BERT:

- Example of the encoder-only paradigm
- **Pretuning**: train w/ self-supervised obj called "masked LM"
- **Finetuning**: process of adapting the pretrained model to a particular downstream task

**Pretuning**:

```
I
II
III
IV
V

Unmasked Multitask Self-Attend
```

[CLS] student opened [MASK] books

**Finetuning**: sentiment

```
I
II
III
IV
V

new softmax layer \( W_s \)
\[ 0 = \text{softmax}(W_s \cdot h_{[CLS]}) \]
⇒ require a labeled training dataset for the downstream task

[CLS] the movie is good

```
Extractive QA

Inputs:
- a question
- a paragraph that contains the answer

Output:
- a span of the paragraph that answers the question

Datasets:
- SQuAD v1, 2
- QAC / CoQA
- Hotpot QA

Q: Who starred in the Matrix as Neo?
P: \( v_1, v_2, v_3 \ldots \) Neo was played by actor Keanu Reeves
A: \((i, j)\)

How do we use BERT for extractive QA?

2 softmax layers on each token in passage
- predict beginning, end index of answer span

[CLS] Who starred in the Matrix [SEP] \( w_1, w_2, w_3 \ldots \) Keanu Reeves ...
how do we select an answer span at test time?

→ find the span \( w_i, \ldots, j \) that maximizes

\[
P_{\text{START}}(i) \cdot P_{\text{END}}(j)
\]

→ exclude spans where \( j < i \)
→ exclude spans longer than a threshold

advanced variants of BERT:

→ pretraining improvements \( \implies \) RoBERTa
  → more data

→ longer sequences during pretraining
  → BERT \( \implies \) 512 tokens max
  → XLNet \( \implies \) 900 tokens

→ more pretraining objectives

→ ELECTRA

→ "students opened [MASK] books"

→ "students opened monster books"

\( \text{real, corrupted, real, corrupted, real} \)

→ smaller models

→ tiny BERT, distilBERT, ALBERT
Transformer LMs at test-time:

- distillation
- pruning

Decoder-only

At training time, all of these computations happen in parallel

Students opened their books

→ these four predictions can happen simultaneously b/c we already know the identity of the gold next tokens (ground-truth)
At test-time, I have to \underline{decode} the output word-by-word, because I don't have access to the gold next tokens.

\(\Rightarrow\) decoding algorithms