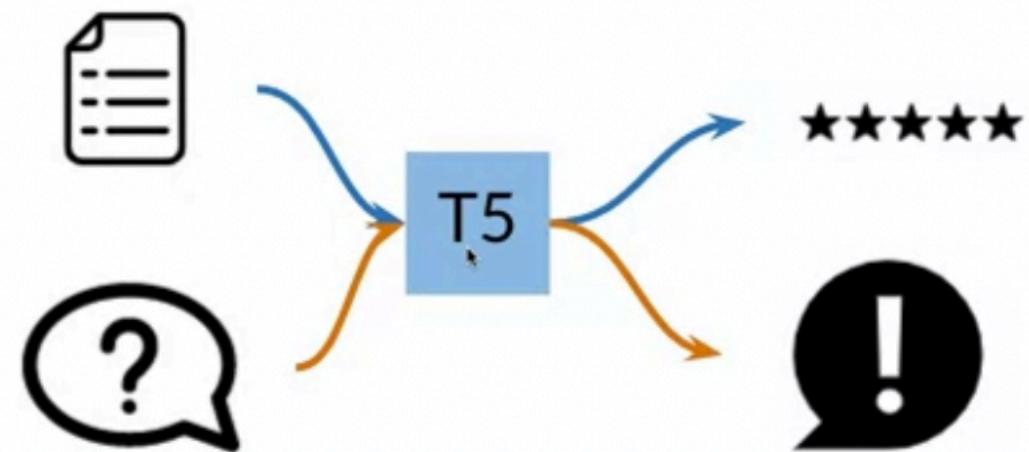


Week 3

Question
Answering

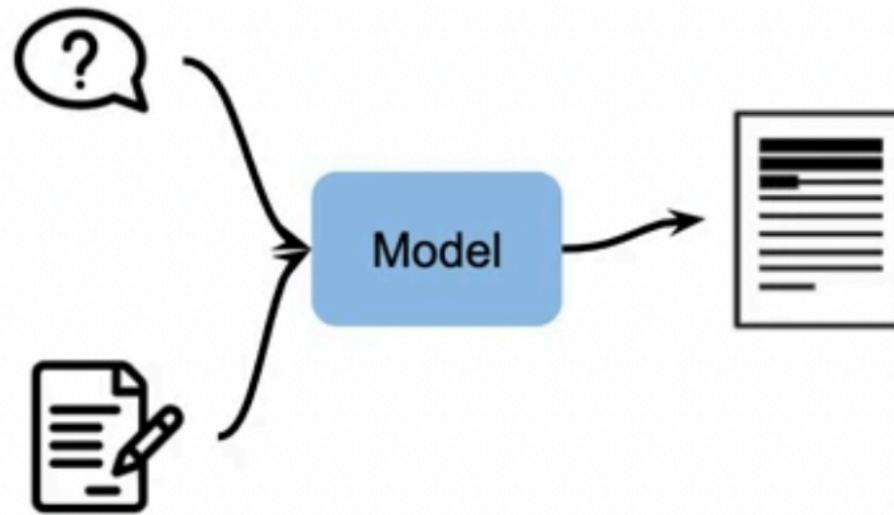


Transfer
learning

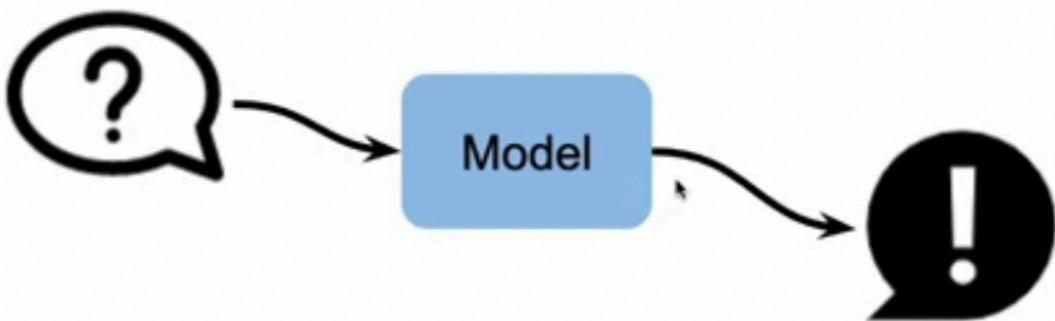


Question Answering

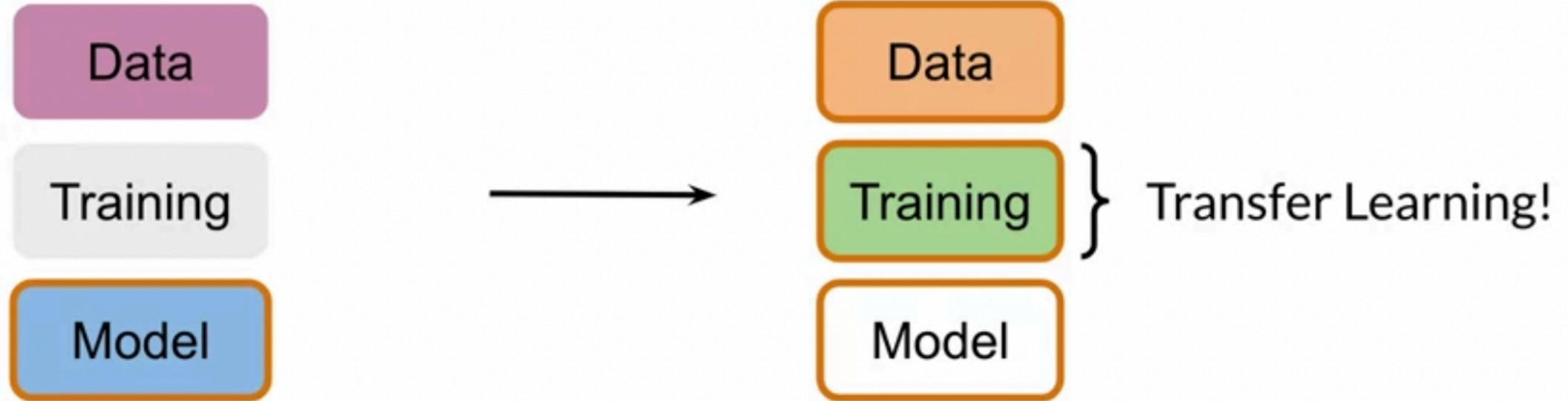
Context-based



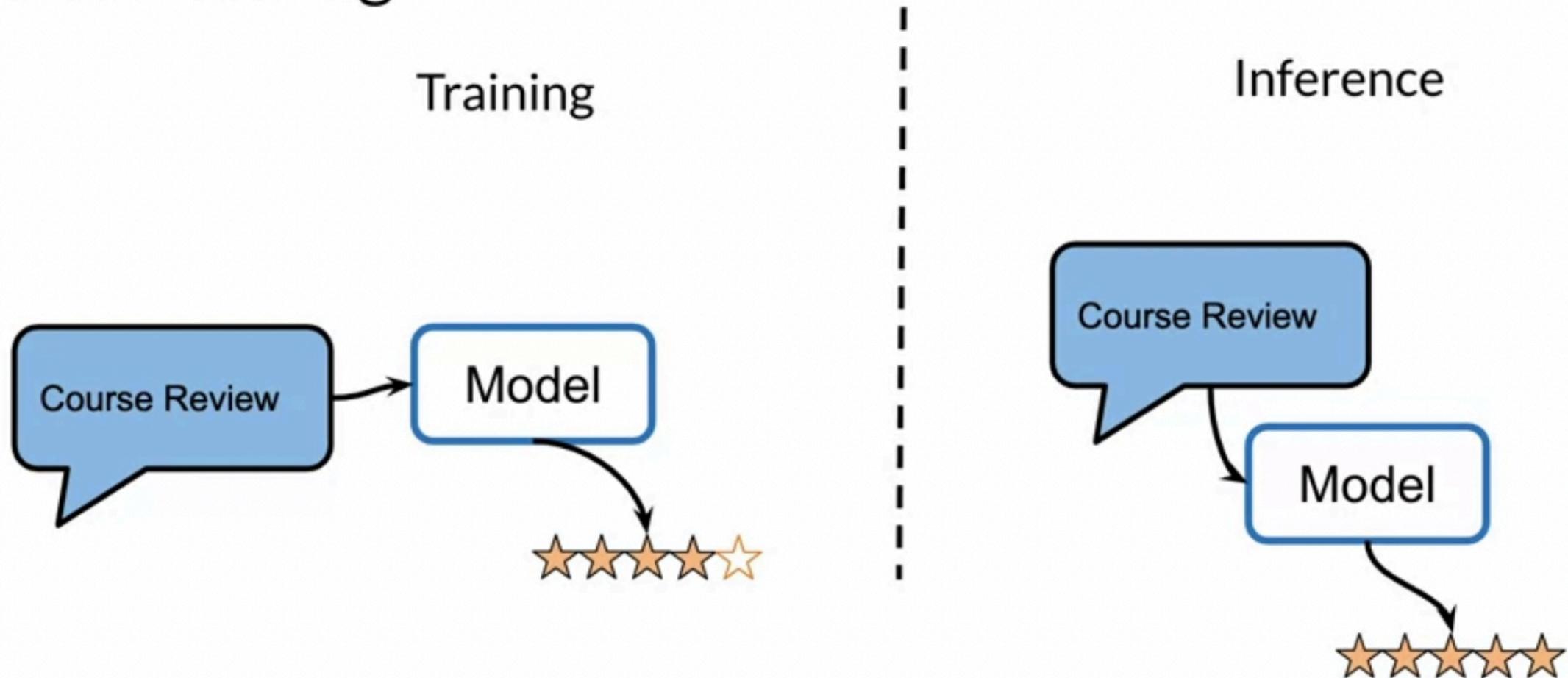
Closed book



Not just the model

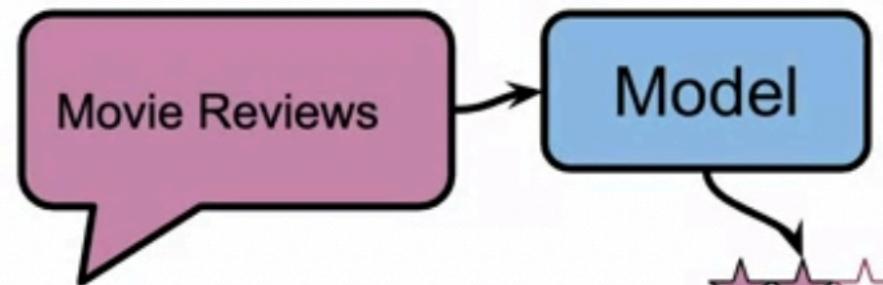


Classical training

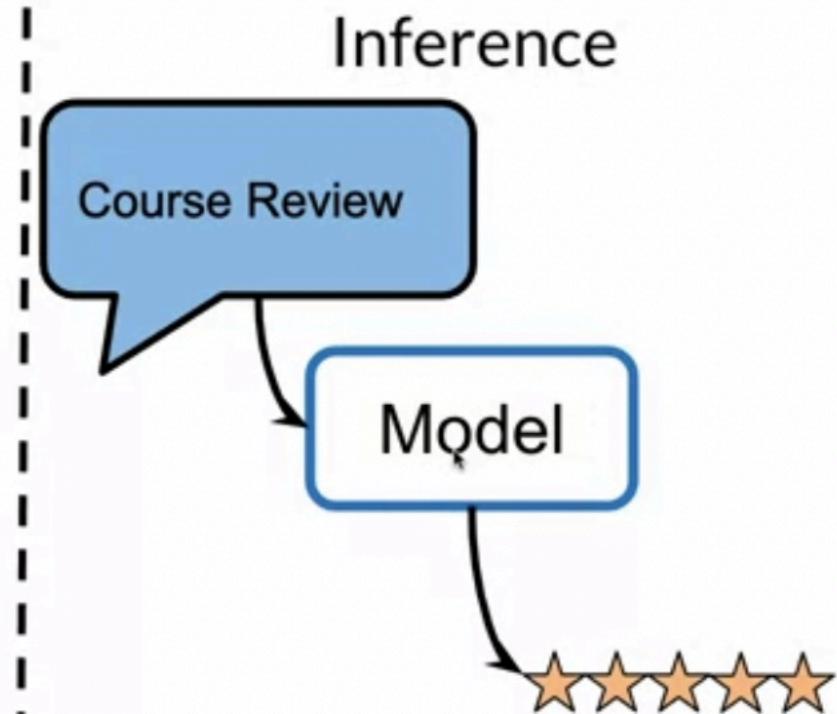
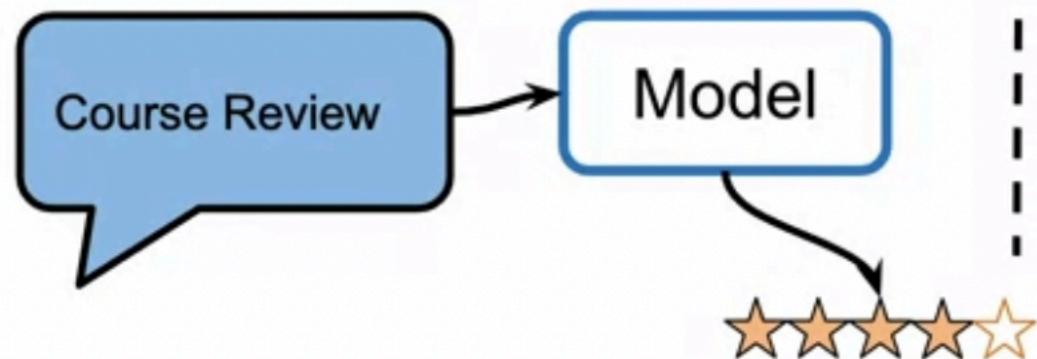


Transfer learning

Pre-training

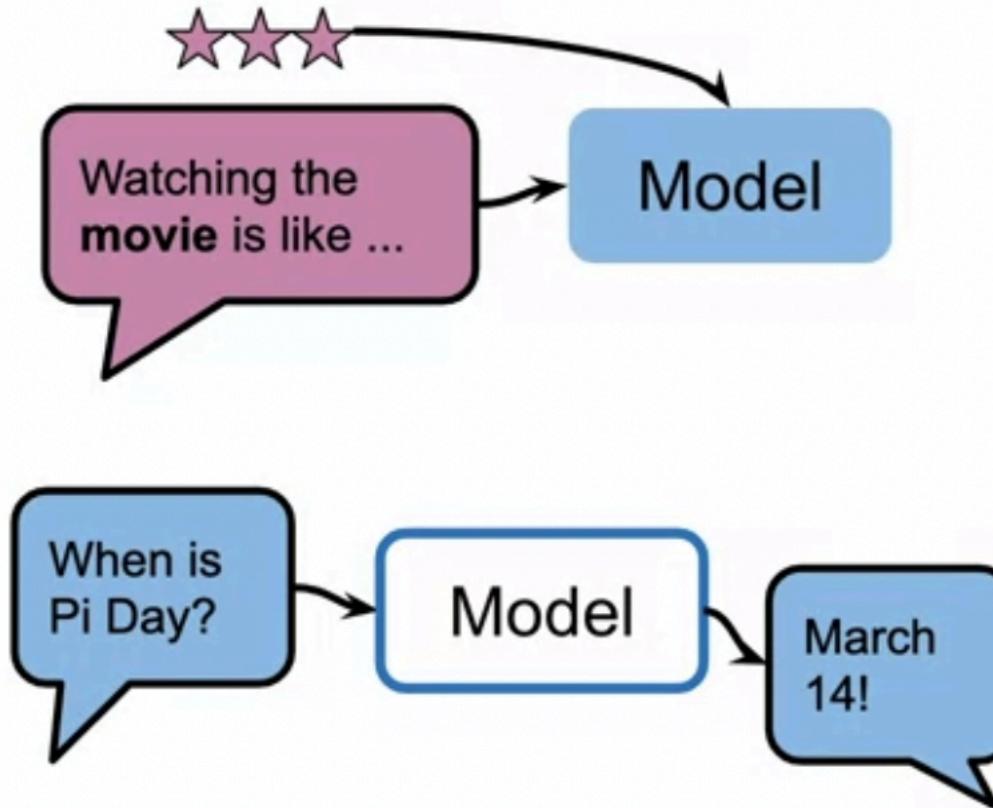


Training
on "Downstream" Task

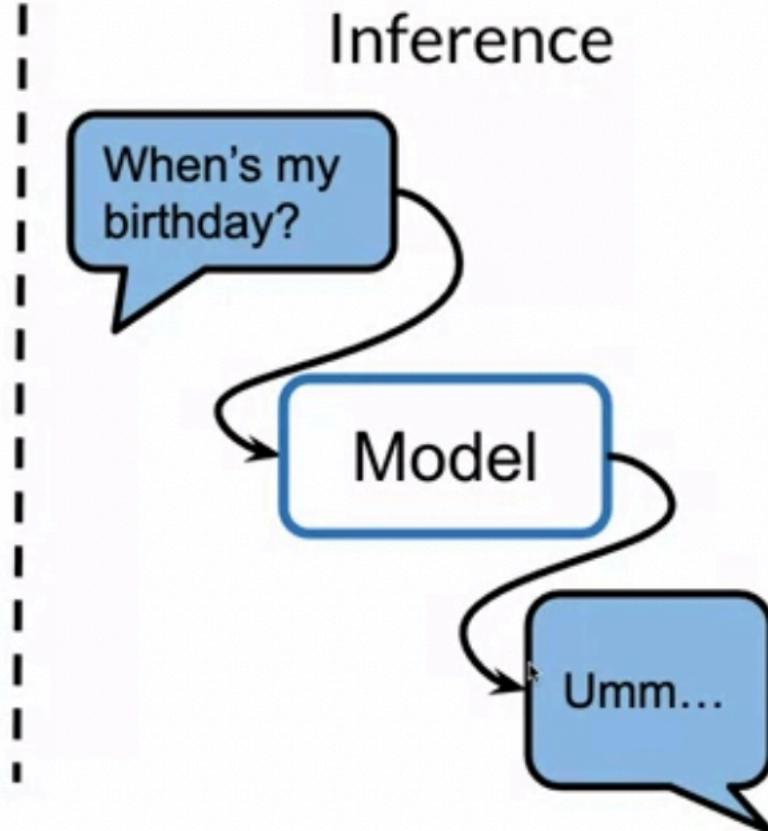


Transfer Learning: Different Tasks

Pre-Training
Sentiment
Classification

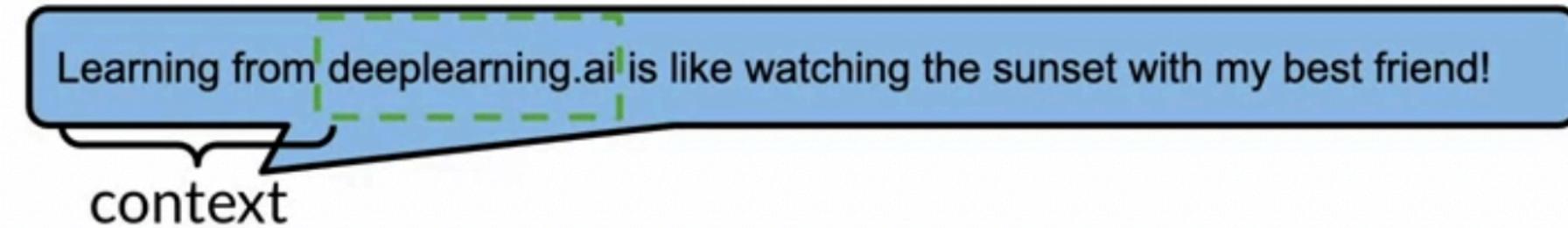


Training
Downstream task:
Question Answering



BERT: Bi-directional Context

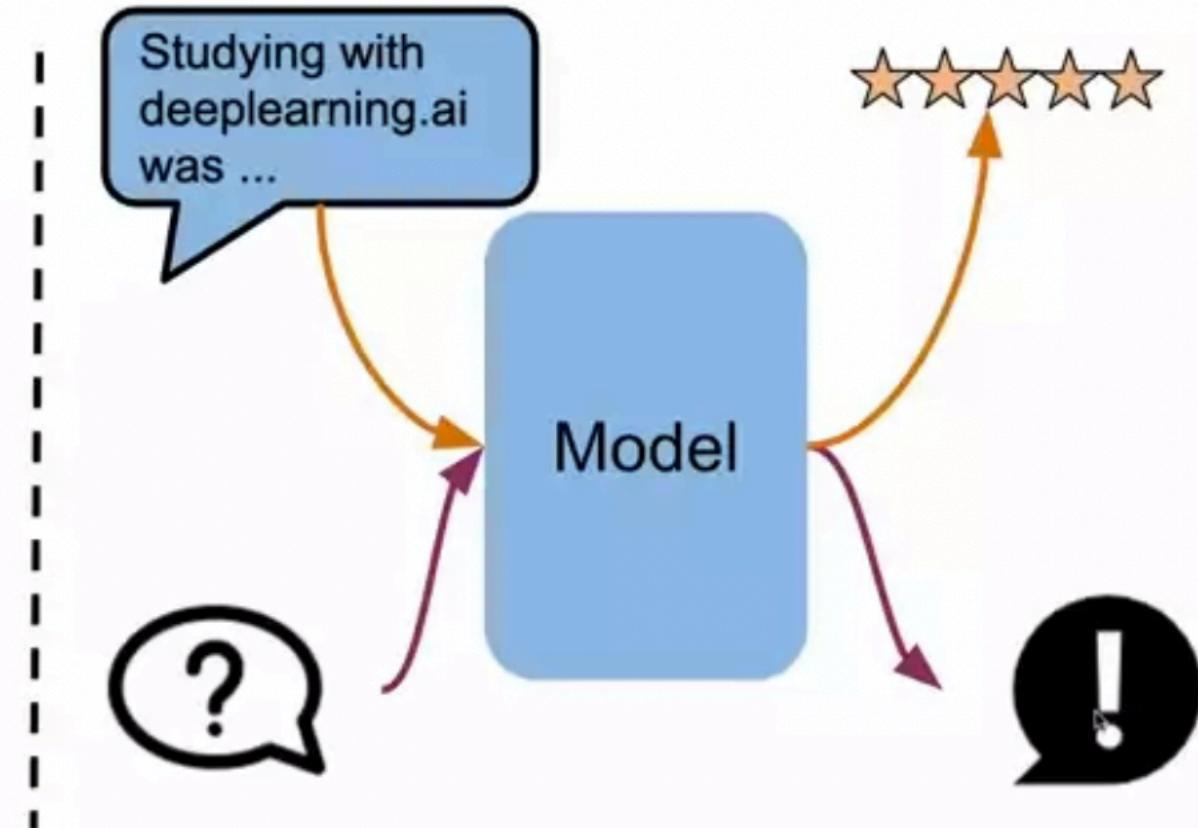
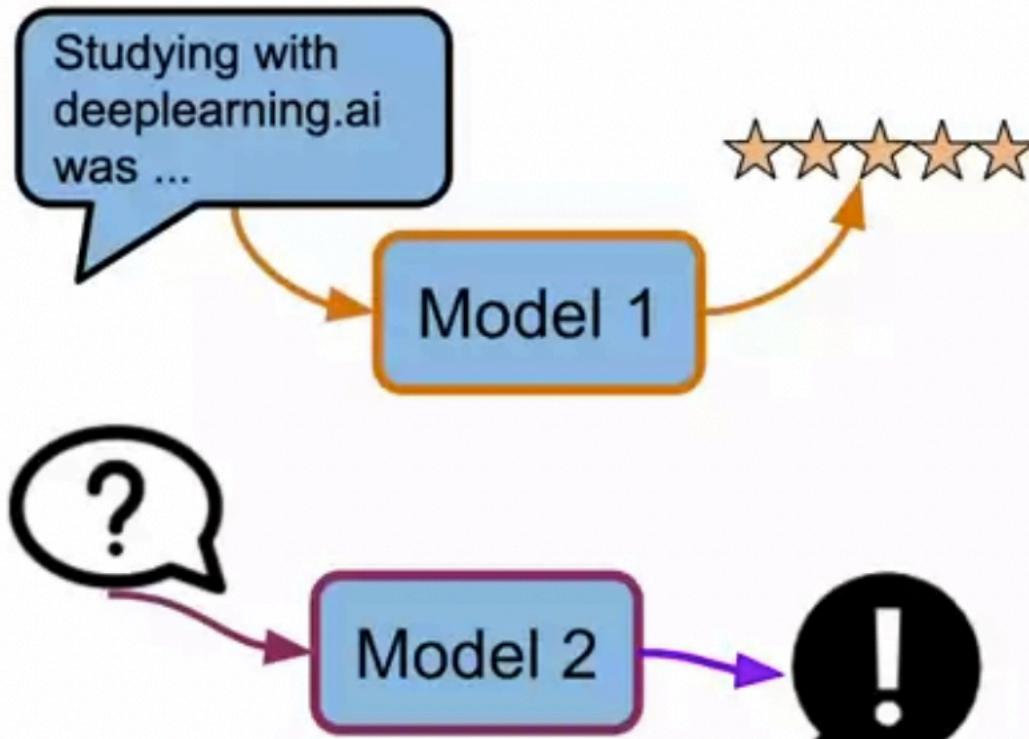
Uni-directional



Bi-directional

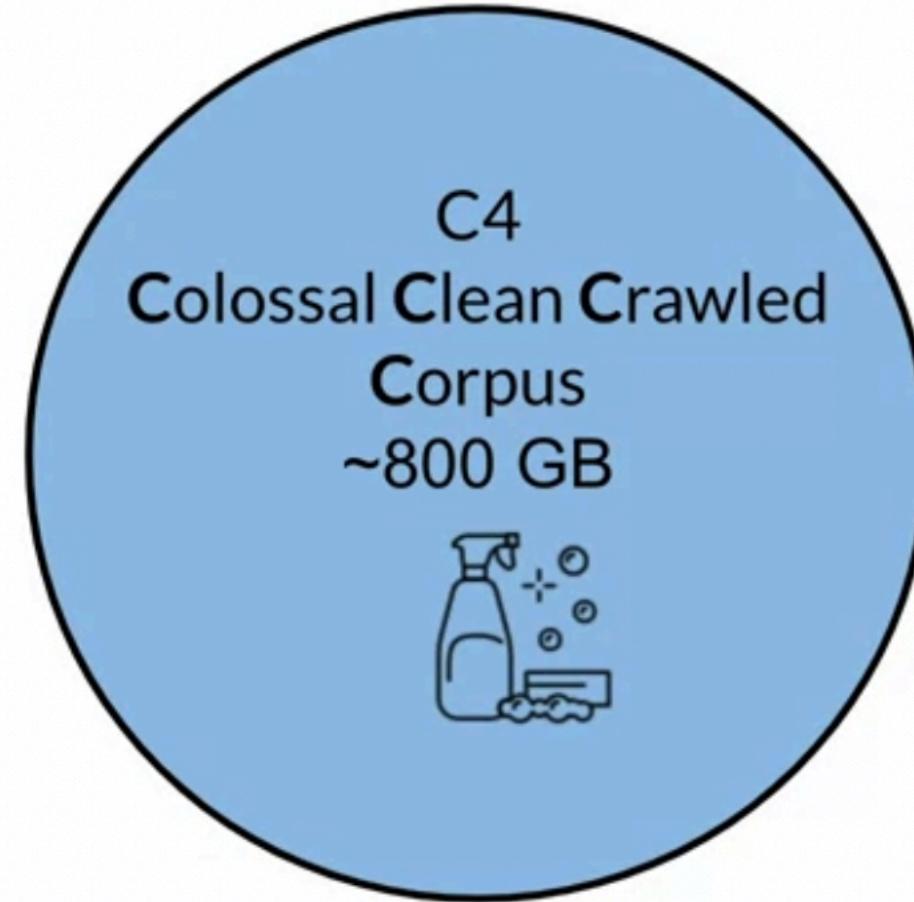


T5: Single task vs. Multi task



T5: more data, better performance

English wikipedia
~13 GB



Desirable Goals



- Reduce training time



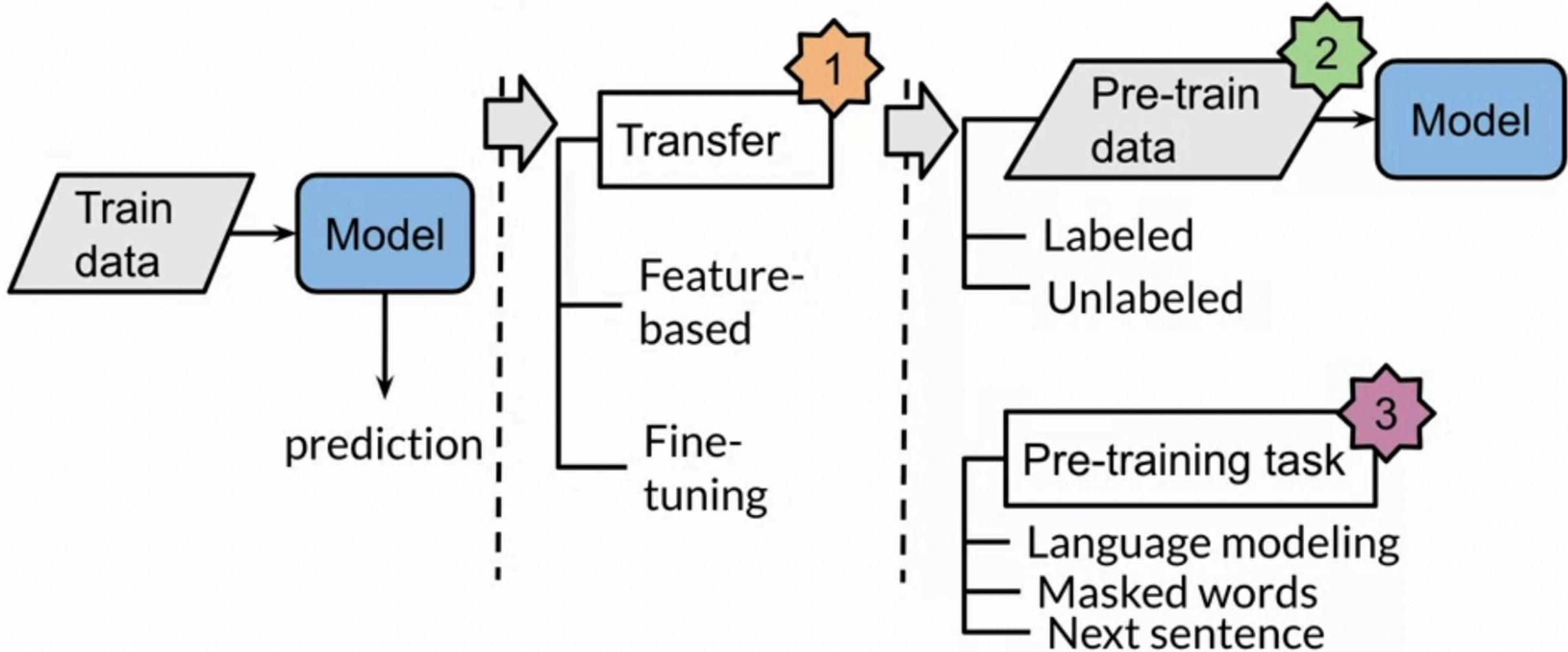
- Improve predictions



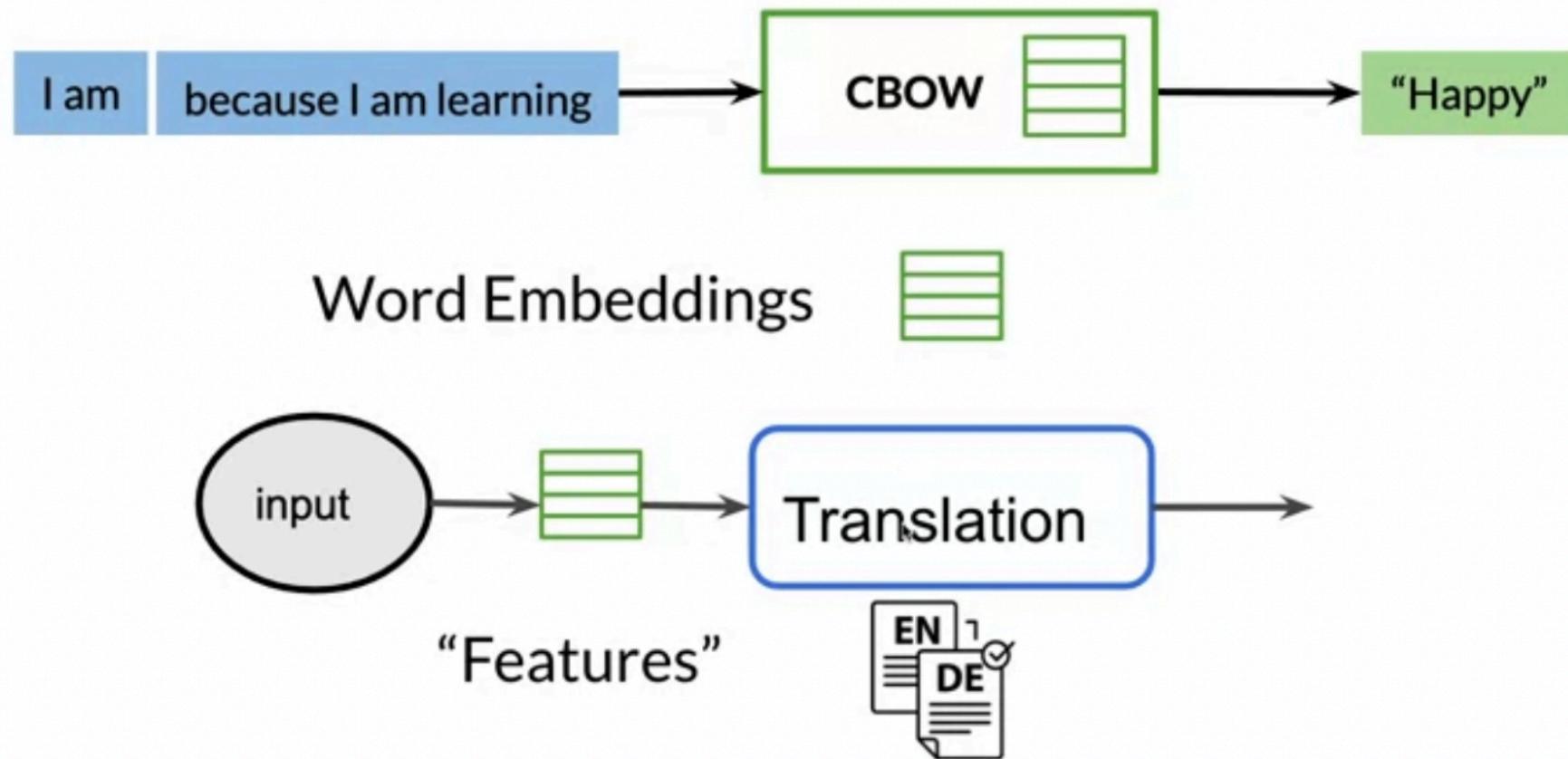
- Small datasets



Transfer learning options

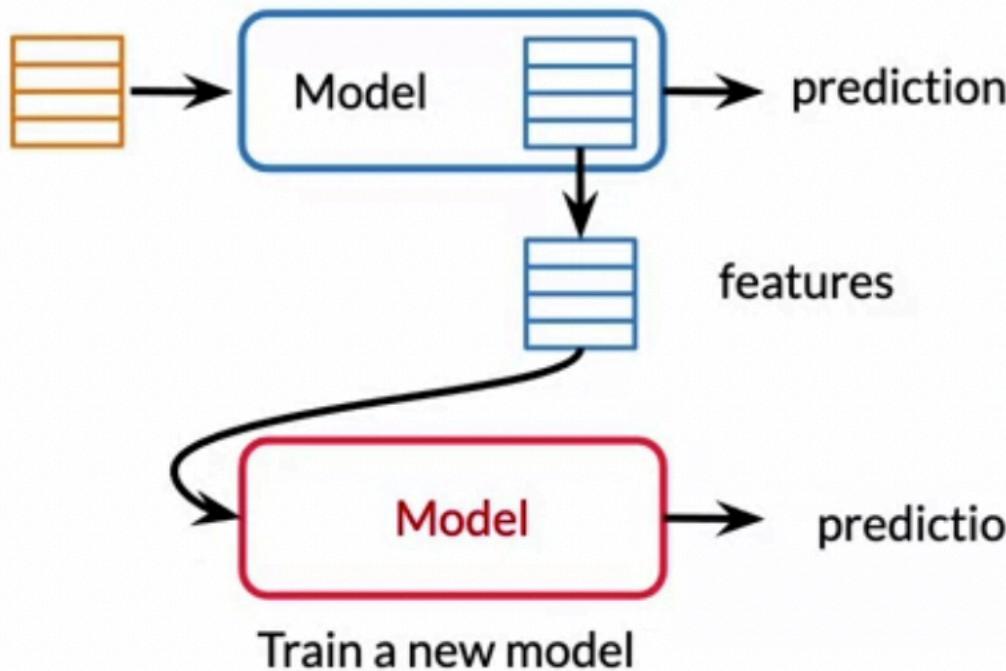


General purpose learning

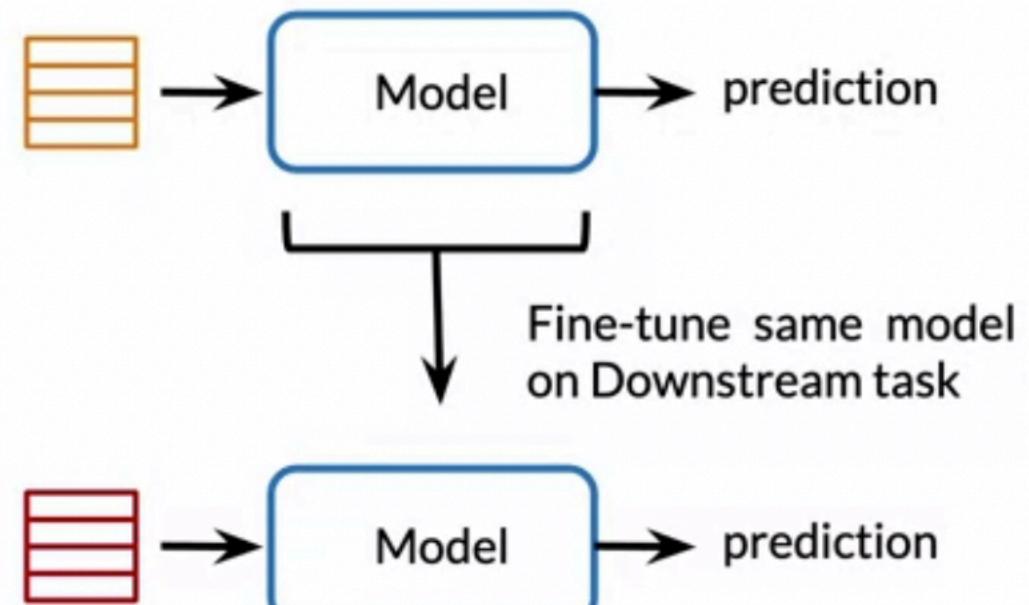


Feature-based vs. Fine-Tuning

Pre-Train



Pre-Train

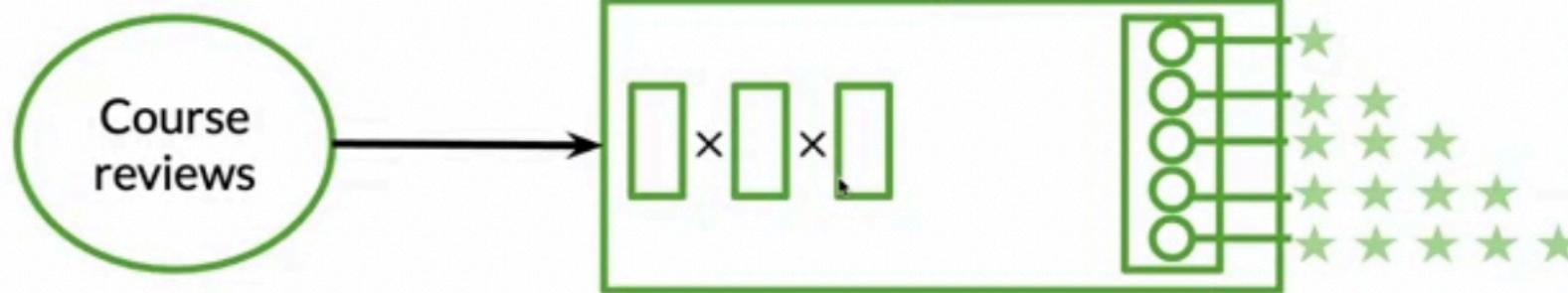
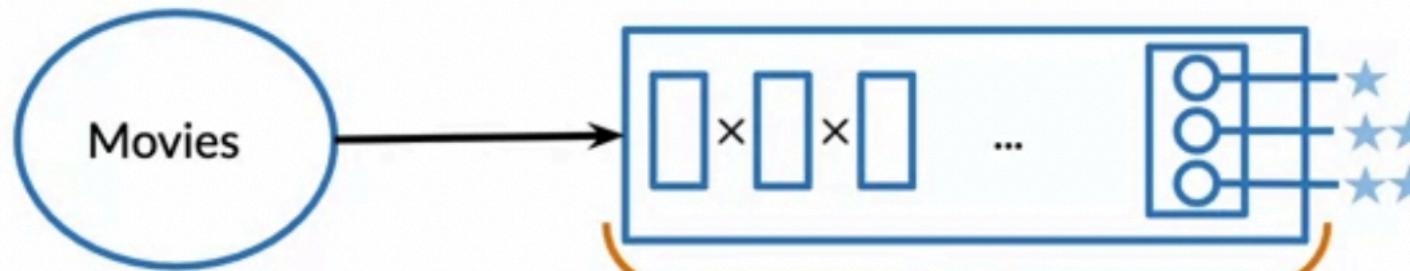


Fine-tune: adding a layer

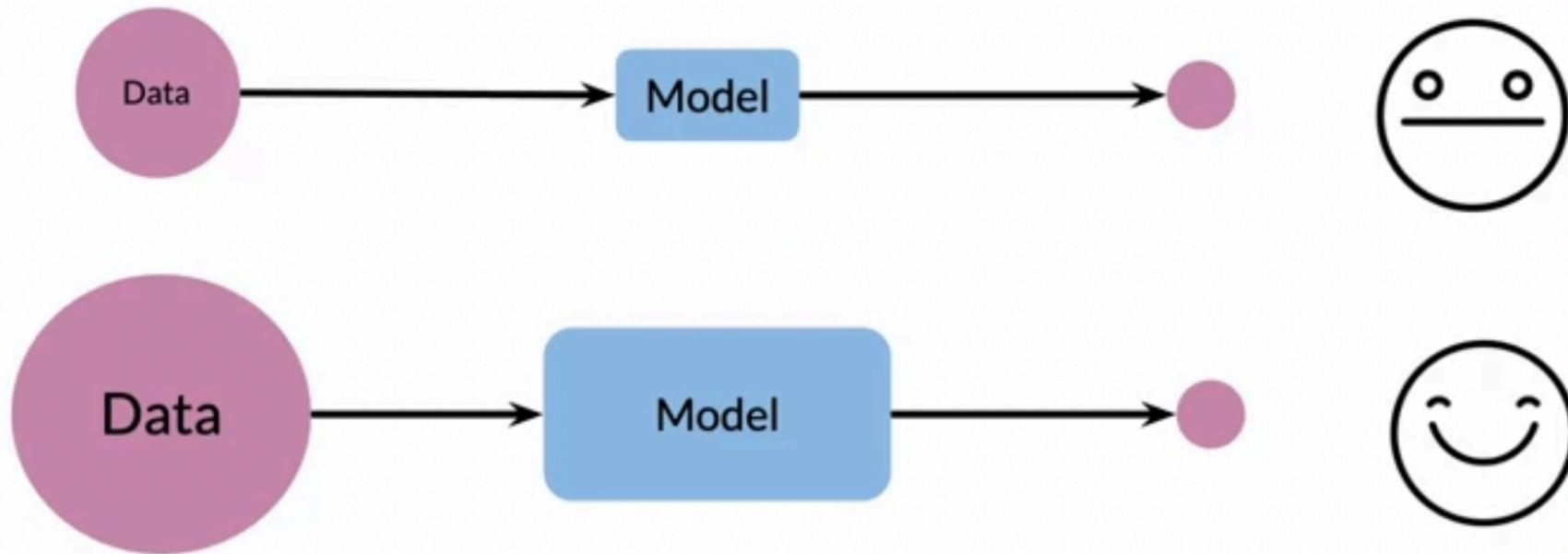
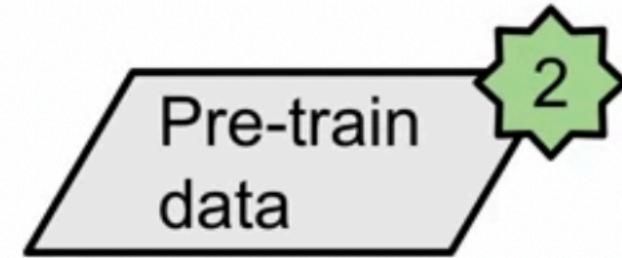
Transfer

1

Pre-Training



Data and performance



Labeled vs Unlabeled Data

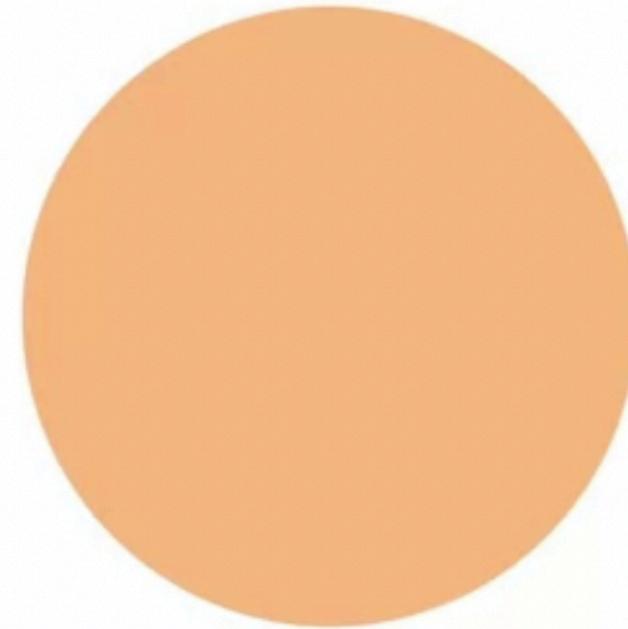
Pre-train
data

2

Labeled text data



Unlabeled text data



Transfer learning with unlabeled data

Pre-train
data

2

Pre-Training



Model

No labels !

Downstream task

What day is Pi day?

Model

March 14

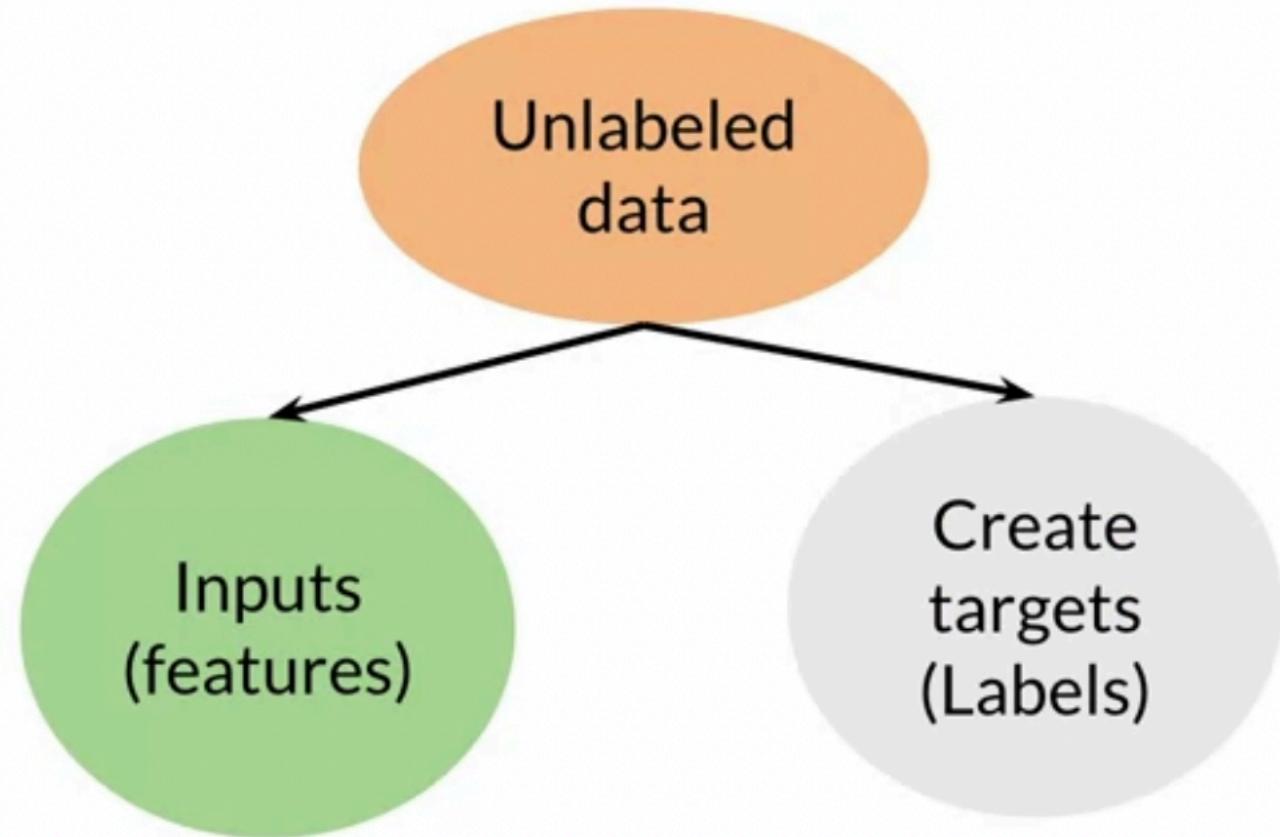
Labeled data

Which tasks work with
unlabeled data?

Self-supervised task

Pre-training task

3



Self-supervised tasks

3

Pre-training task

Unlabeled Data

Learning from deeplearning.ai
is like watching the sunset
with my best friend.

Input

Learning from deeplearning.ai
is like watching the sunset
with my best _____

Self-supervised tasks

3

Pre-training task

Unlabeled Data

Learning from deeplearning.ai
is like watching the sunset
with my best friend.

Input

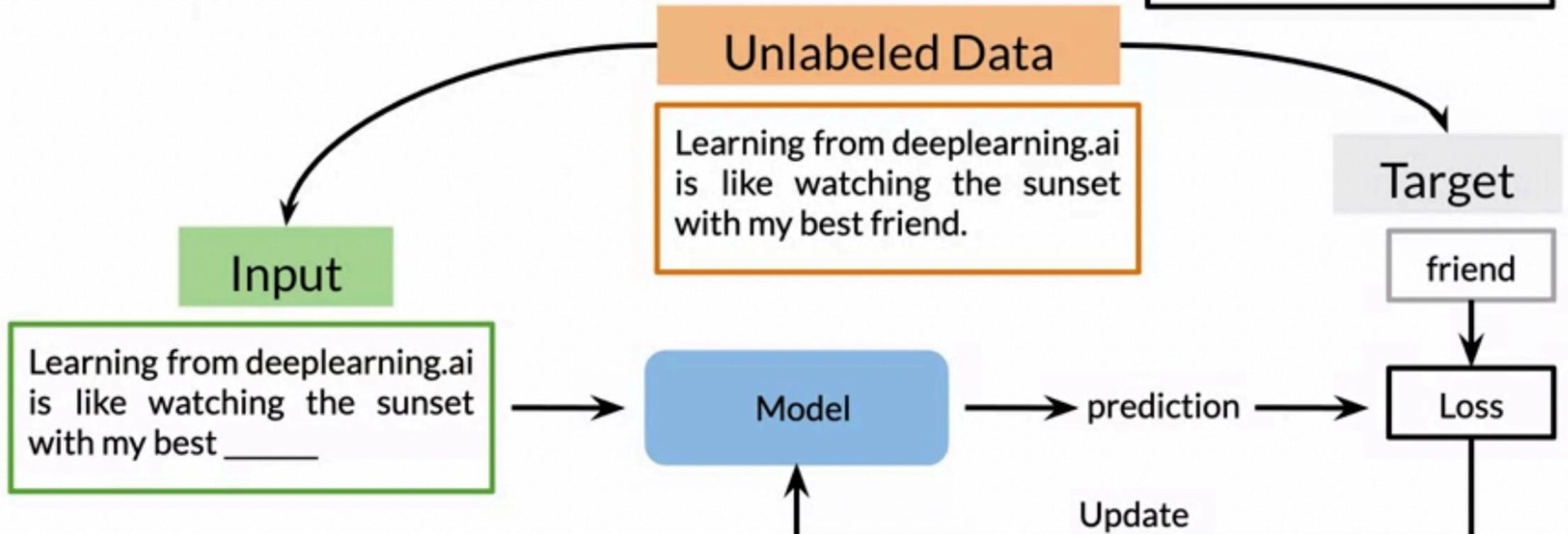
Learning from deeplearning.ai
is like watching the sunset
with my best _____



Model

Self-supervised tasks

3

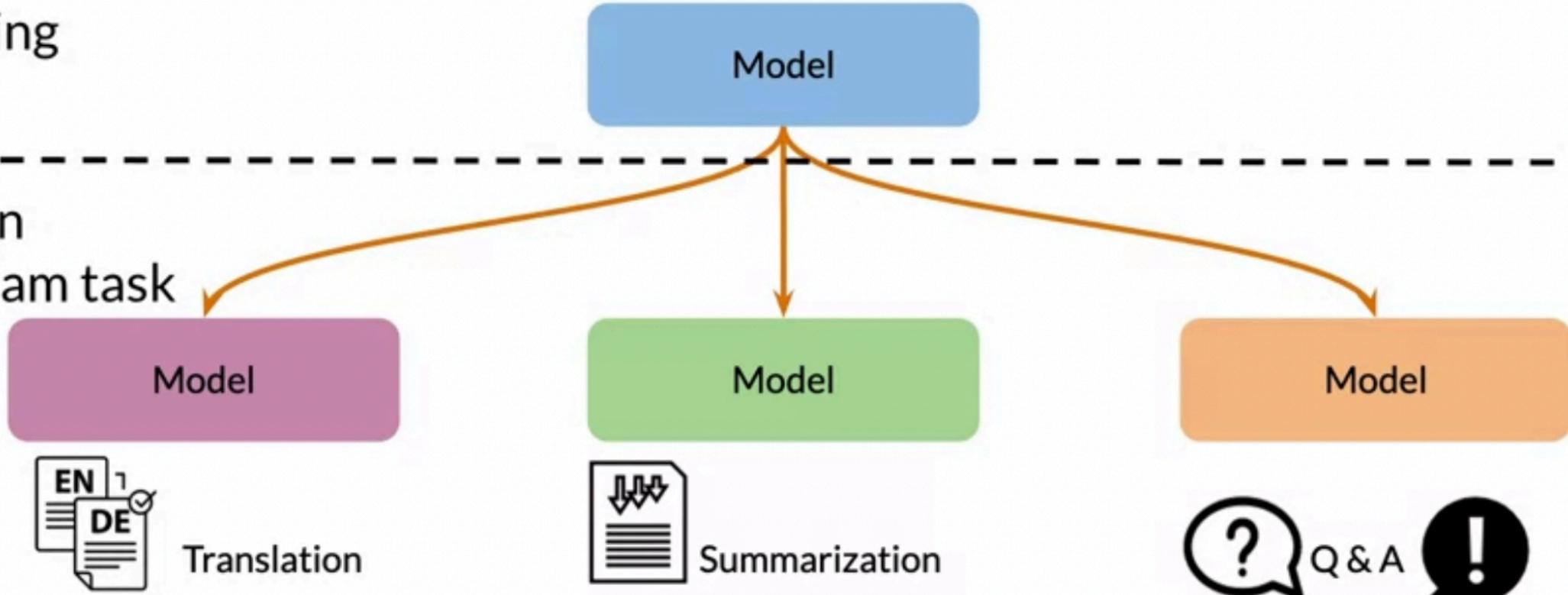


Language modeling

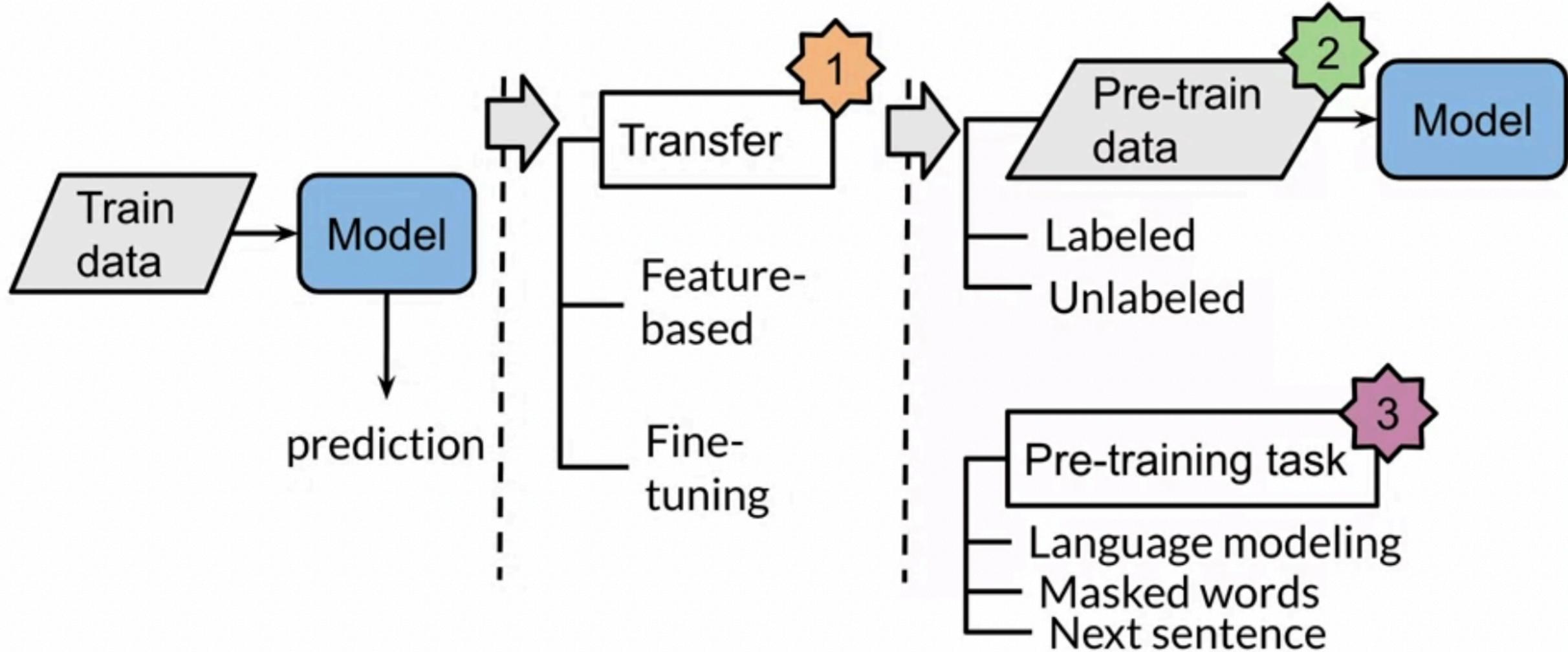
Fine-tune a model for each downstream task

Pre Training

Training on
Downstream task



Summary



Outline

CBOW ELMo GPT BERT T5



Context

... right ...

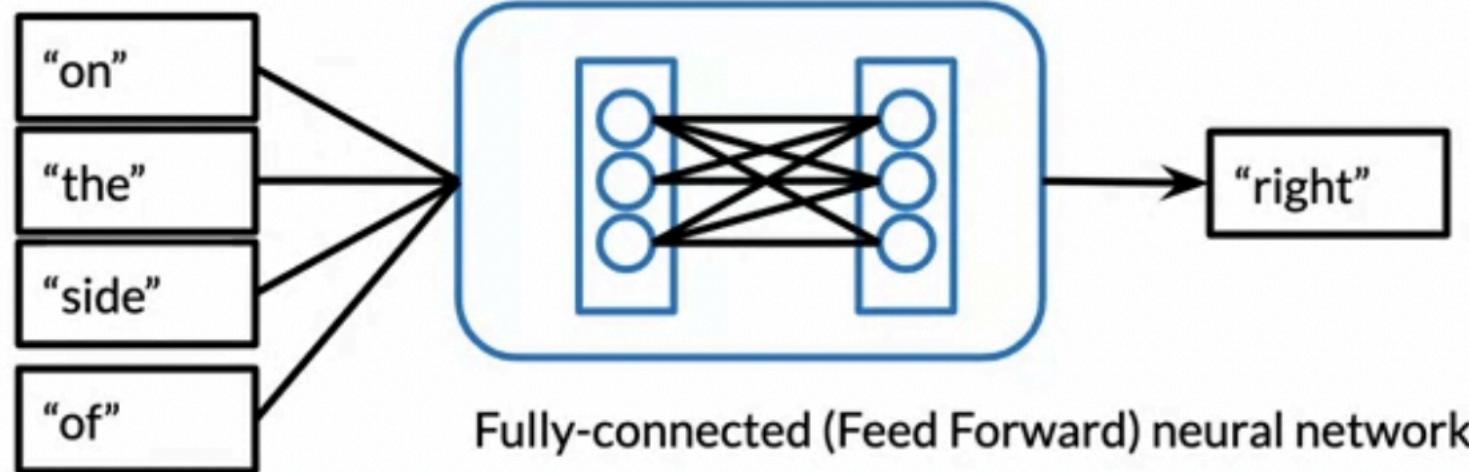
... they were on the right ...

... they were on the right side of the street

Continuous Bag of Words

... they were on the right side of the street

Fixed window Fixed window



Need more context?

... they were on the right side of **the street.**

Fixed window Fixed window

... they were on the right side of **history.**

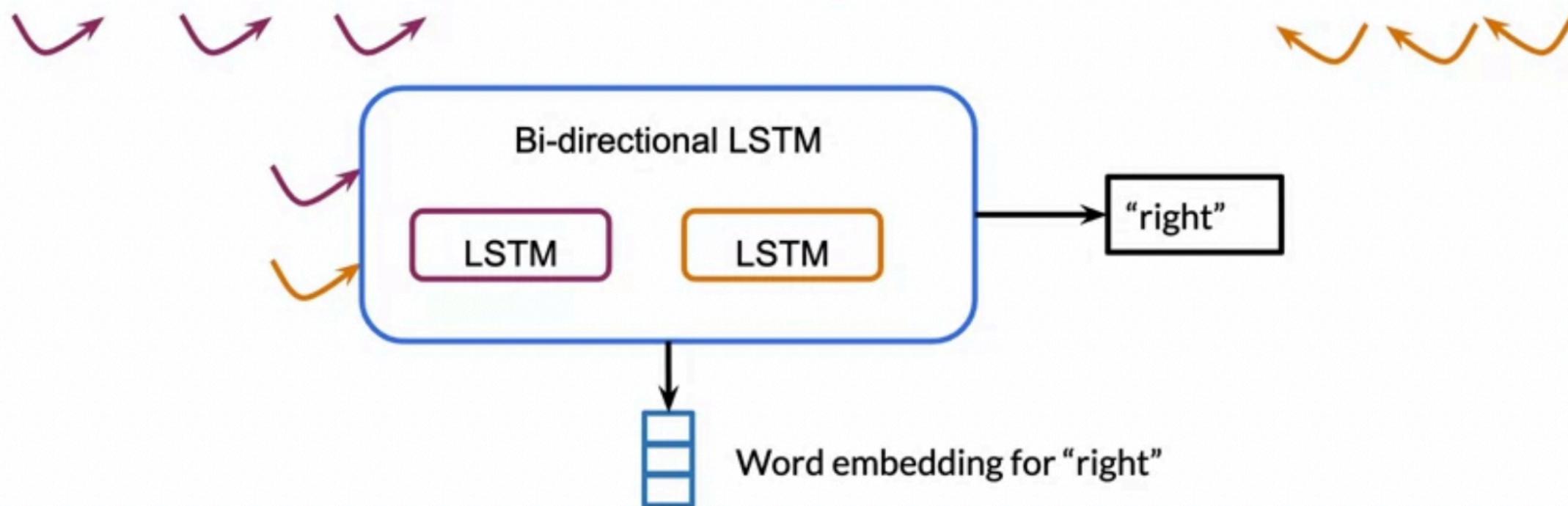
Use all context words

The legislators believed that they were on the **right** side of history, so they changed the law.



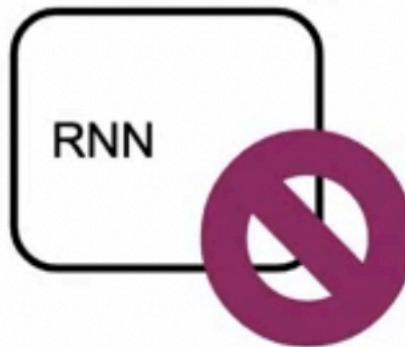
ELMo: Full context using RNN

The legislators believed that they were on the _____ side of history so they changed the law.

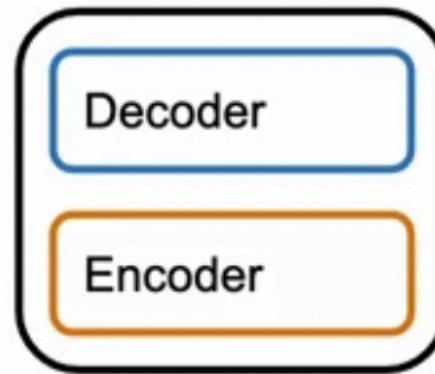


Open AI GPT

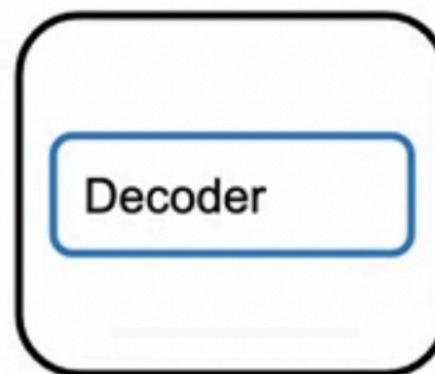
ELMo



Transformer



GPT

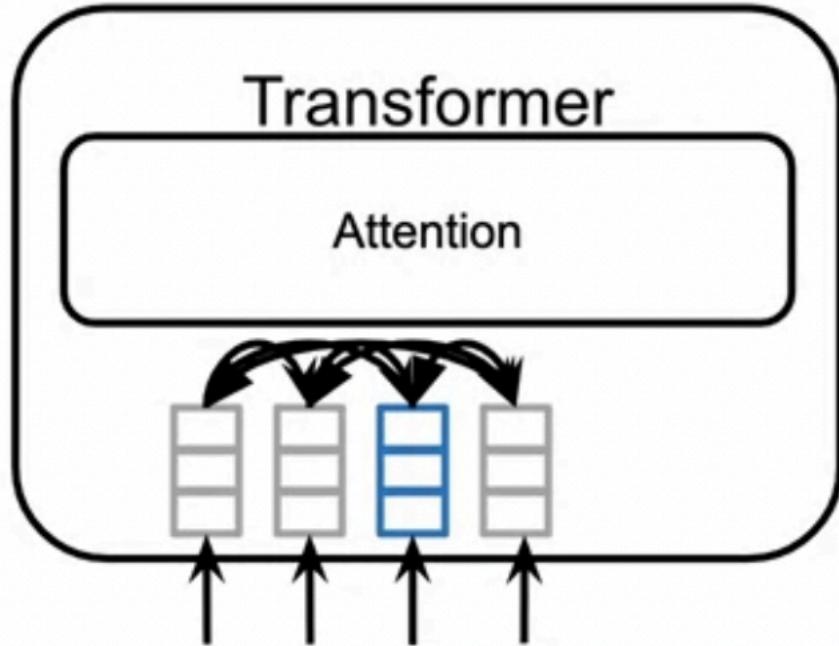


The legislators believed that they were on the _____



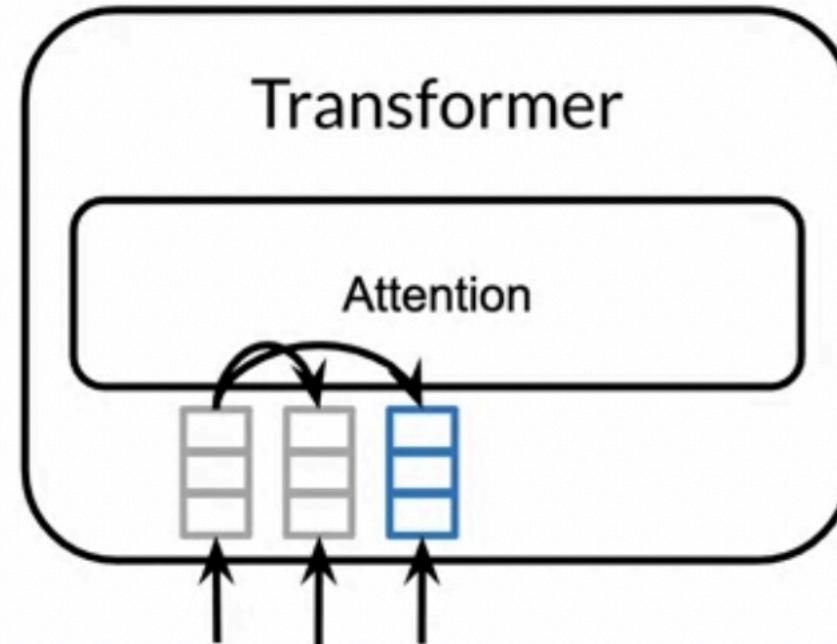
Uni-directional

GPT: Uni-directional



... on the right side...

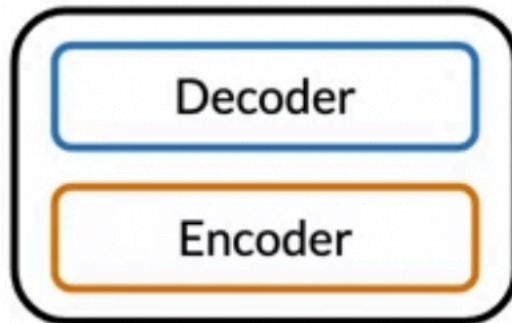
Each word can peek at itself!



... on the right
No peeking!

BERT

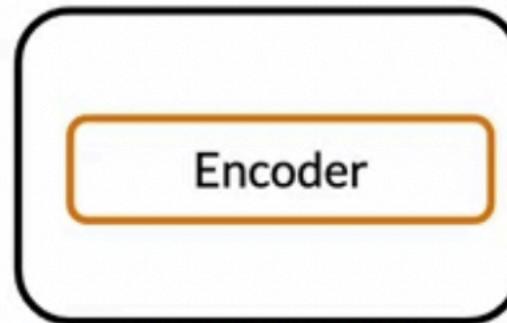
Transformer



GPT



BERT

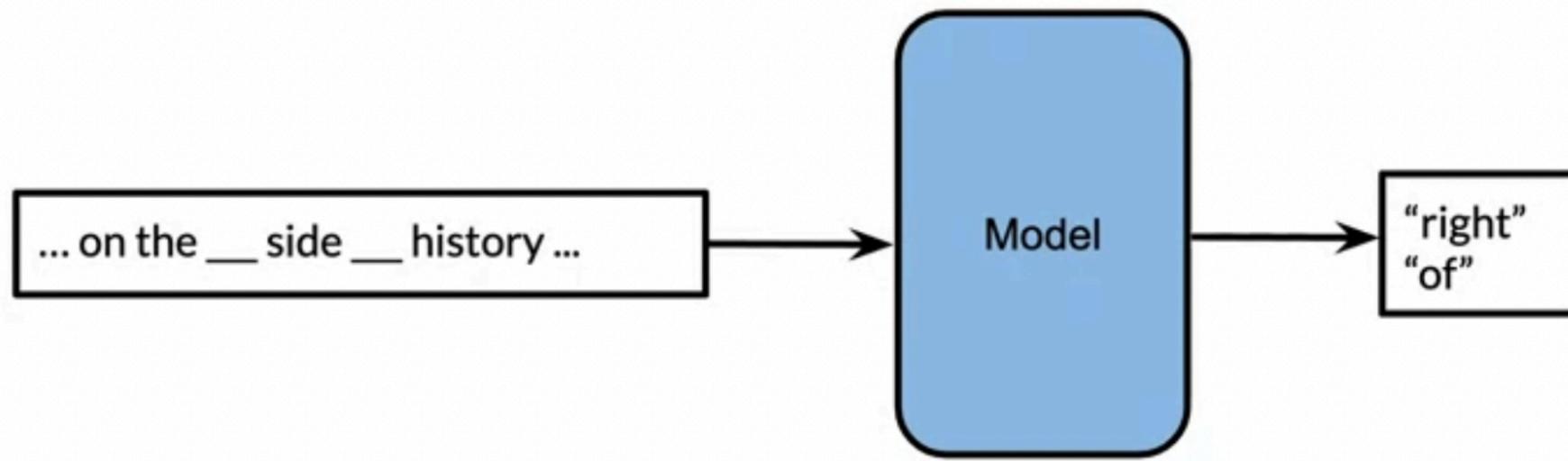


The legislators believed that they were on the ___ side of history, so they changed the law.



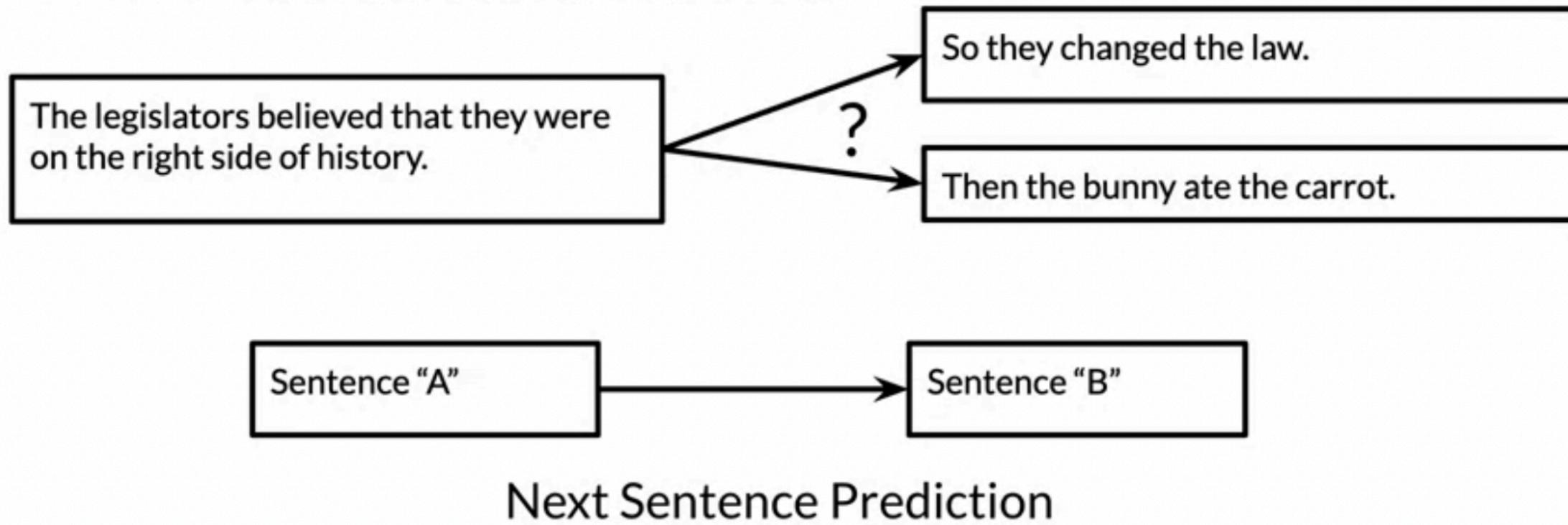
Bi-directional

Transformer + Bi-directional Context



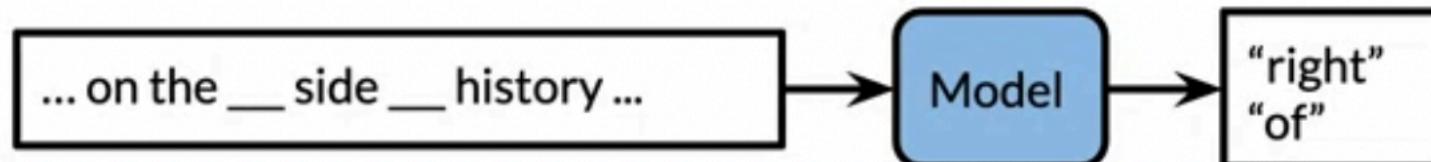
Multi-Mask Language Modeling

BERT: Words to Sentences

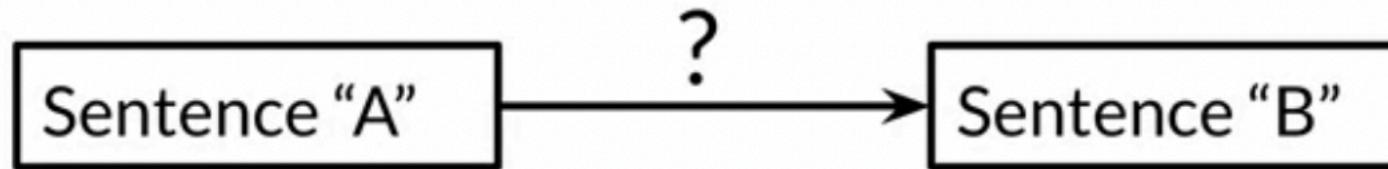


BERT Pre-training Tasks

Multi-Mask Language Modeling

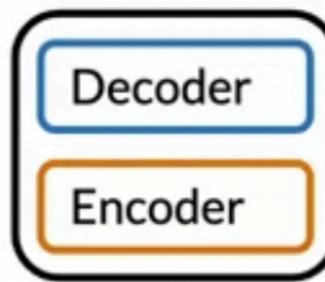


Next Sentence Prediction



T5: Encoder vs. Encoder-Decoder

Transformer



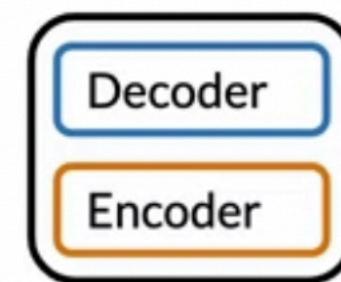
GPT



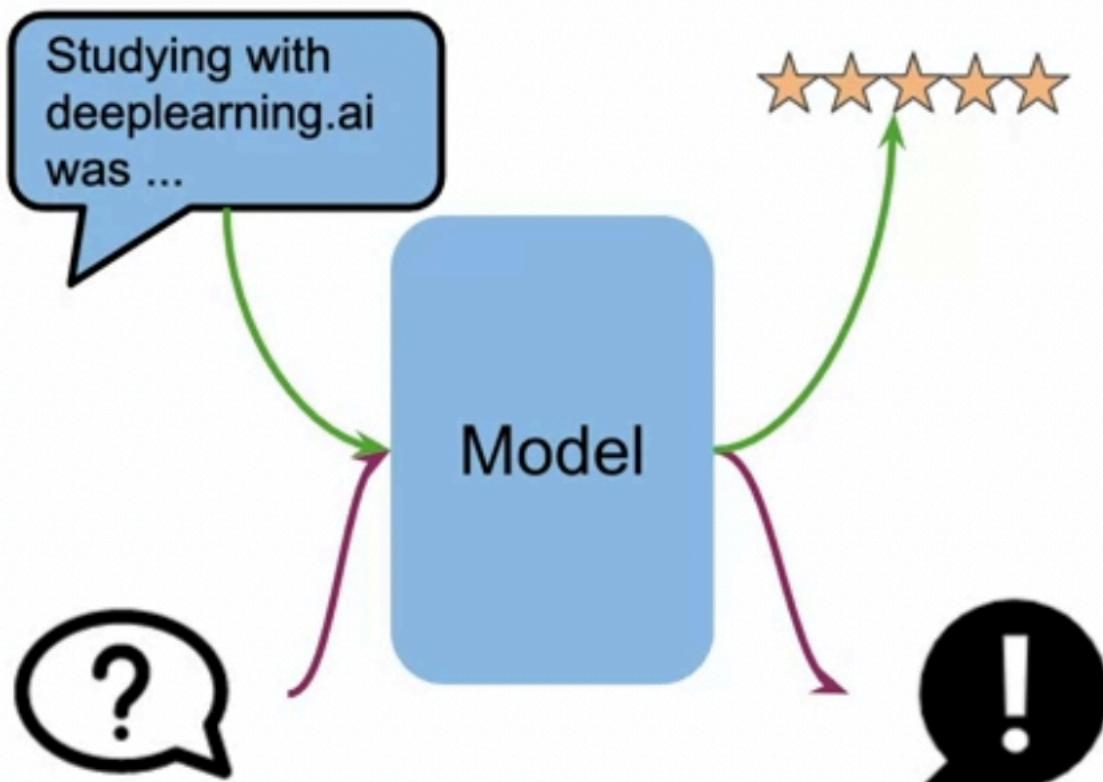
BERT



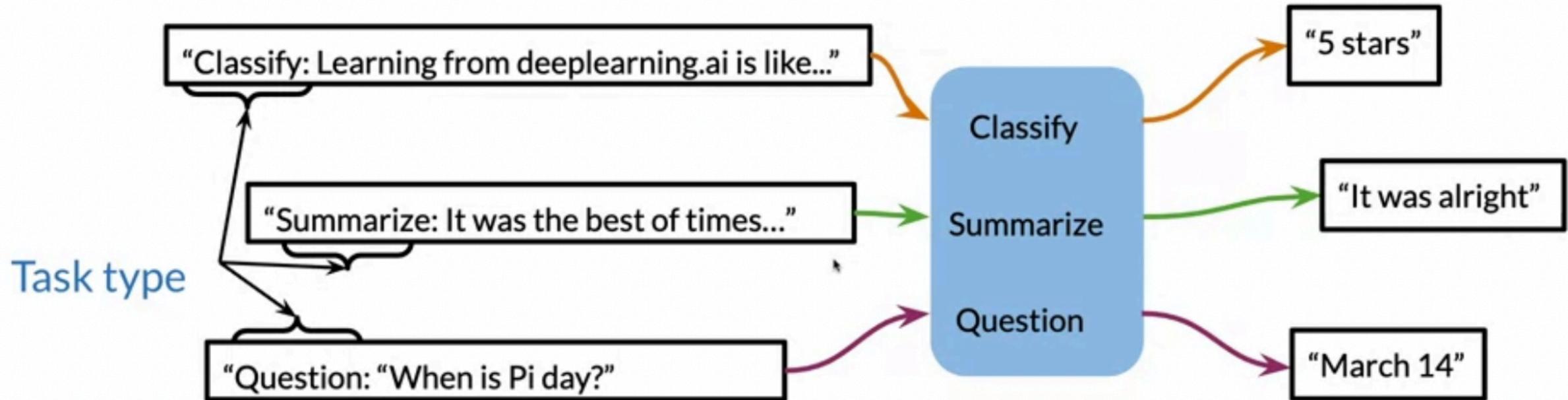
T5



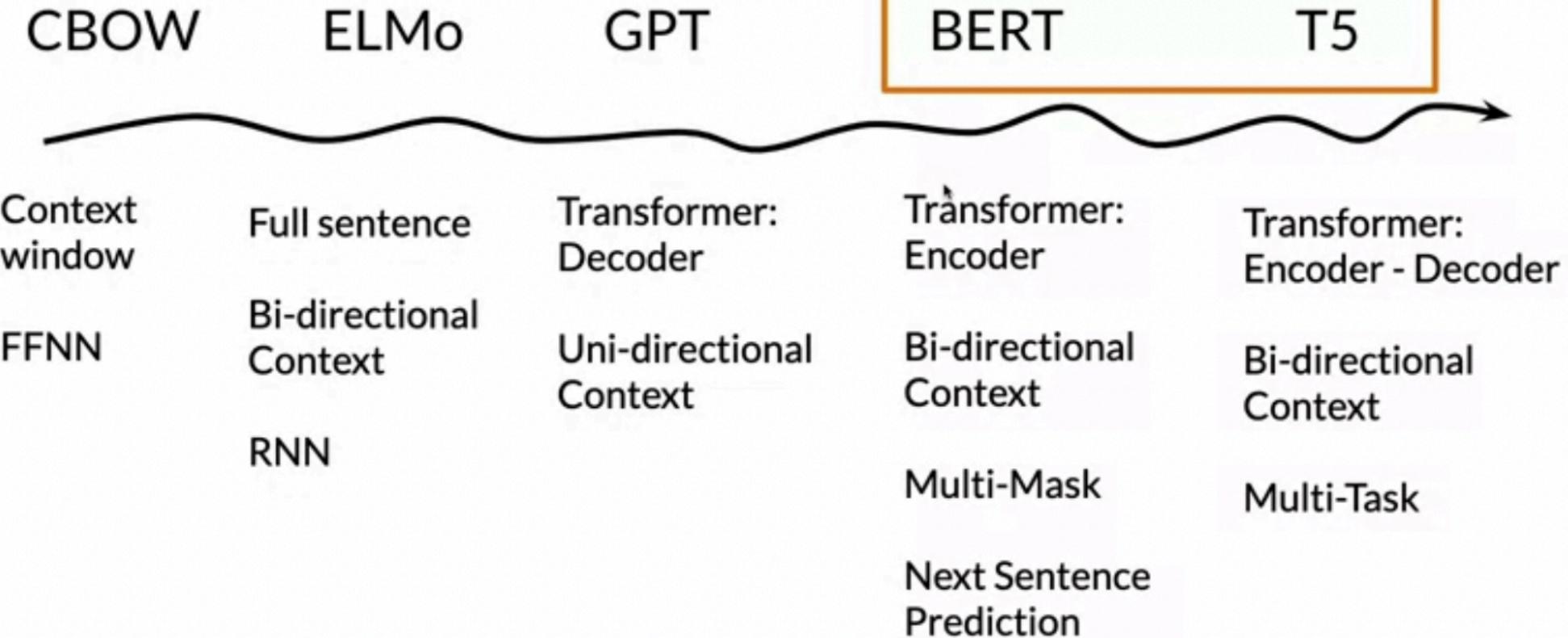
T5: Multi-task



T5: Text-to-Text



Summary

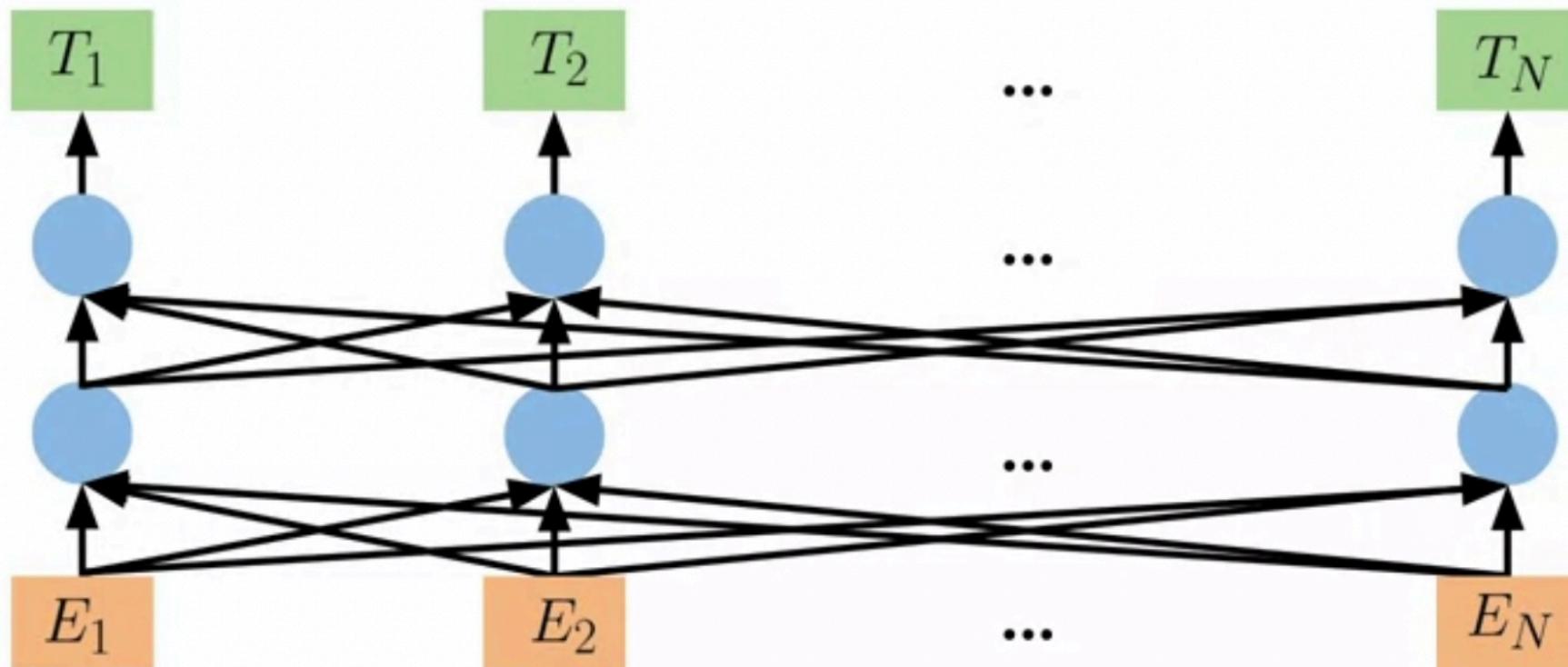


Outline

- Learn about the BERT architecture
- Understand how BERT pre-training works

BERT

- Makes use of transfer learning/pre-training:



BERT

- A multi layer bidirectional transformer
- Positional embeddings
- BERT_base:
 - 12 layers (12 transformer blocks)
 - 12 attentions heads
 - 110 million parameters

BERT pre-training

After school Lukasz does his _____ in the library.

- Masked language modeling (MLM)

BERT pre-training

After school Lukasz does his homework in the library.

After school _____ his homework in the _____ .

Summary

- Choose 15% of the tokens at random: mask them 80% of the time, replace them with a random token 10% of the time, or keep as is 10% of the time.
- There could be multiple masked spans in a sentence
- Next sentence prediction is also used when pre-training.

Outline

- Understand how BERT inputs are fed into the model
- Visualize the output
- Learn about the BERT objective

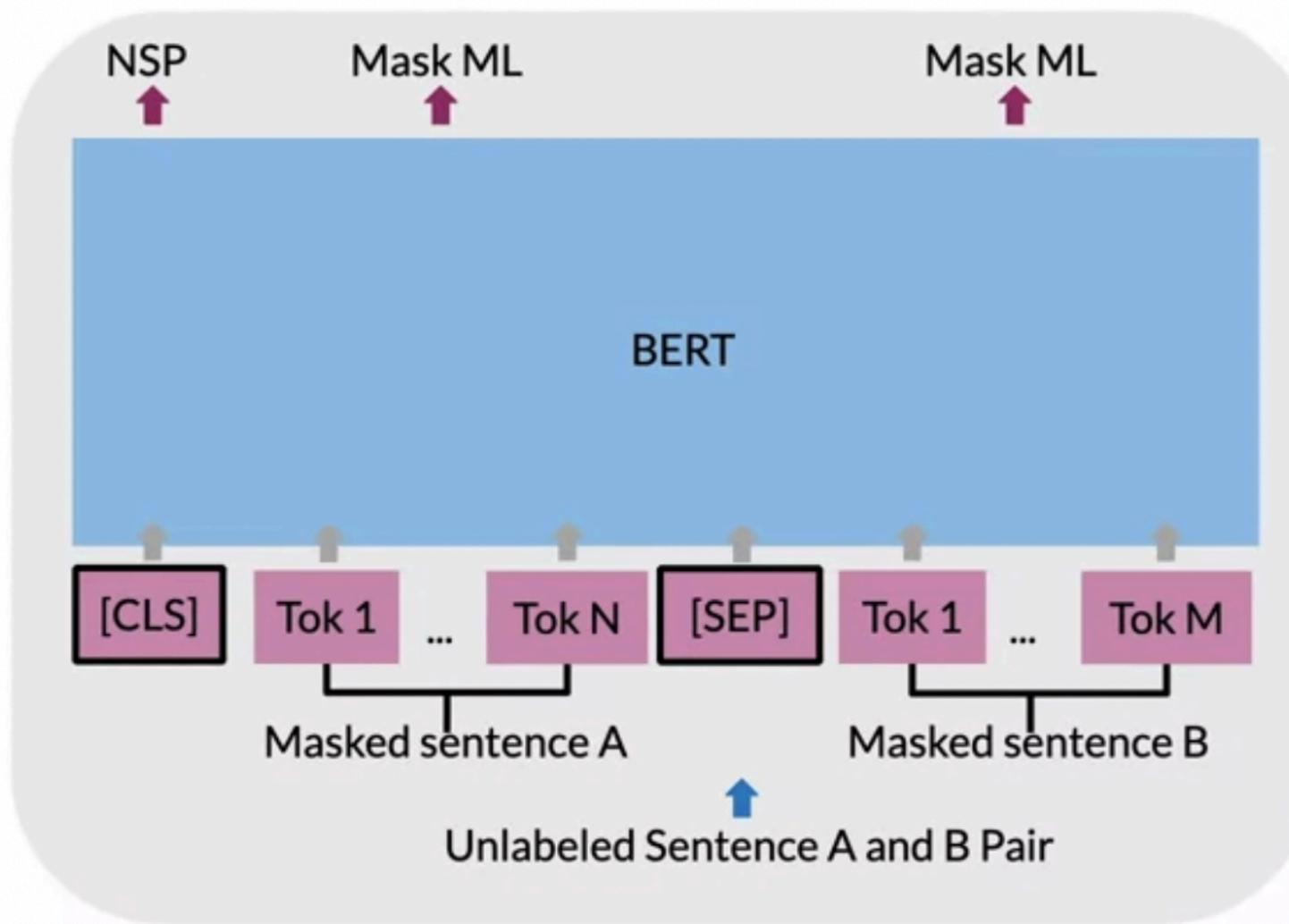
Formalizing the input

Token Embeddings	E [CLS]	E my	E dog	E is	E cute	E [SEP]	E he	E likes	E play	E ##ing	E [SEP]
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_10

Formalizing the input

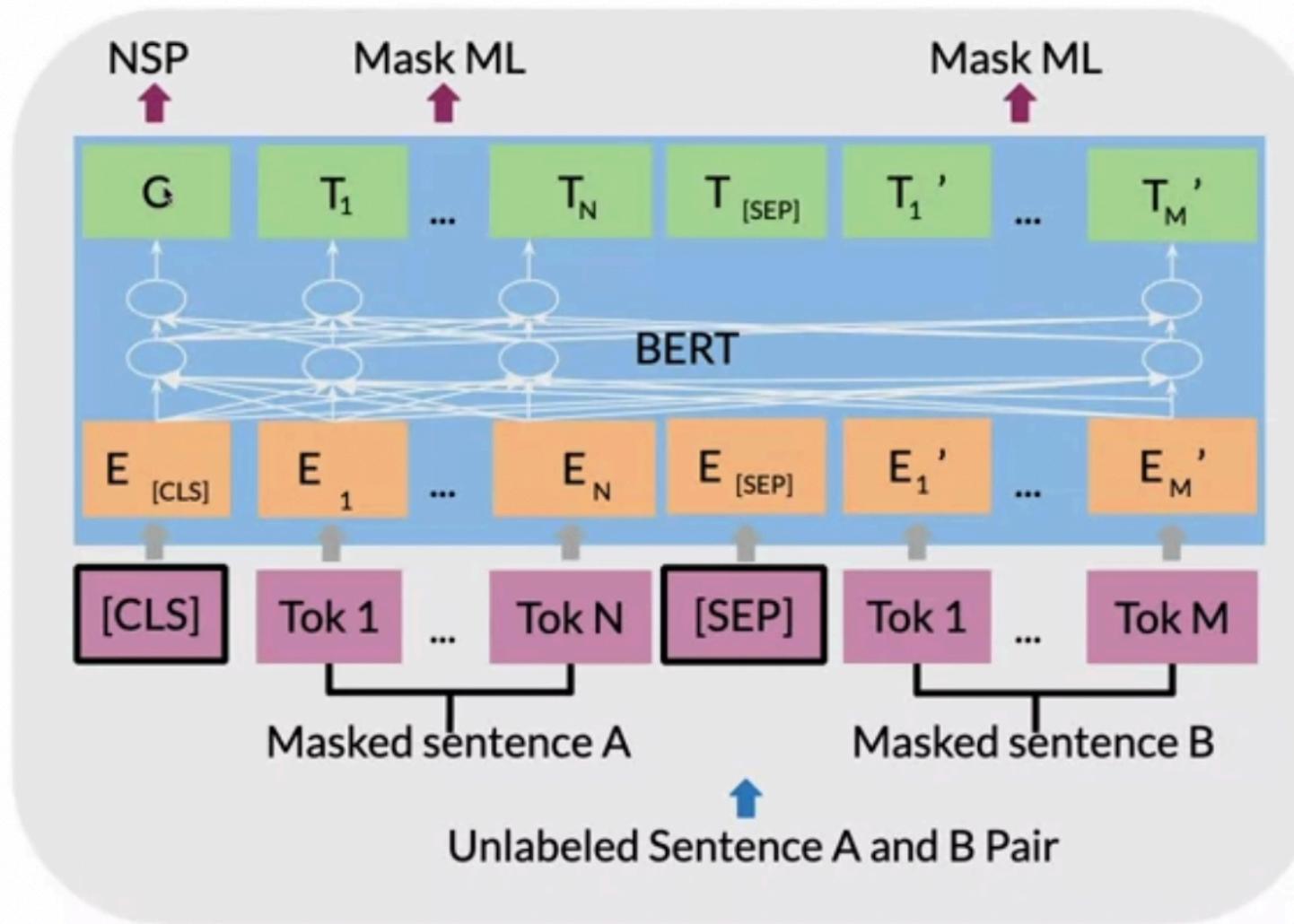
Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	E [CLS]	E my	E dog	E is	E cute	E [SEP]	E he	E likes	E play	E ##ing	E [SEP]
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B	E_B
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_10

Visualizing the output



- [CLS]: a special classification symbol added in front of every input
- [SEP]: a special separator token

Visualizing the output

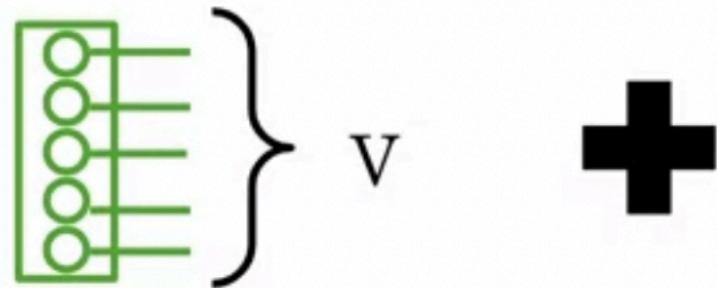


- **[CLS]**: a special classification symbol added in front of every input
- **[SEP]**: a special separator token

BERT Objective

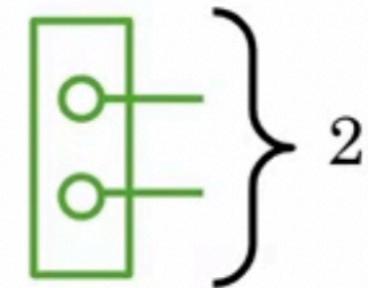
Objective 1:
Multi-Mask LM

Loss: Cross Entropy Loss



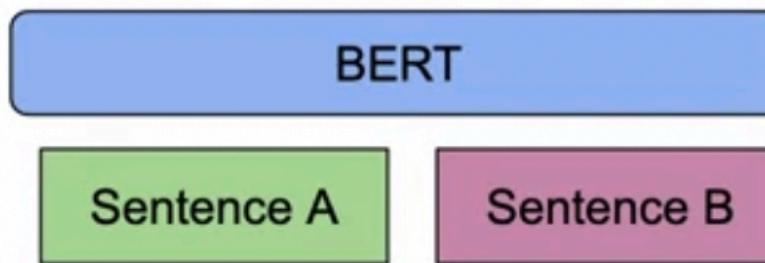
Objective 2:
Next Sentence Prediction

Loss: Binary Loss

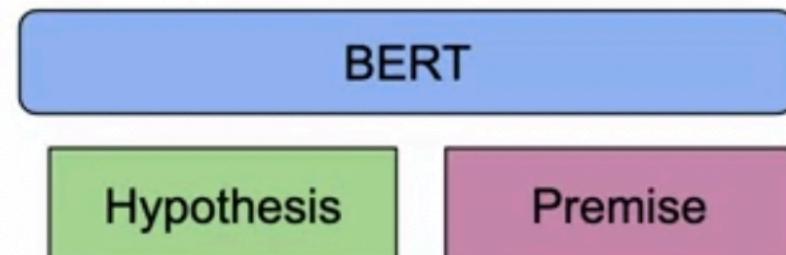


Fine-tuning BERT: Outline

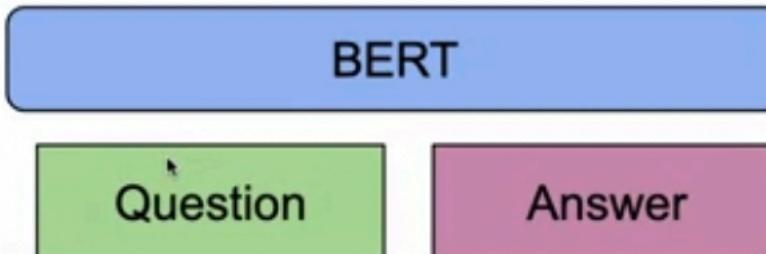
Pre-train



MNLI



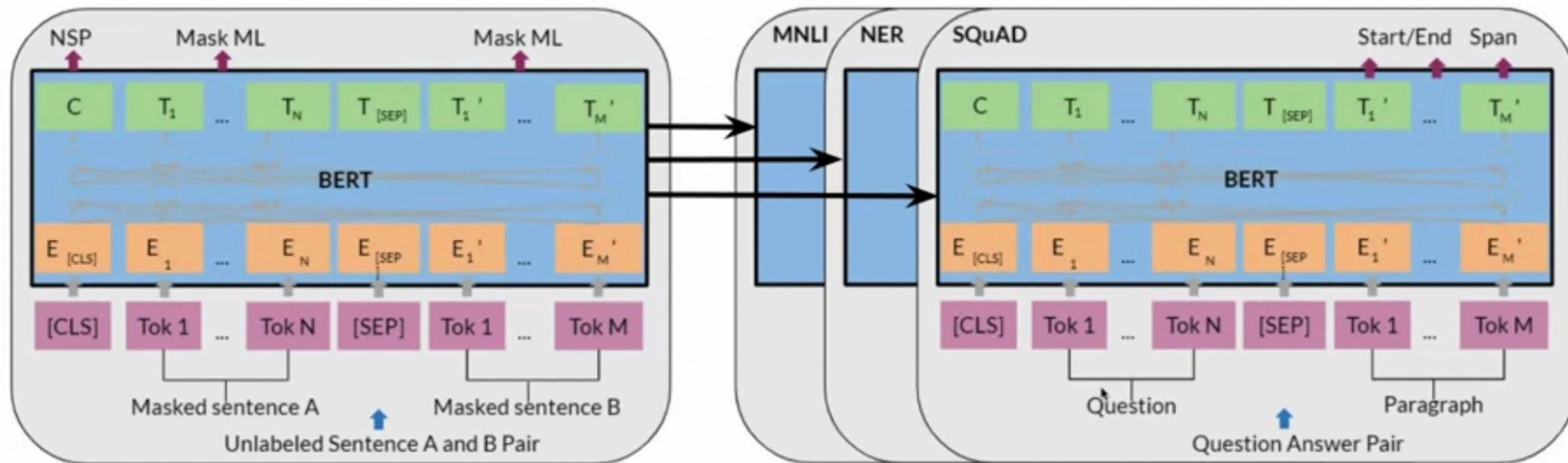
SQuAD



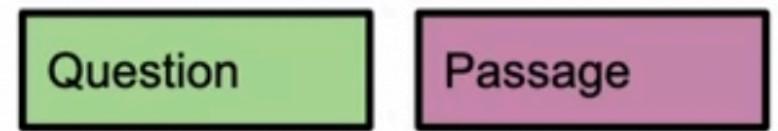
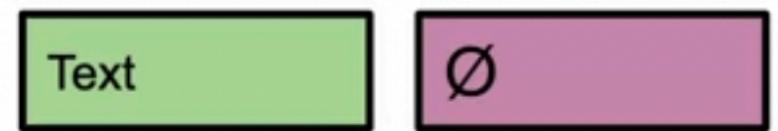
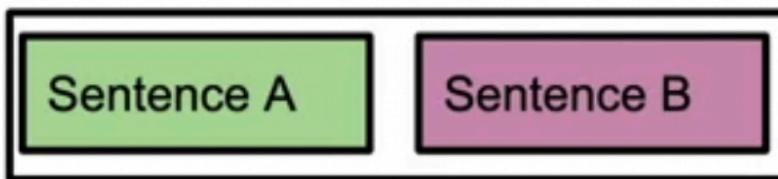
NER



Inputs



Summary



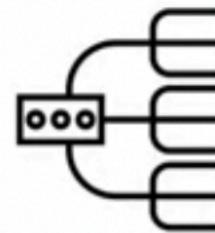
⋮

Outline

- Understand how T5 works
- Recognize the different types of attention used
- Overview of model architecture

Transformer - T5 Model

Text to Text



Classification



Question

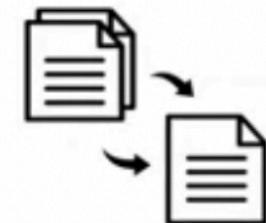


Answering (Q&A)

Machine Translation



Summarization



Sentiment



Transformer - T5 Model

Original text

Thank you for inviting me to your party last week.

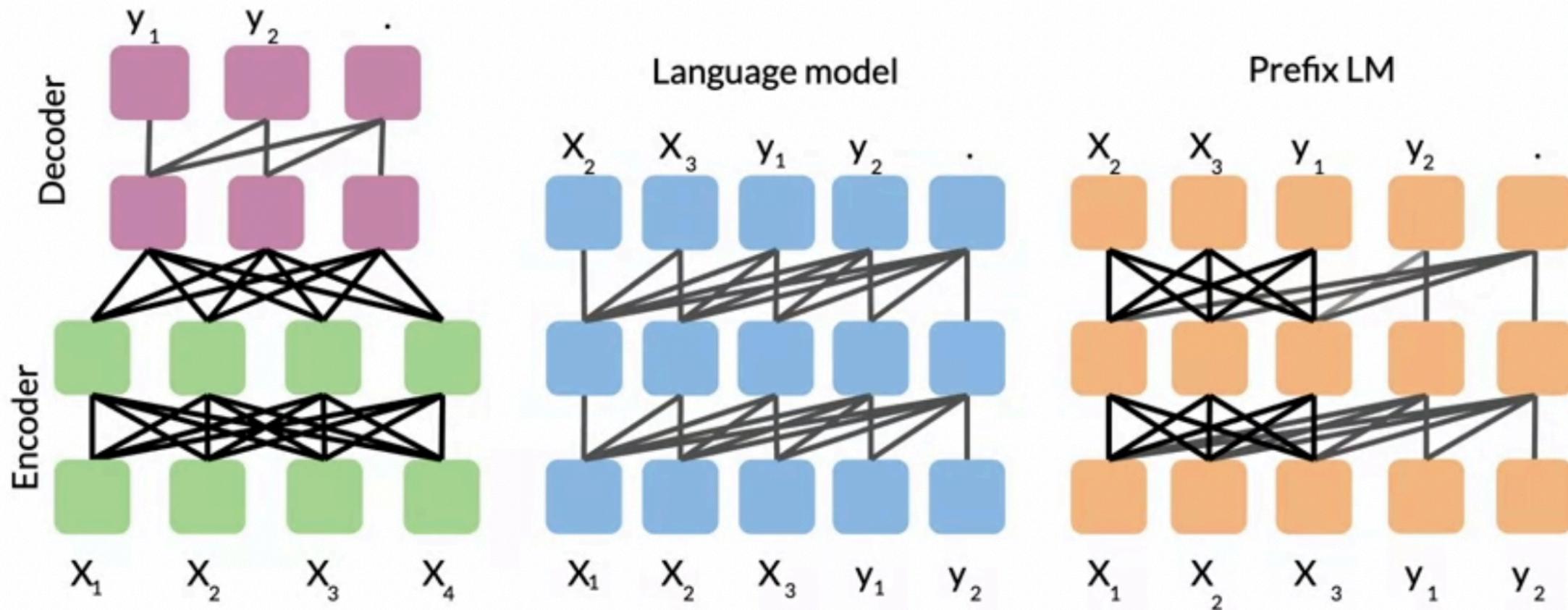
Inputs

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

Targets

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

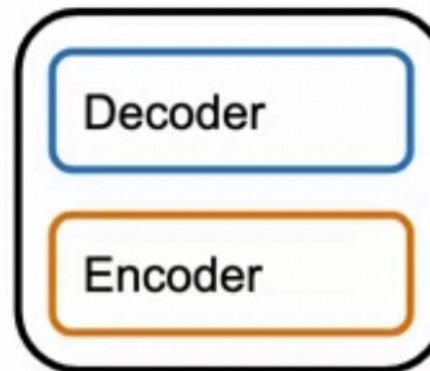
Model Architecture



©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture

- Encoder/decoder
- 12 transformer blocks each
- 220 million parameters

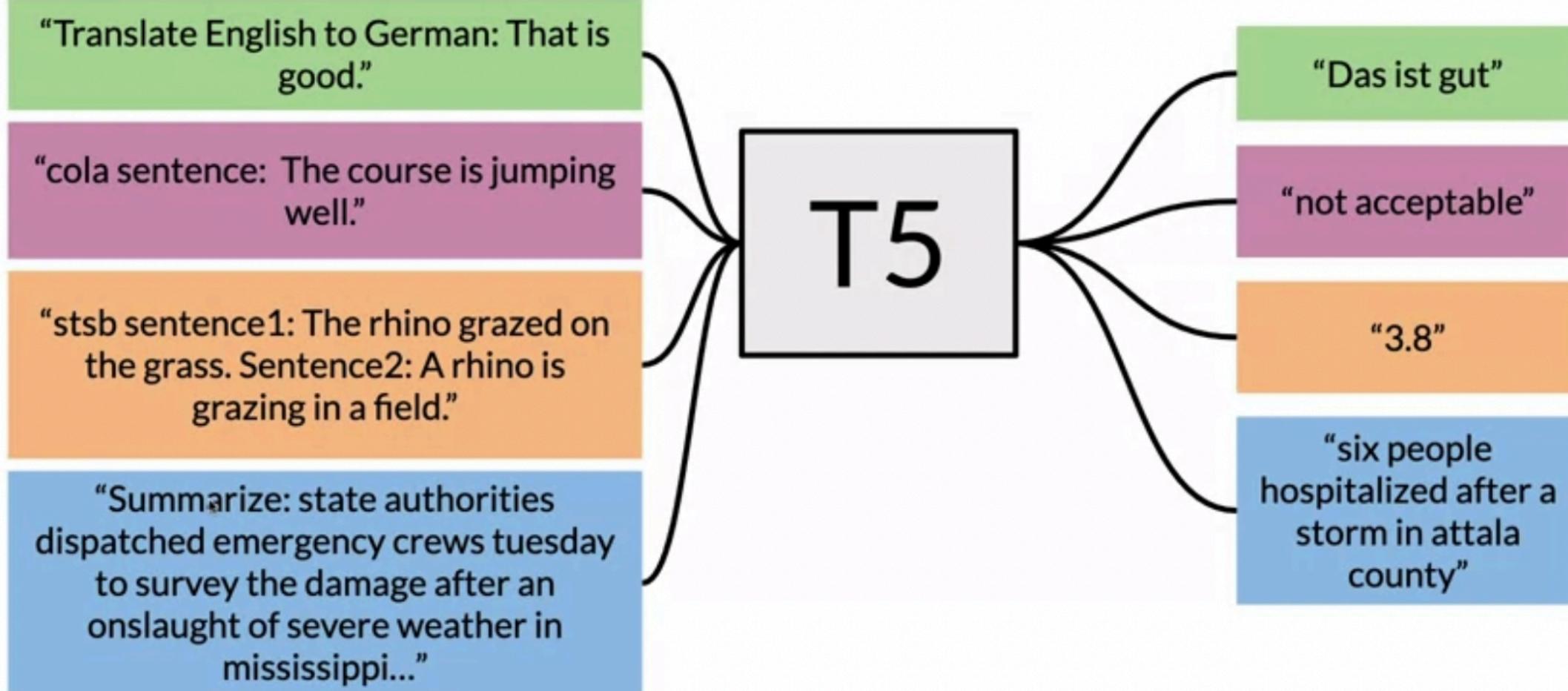


©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Summary

- Prefix LM attention
- Model architecture
- Pre-training T5 (MLM)

Multi-task training strategy



©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Input and Output Format

Machine translation:

- translate English to German: That is good.
- Predict entailment, contradiction , or neutral
 - mnli premise: I hate pigeons hypothesis: My feelings towards pigeons are filled with animosity. target: entailment
- Winograd schema
 - The city councilmen refused the demonstrators a permit because *they* feared violence

Multi-task Training Strategy

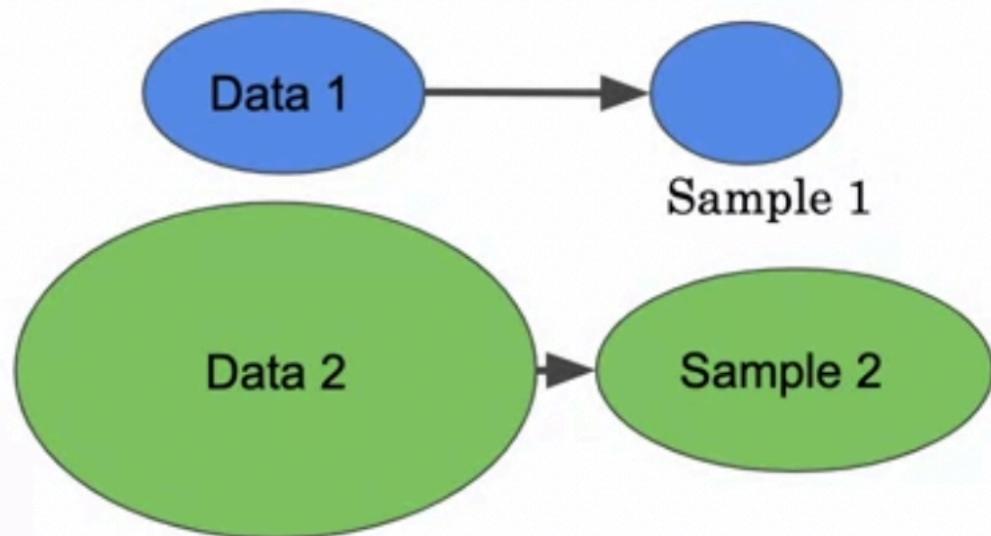
Fine-tuning method	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
* All parameters	83.28	19.24	80.88	71.36	26.98	39.82	27.65
Adapter layers, $d = 32$	80.52	15.08	79.32	60.40	13.84	17.88	15.54
Adapter layers, $d = 128$	81.51	16.62	79.47	63.03	19.83	27.50	22.63
Adapter layers, $d = 512$	81.54	17.78	79.18	64.30	23.45	33.98	25.81
Adapter layers, $d = 2048$	81.51	16.62	79.47	63.03	19.83	27.50	22.63
Gradual unfreezing	82.50	18.95	79.17	70.79	26.71	39.02	26.93

How much data from each task to train on?

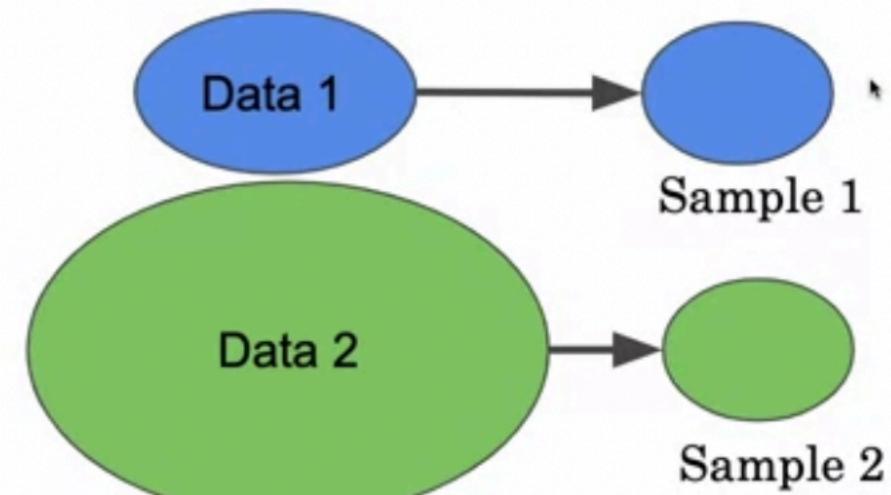
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Data Training Strategies

Examples-proportional mixing

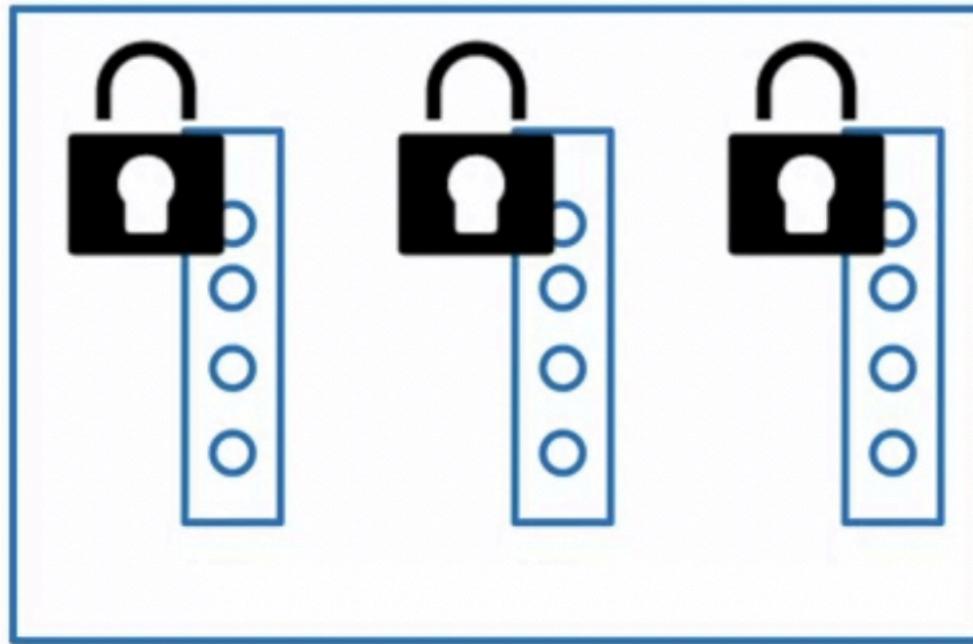


Equal mixing



Temperature-scaled mixing

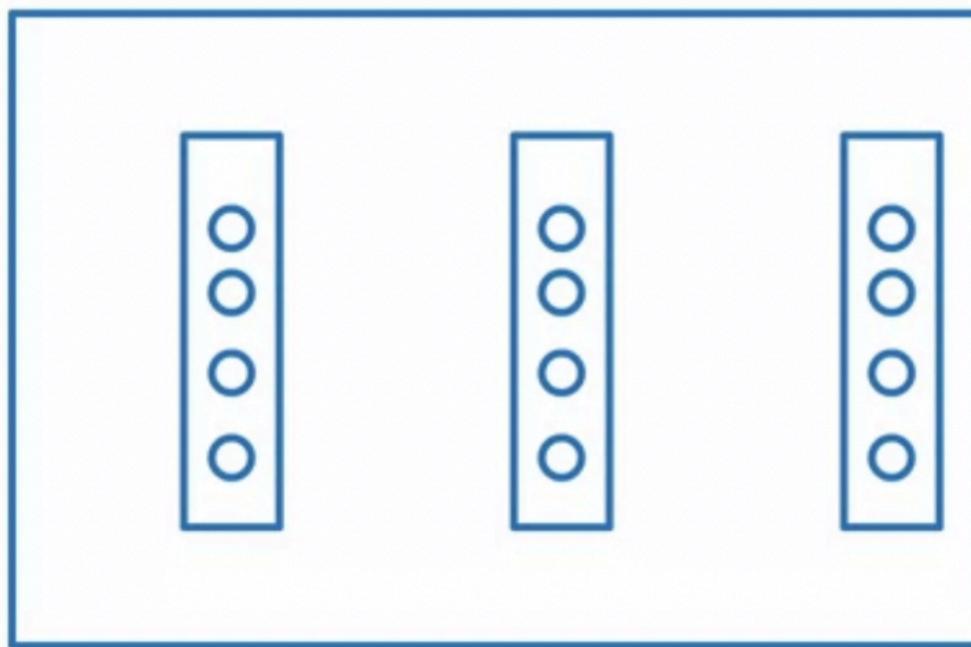
Gradual unfreezing vs. Adapter layers



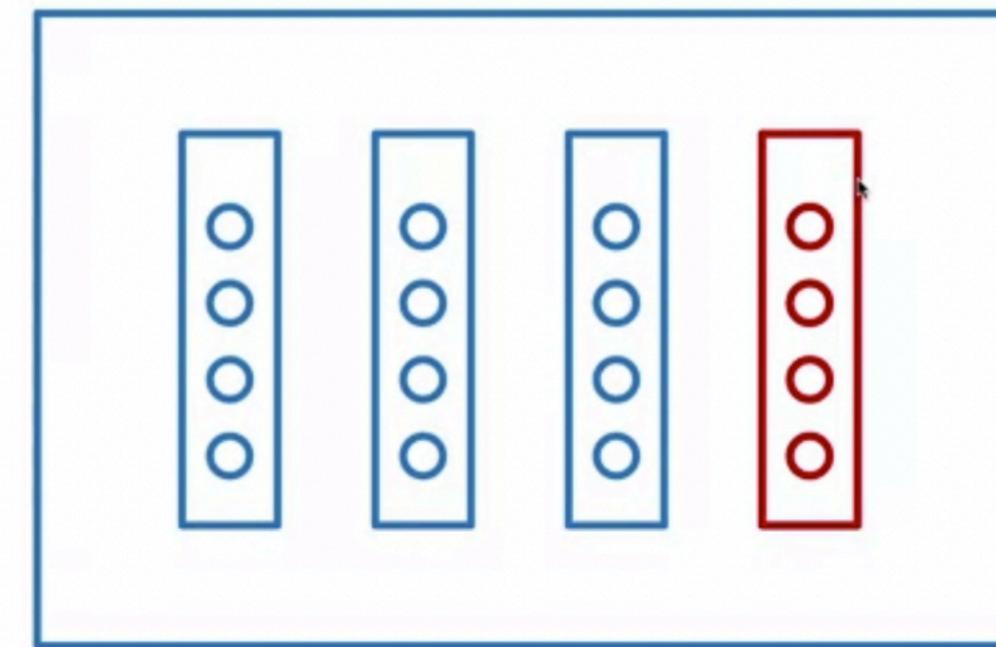
Gradual unfreezing

©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Gradual unfreezing vs. Adapter layers



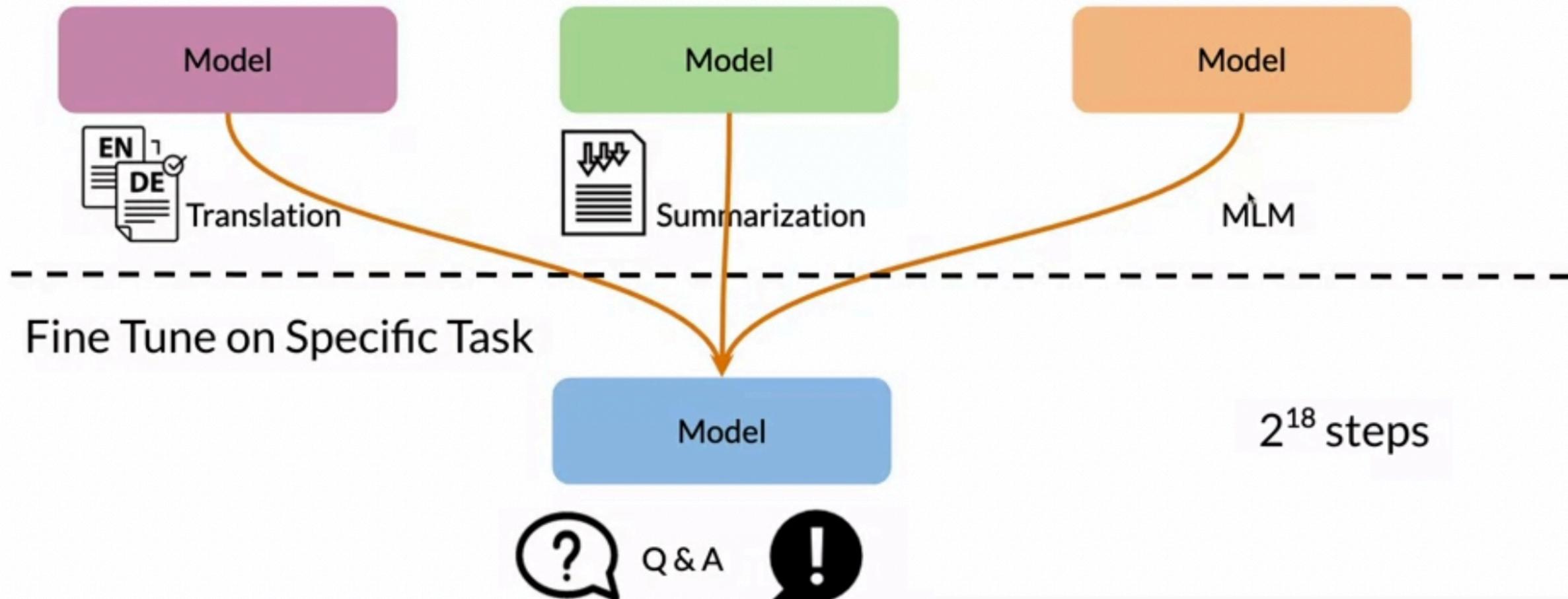
Gradual unfreezing



Adapter layers

Fine-tuning

Pre Training



General Language Understanding Evaluation

- A collection used to train, evaluate, analyze natural language understanding systems
- Datasets with different genres, and of different sizes and difficulties
- Leaderboard

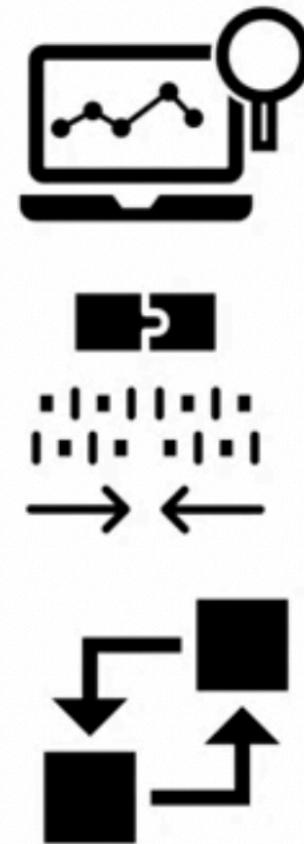
Tasks Evaluated on

- Sentence grammatical or not?
- Sentiment
- Paraphrase
- Similarity
- Questions duplicates
- Answerable
- Contradiction
- Entailment
- Winograd (co-ref)

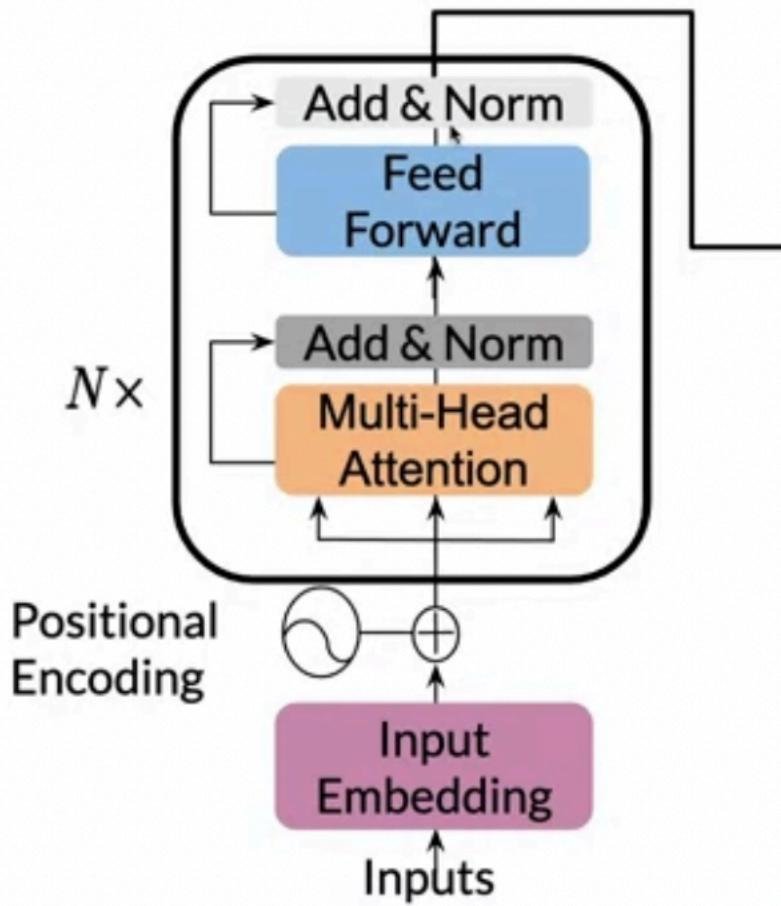


General Language Understanding Evaluation

- Drive research
- Model agnostic
- Makes use of transfer learning



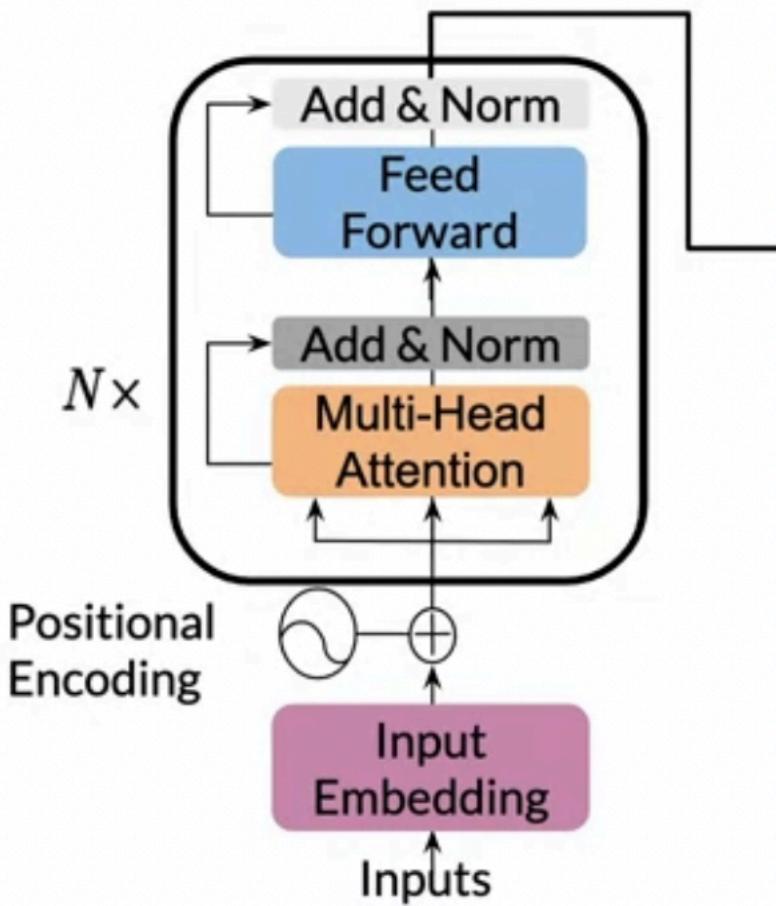
Transformer encoder



Feedforward:

```
[  
    LayerNorm,  
    dense,  
    activation,  
    dropout_middle,  
    dense,  
    dropout_final  
]
```

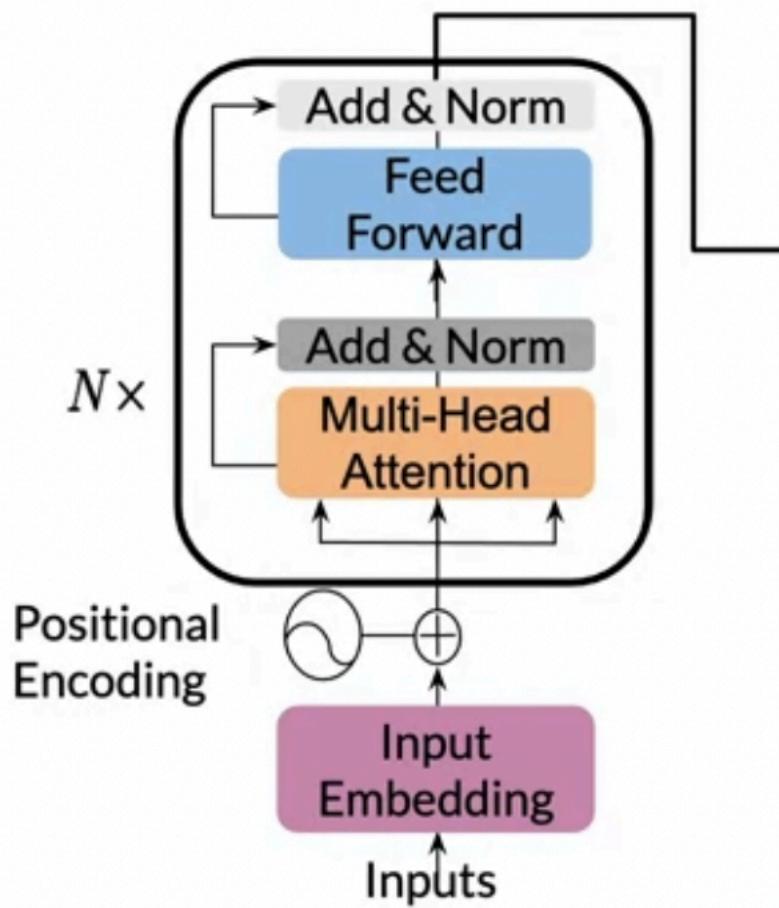
Transformer encoder



Encoder block:

```
[  
    Residual(  
        LayerNorm,  
        attention,  
        dropout_,  
    ),  
    Residual(  
        feed_forward,  
    ),  
]
```

Transformer encoder



Feedforward:

```
[ LayerNorm,  
  dense,  
  activation,  
  dropout_middle,  
  dense,  
  dropout_final ]
```

Encoder block:

```
[ Residual(  
  LayerNorm,  
  attention,  
  dropout_,  
 ),  
 Residual(  
  feed_forward,  
 ) ]
```

Data examples

Question: What percentage of the French population today is non - European ?

Context: Since the end of the Second World War , France has become an ethnically diverse country . Today , **approximately five percent** of the French population is non - European and non - white . This does not approach the number of non - white citizens in the United States (roughly 28 – 37 % , depending on how Latinos are classified ; see Demographics of the United States) . Nevertheless , it amounts to at least three million people , and has forced the issues of ethnic diversity onto the French policy agenda . France has developed an approach to dealing with ethnic problems that stands in contrast to that of many advanced , industrialized countries . Unlike the United States , Britain , or even the Netherlands , France maintains a " color - blind " model of public policy . This means that it targets virtually no policies directly at racial or ethnic groups . Instead , it uses geographic or class criteria to address issues of social inequalities . It has , however , developed an extensive anti - racist policy repertoire since the early 1970s . Until recently , French policies focused primarily on issues of hate speech — going much further than their American counterparts — and relatively less on issues of discrimination in jobs , housing , and in provision of goods and services .

Target: **Approximately five percent**

Implementing Q&A with T5

- Load a pre-trained model
- Process data to get the required inputs and outputs: "question: Q context: C" as input and "A" as target
- Fine tune your model on the new task and input
- Predict using your own model

