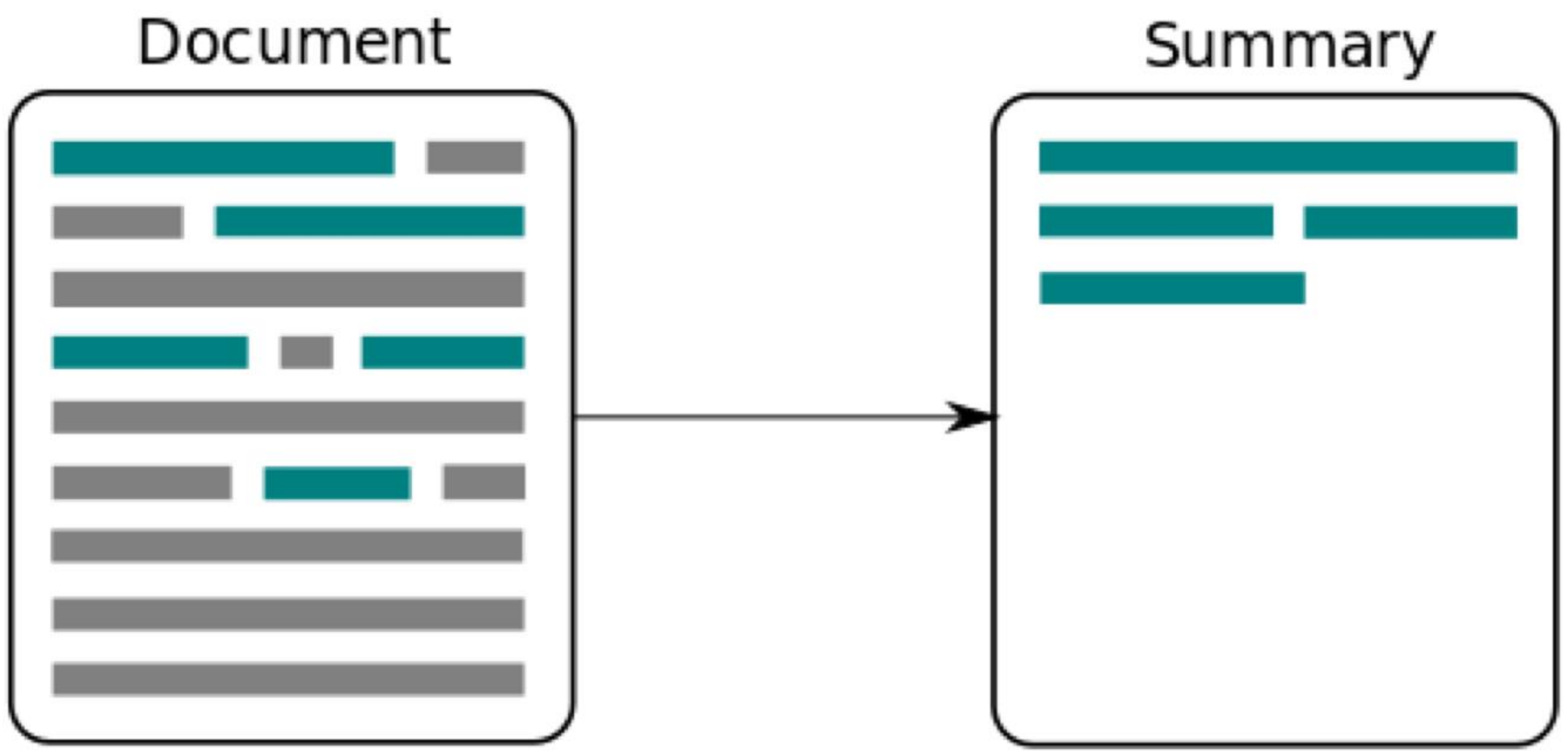


Prompting

Q: how big is France?
A:

→

543,940 km²



Few-shot Learning

```
A "whatpu" is a small, furry animal native to Tanzania.  An example of a sentence that uses
the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.
```

```
To do a "farduddle" means to jump up and down really fast.  An example of a sentence that uses
the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she
started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin.  An example of a sentence
that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden
there.  It was delicious.
```

```
A "Burringo" is a car with very fast acceleration.  An example of a sentence that uses the
word Burringo is:
In our garage we have a Burringo that my father drives to work every day.
```

```
A "Gigamuru" is a type of Japanese musical instrument.  An example of a sentence that uses the
word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift.  I love to play it at home.
```

```
To "screeg" something is to swing a sword at it.  An example of a sentence that uses the word
screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.
```

Few-shot learning: using a new word in a sentence

Language Models are Few-Shot Learners
Brown et al., 2020
https://arxiv.org/pdf/2005.14165.pdf

Zero-shot Learning

- Zero-shot capabilities: a domain-general language model (can do anything if asked to)
- More of an ideal than an accomplished goal

Context →	Please unscramble the letters into a word, and write that word: asinoc =
Target Completion →	casino

Figure G.19: Formatted dataset example for Cycled Letters

Language Models are Few-Shot Learners
Brown et al., 2023

Zero-shot Learning

- Zero-shot capabilities: a domain-general language model (can do anything if asked to)
- More of an ideal than an accomplished goal

Context →	Passage: Saint Jean de Brébeuf was a French Jesuit missionary who travelled to New France in 1625. There he worked primarily with the Huron for the rest of his life, except for a few years in France from 1629 to 1633. He learned their language and culture, writing extensively about each to aid other missionaries. In 1649, Brébeuf and another missionary were captured when an Iroquois raid took over a Huron village . Together with Huron captives, the missionaries were ritually tortured and killed on March 16, 1649. Brébeuf was beatified in 1925 and among eight Jesuit missionaries canonized as saints in the Roman Catholic Church in 1930. Question: How many years did Saint Jean de Brébeuf stay in New France before he went back to France for a few years? Answer:
Target Completion →	4

Figure G.20: Formatted dataset example for DROP

Language Models are Few-Shot Learners
Brown et al., 2023

Instruct Models

- Instruction models (like ChatGPT) do more than complete the next word – they are tuned to follow instructions

Prompt *Explain the moon landing to a 6 year old in a few sentences.*

Completion GPT-3

```
Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.
```

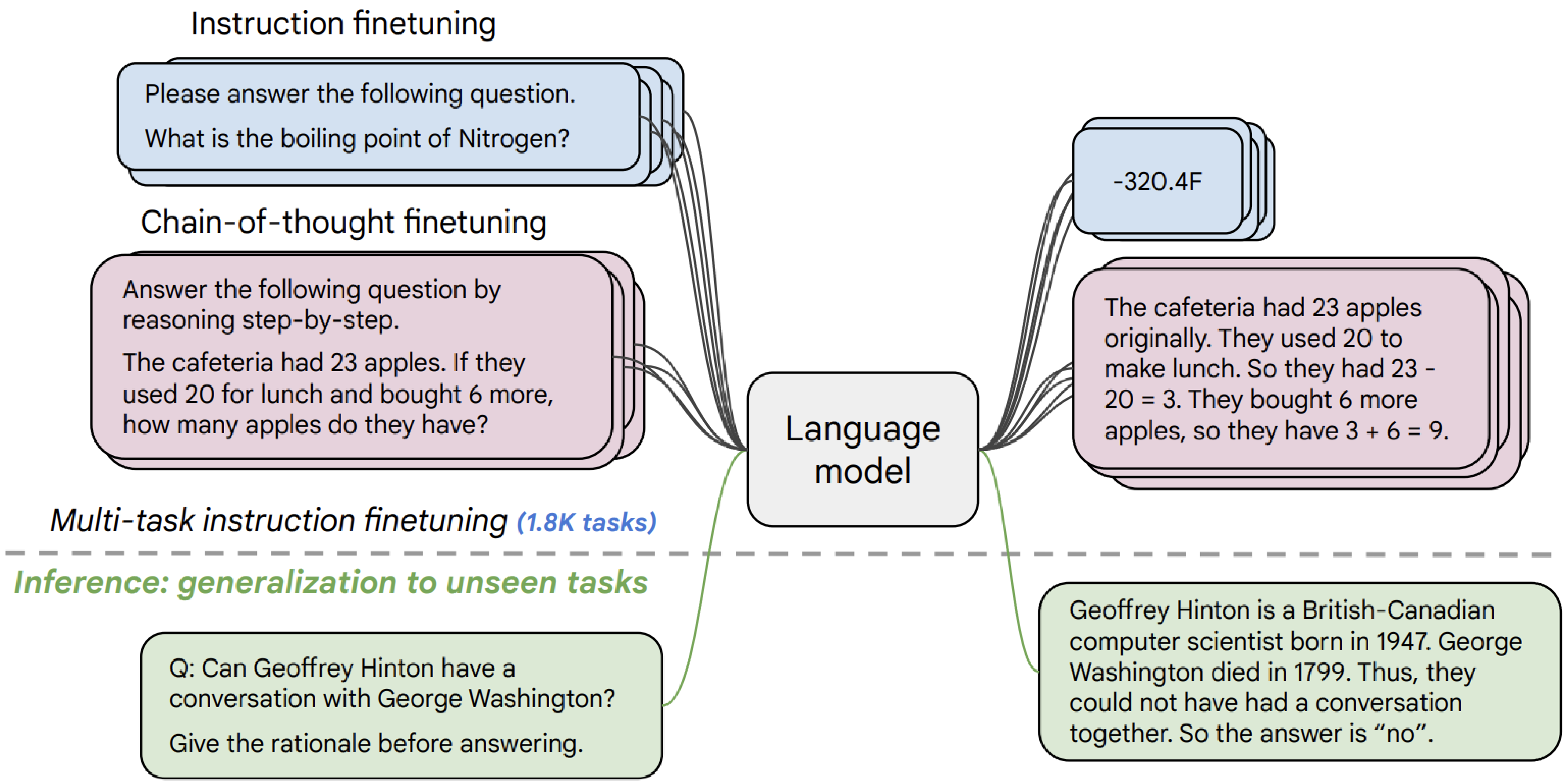
InstructGPT

```
People went to the moon, and they took pictures of what they saw, and sent
them back to the earth so we could all see them.
```

https://openai.com/research/instruction-following#sample1

Instruct Models

- This is achieved by a combination of fine-tuning them on many annotated instructions, as well as a reinforcement learning procedure
- More on this in the next lesson



Scaling Instruction-Finetuned Language Models
https://arxiv.org/pdf/2210.11416.pdf

Language Model Evaluation

- There are no good ways to evaluate language models

- Because
  - language models are used for very diverse goals
  - there are generally many very different ways to correctly follow an instruction or a set of examples
  - there are no good metrics to compare whether two texts are near-equivalent as responses to a task (e.g., similar translations, similar summaries etc.)

Language Model Evaluation

- Several directions are employed:
  - Multiple-choice tests
  - Building embedding-based metrics for comparing between machine output and a reference

Language Model Evaluation

- Several directions are employed:
  - Multiple-choice tests

Microeconomics

One of the reasons that the government discourages and regulates monopolies is that
(A) producer surplus is lost and consumer surplus is gained. ✗
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency. ✗
(C) monopoly firms do not engage in significant research and development. ✗
(D) consumer surplus is lost with higher prices and lower levels of output. ✓

Figure 3: Examples from the Microeconomics task.

Conceptual Physics

When you drop a ball from rest it accelerates downward at $9.8\ m/s^2$. If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) $9.8\ m/s^2$ ✓
(B) more than $9.8\ m/s^2$ ✗
(C) less than $9.8\ m/s^2$ ✗
(D) Cannot say unless the speed of throw is given. ✗

College Mathematics

In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a
(A) pair of points ✗
(B) circle ✗
(C) half-line ✗
(D) line ✓

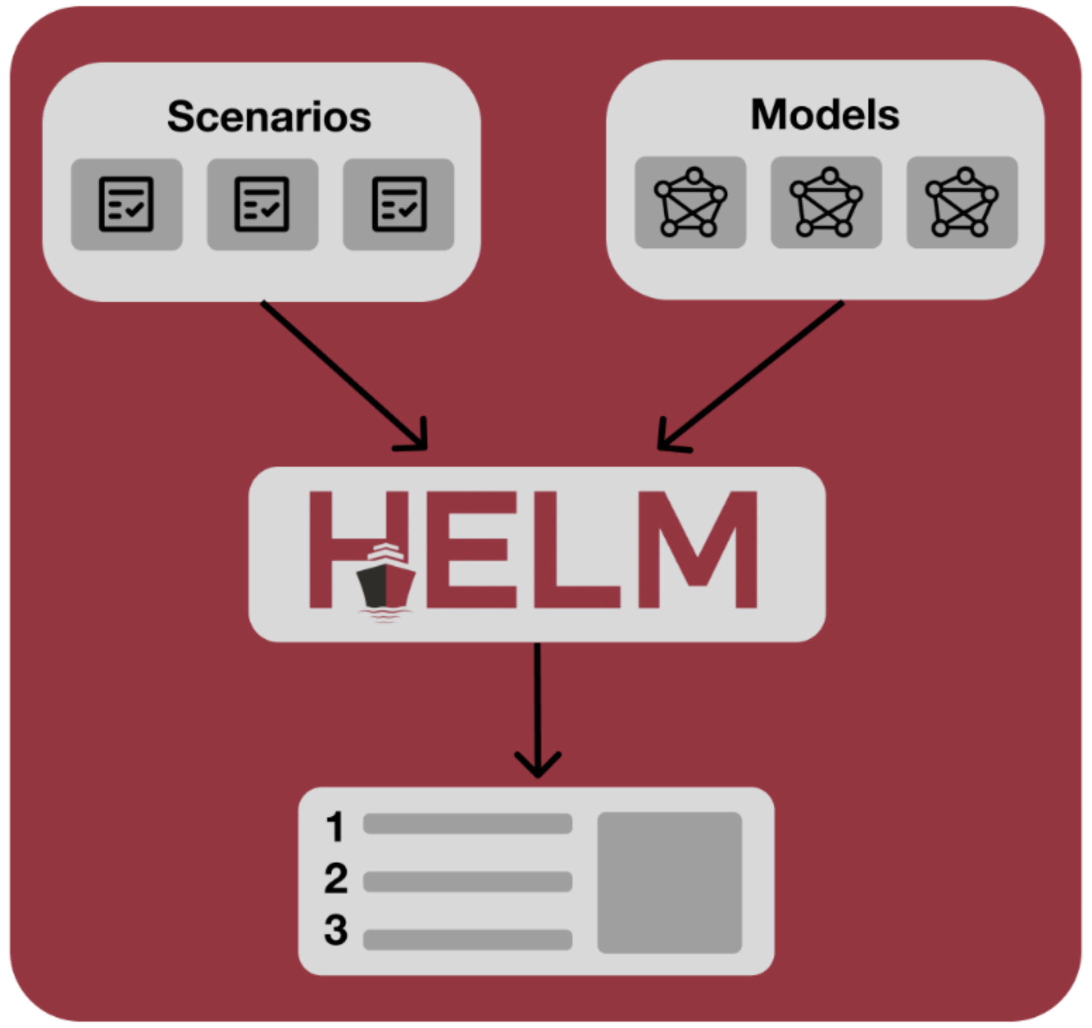
Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

https://arxiv.org/abs/2009.03300

Language Model Evaluation

- Benchmarks of many tests
    - HELM

A holistic framework for evaluating foundation models.



Model	Mean win rate
GPT-4 (0613)	0.965
GPT-4 Turbo (1106 preview)	0.842
Palmyra X V3 (72B)	0.832
Palmyra X V2 (33B)	0.794
PaLM-2 (Unicorn)	0.784
Yi (34B)	0.781

SEE MORE

https://crfm.stanford.edu/helm/lite/latest/

Language Model Evaluation

- Embedding-based reference-based metrics

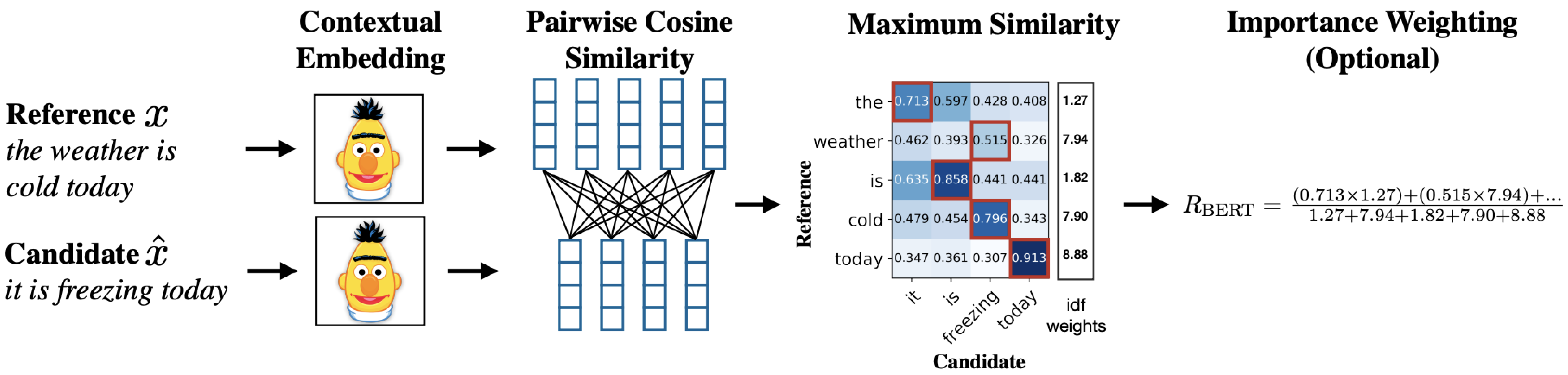


Figure 1: Illustration of the computation of the recall metric R_{BERT}. Given the reference x and candidate $\hat{x}$, we compute BERT embeddings and pairwise cosine similarity. We highlight the greedy matching in red, and include the optional idf importance weighting.

https://arxiv.org/pdf/1904.09675.pdf

Conclusion

- Pretraining has been a game changer in NLP in the last couple of years
- The first steps were contextualized word embeddings, that create word embeddings specific to the text where the word is situated
  - These are produced by a model that is later fine-tuned in a supervised manner
- The next step uses larger models allowed zero-shot/few-shot learning
- The latest are instruct models, which follow instructions in natural language