Fine-tuning an LLM

What does pre-training mean?

What does fine-tuning mean?

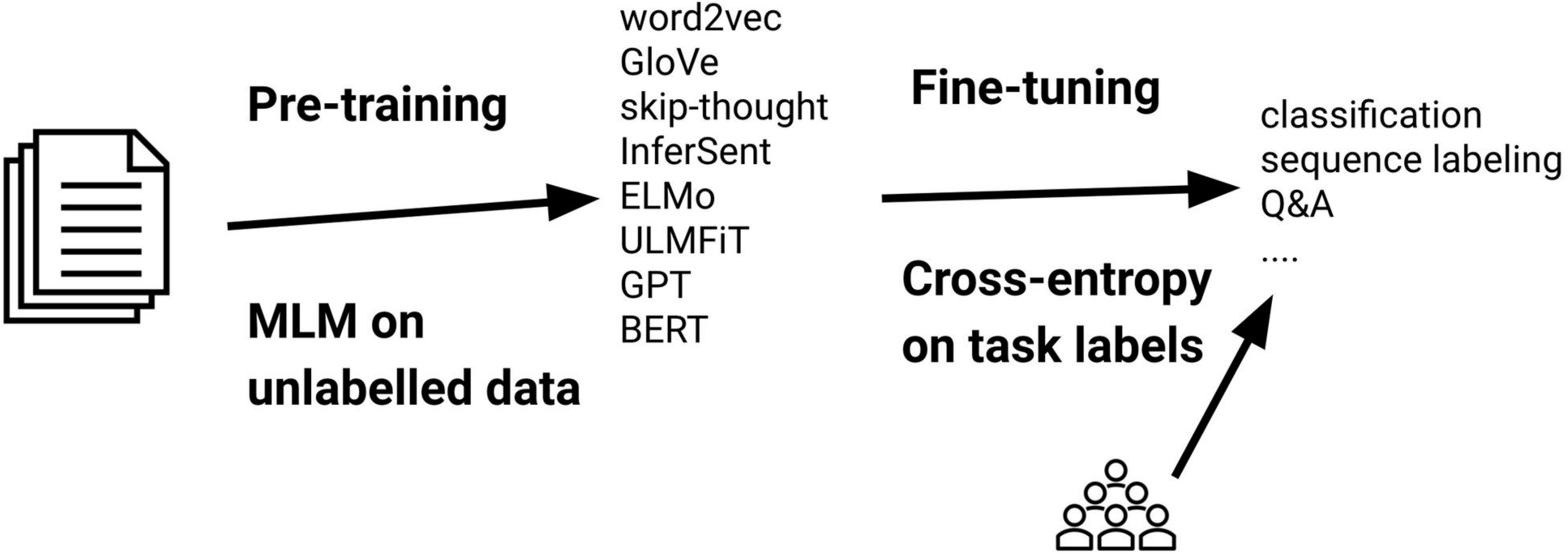
How many parameters does BERT have?

Is BERT much smaller than GPT?

How was the BERT model pre-trained?

How does MLM pre-training objective work?

How does NSP pre-training objective work?



Pre-training

MLM on unlabelled data

word2vec
GloVe
skip-thought
InferSent
ELMo
ULMFiT
GPT
BERT

Fine-tuning

Cross-entropy on task labels

classification
sequence labeling
Q&A
....

Pre-training

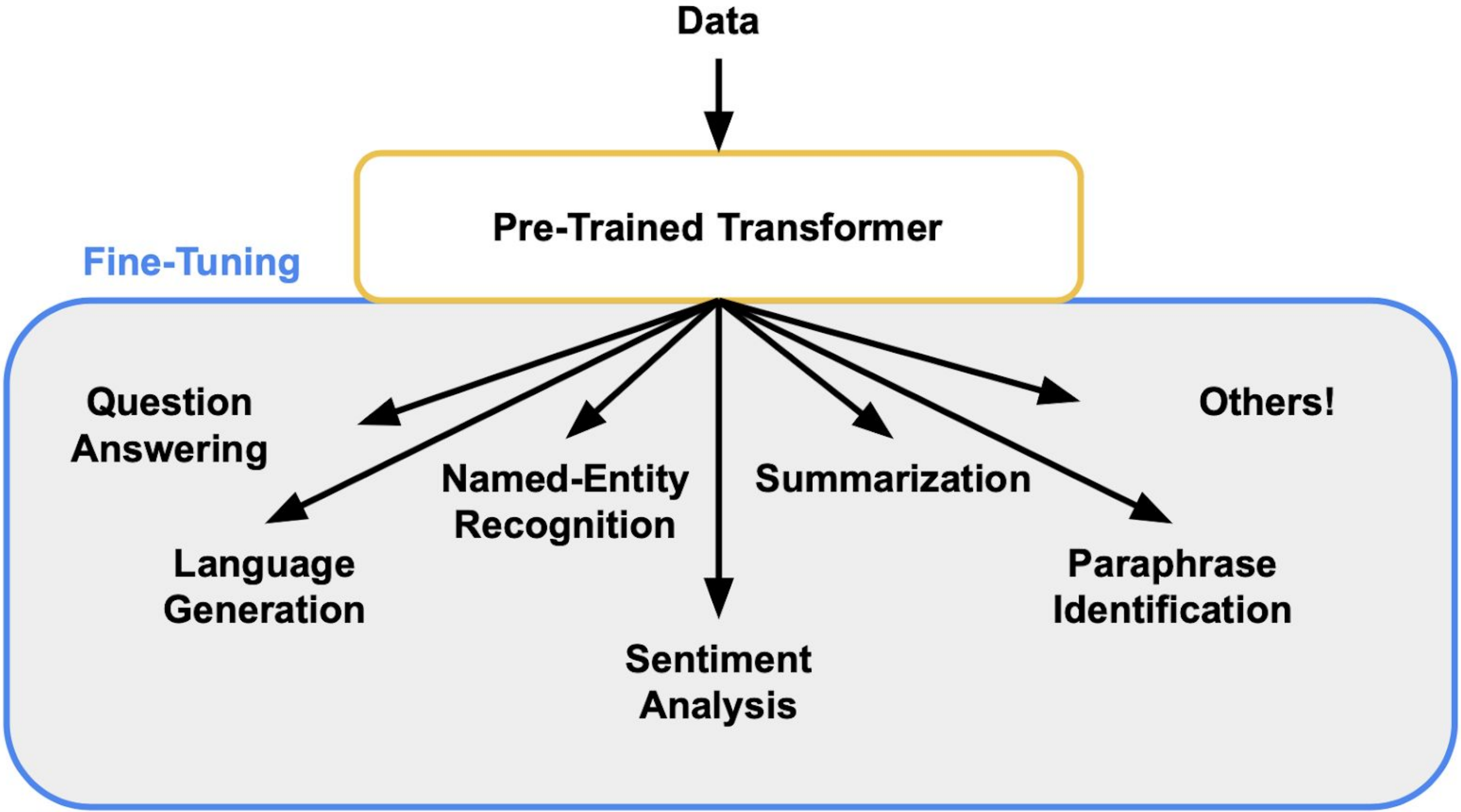
is like a child learning to
read and write his/her mother tongue.

Fine Tuning

is like a student learning to use language to
perform complex tasks in high school and college.

In-Context Learning

is like a working professional trying to
figure out his/her manager’s instructions
Zero Shot vs Few Shot



Data

Pre-Trained Transformer

Fine-Tuning

Question Answering

Language Generation

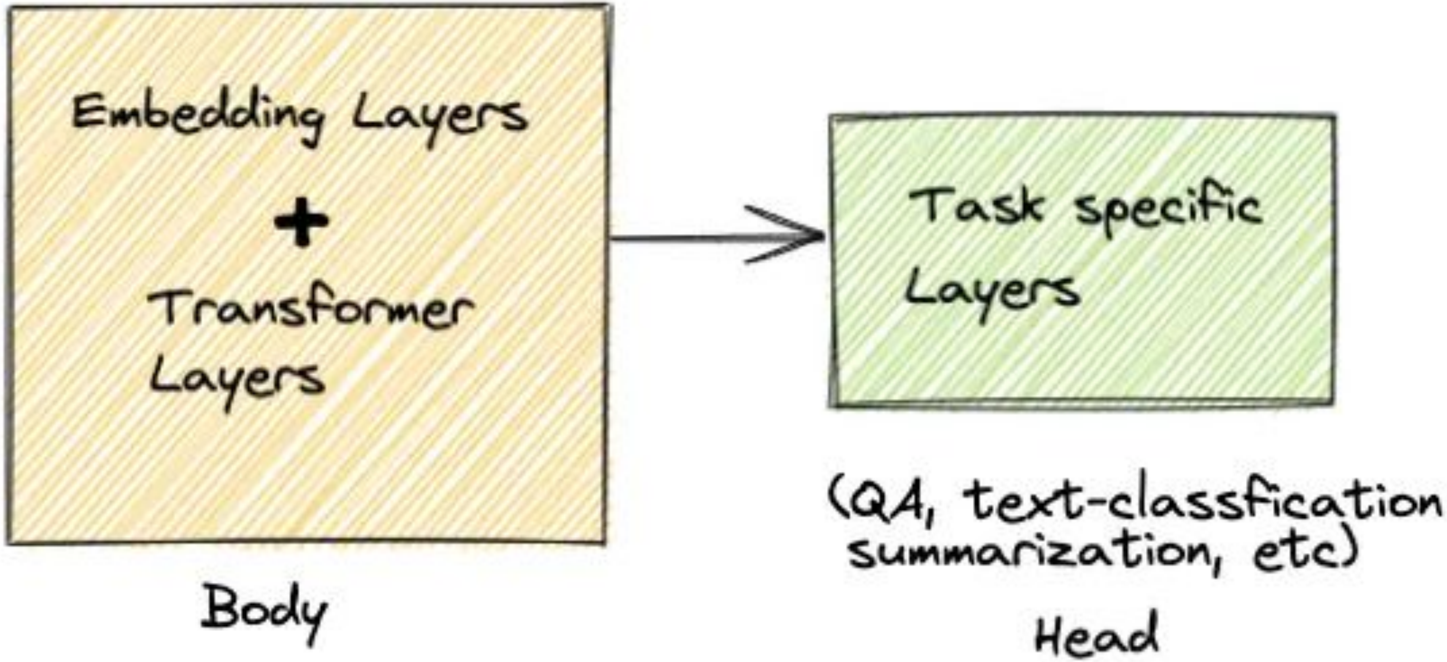
Named-Entity Recognition

Sentiment Analysis

Summarization

Paraphrase Identification

Others!



Embedding Layers
+
Transformer Layers

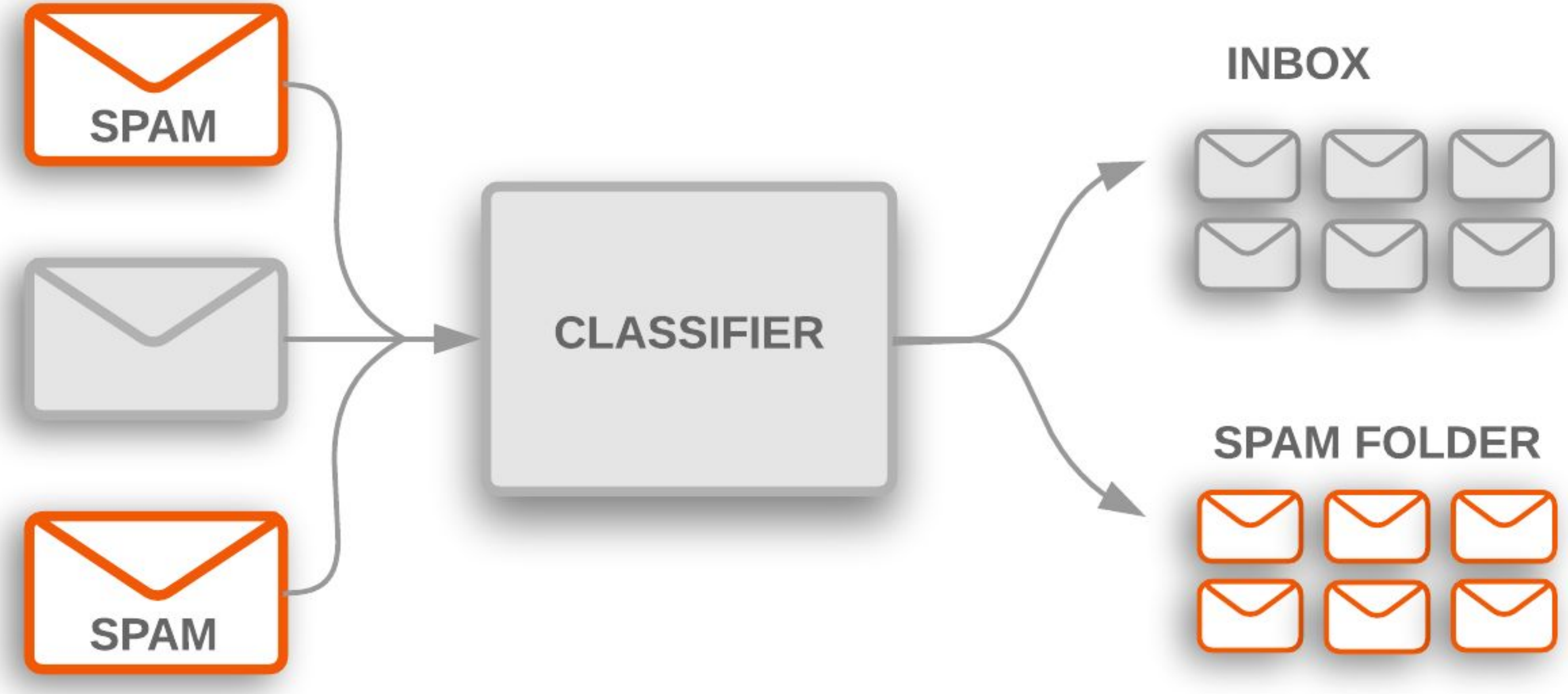
Body

Task specific Layers

(QA, text-classfication summarization, etc)

Head

TEXT CLASSIFICATION


SPAM
SPAM
CLASSIFIER
INBOX
SPAM FOLDER

My experience so far has been **fantastic!**

POSITIVE

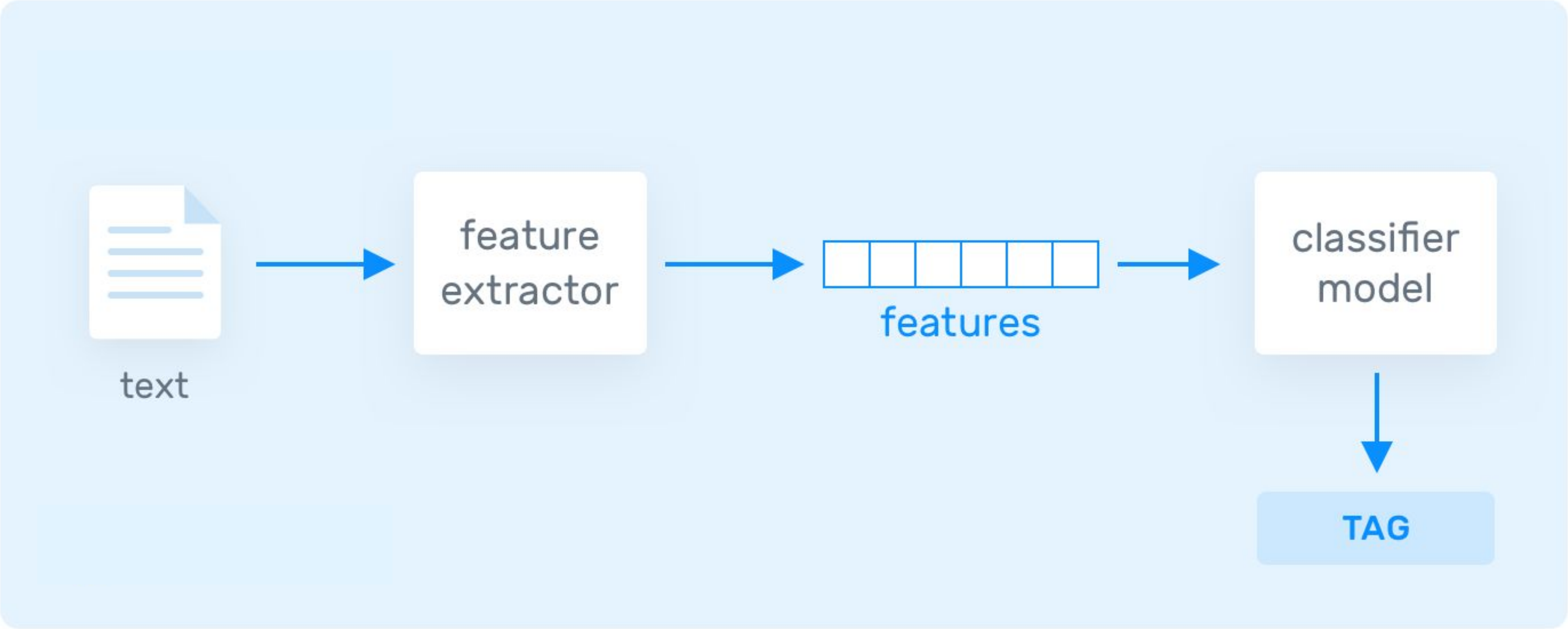
The product is **okay I guess.**

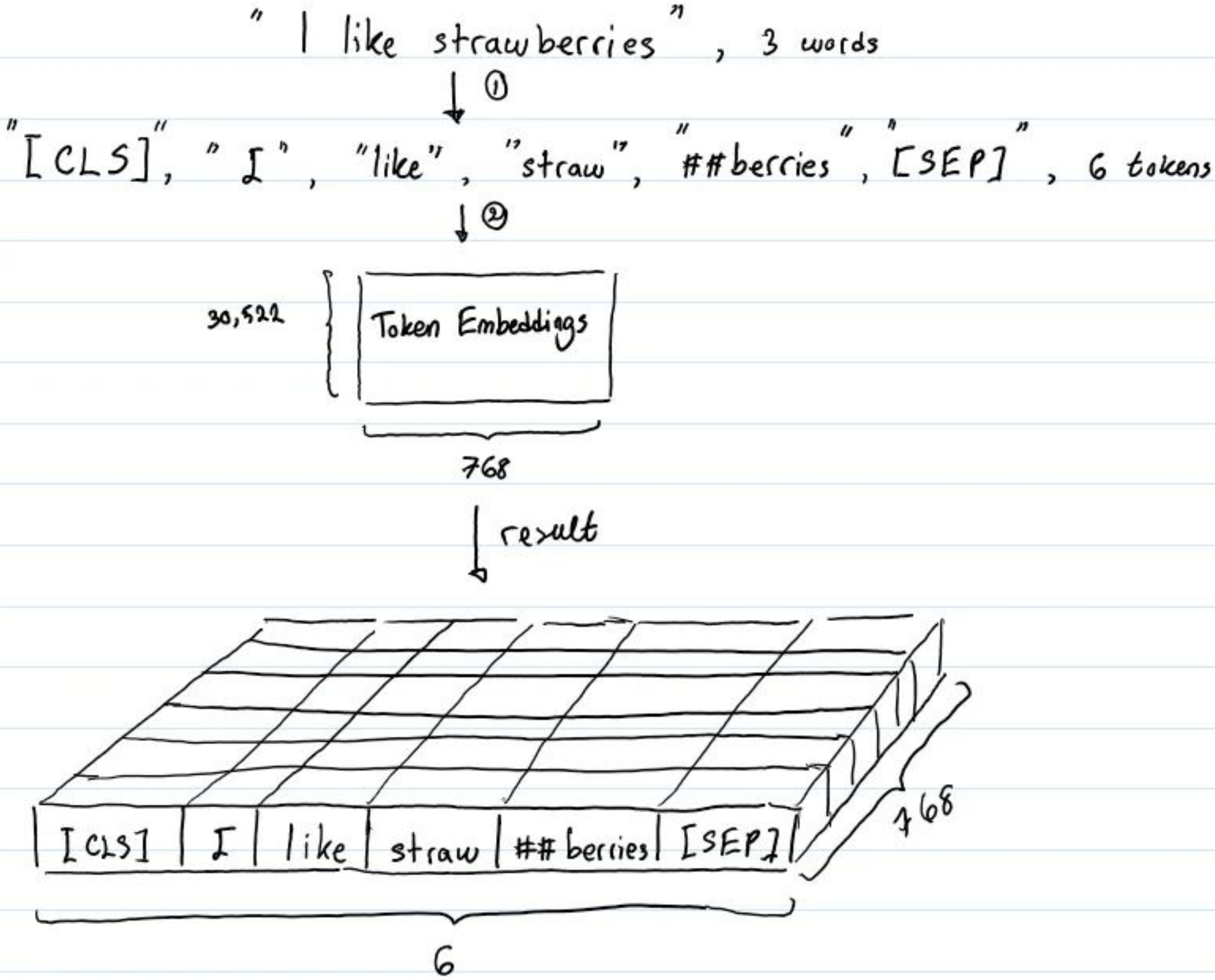
NEUTRAL

Your support team is **useless.**

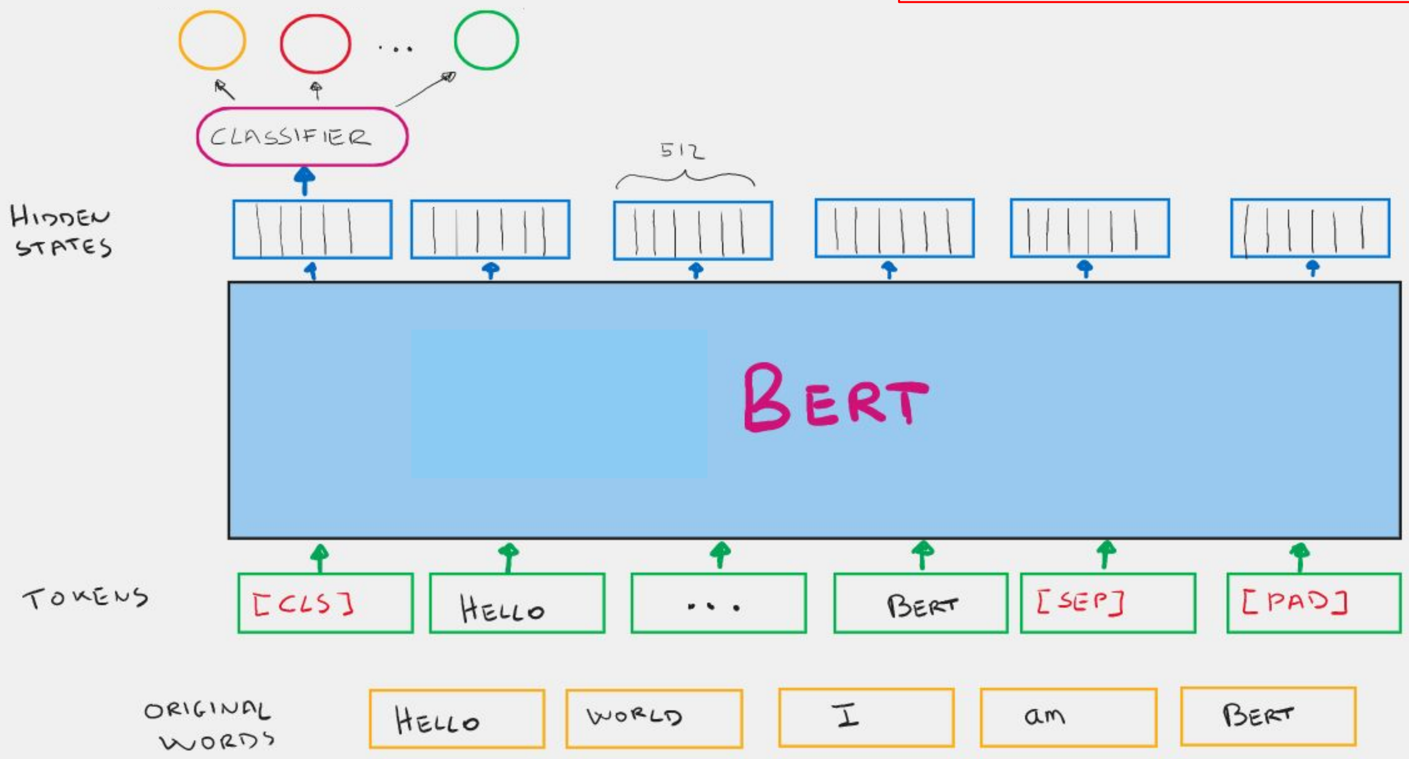
NEGATIVE

Classical NLP Approach

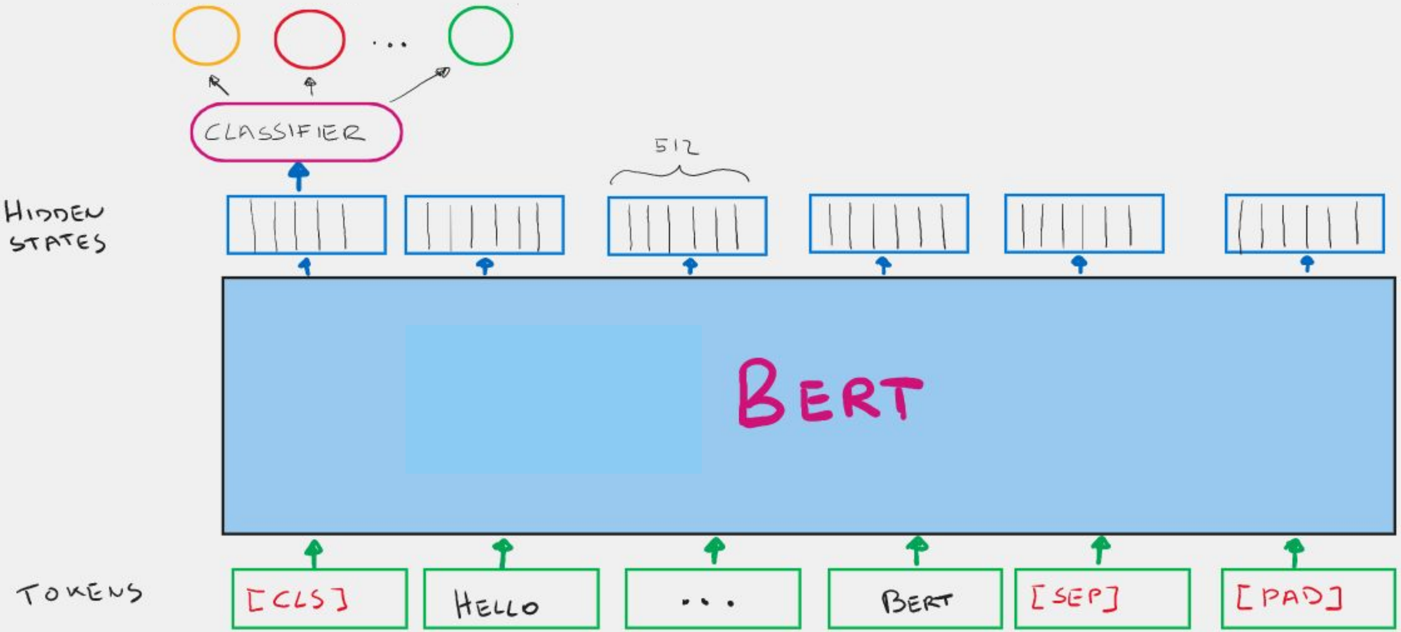



" I like strawberries" , 3 words
①
"[CLS]", " I ", "like", "straw", "##berries", "[SEP]" , 6 tokens
②
30,522
Token Embeddings
768
result
[CLS]
I
like
straw
##berries
[SEP]
768
6

Requires Fine Tuning

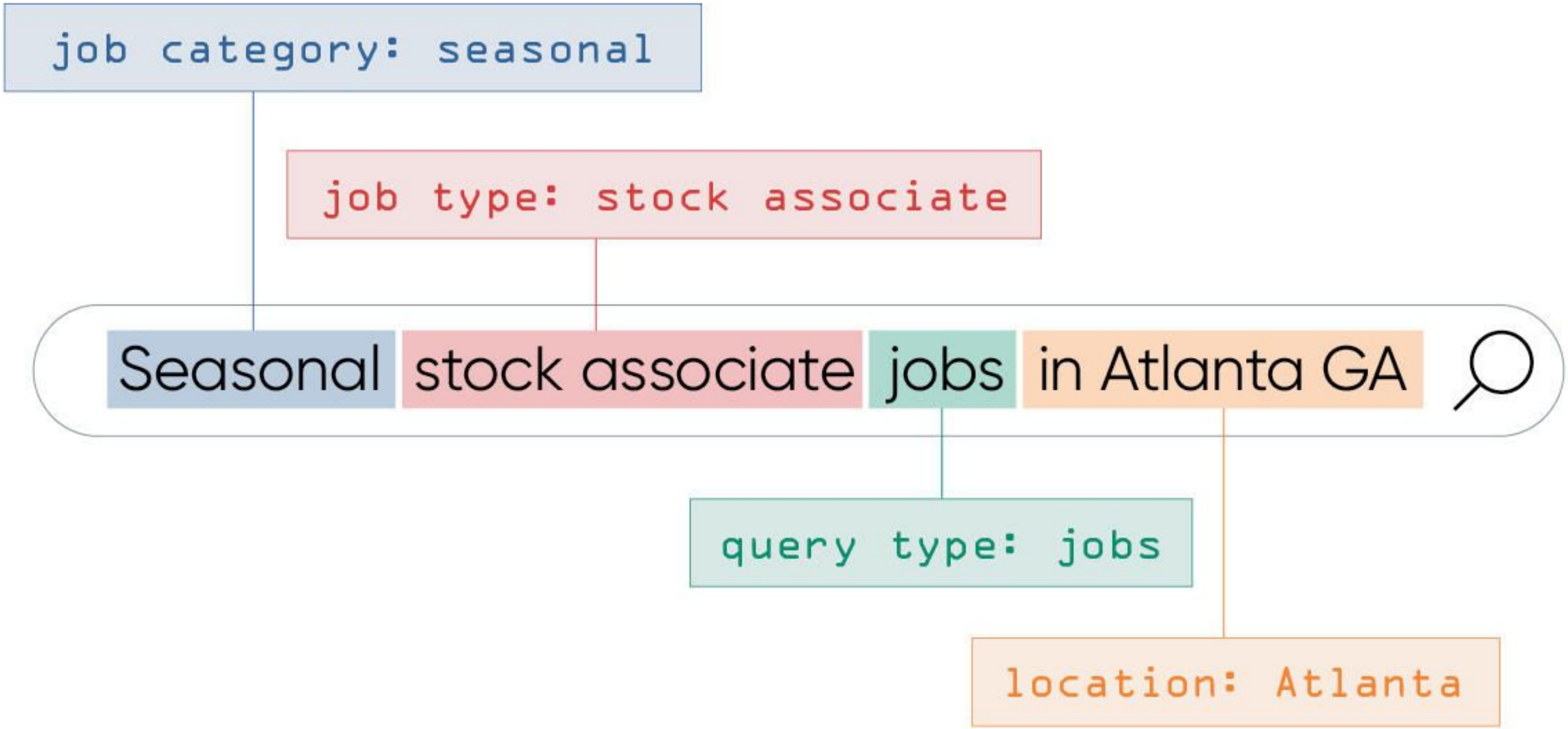


Requires Fine Tuning



Is only the classifier layer on top trained or
are the BERT parameters also updated during fine-tuning?

NAMED ENTITY RECOGNITION



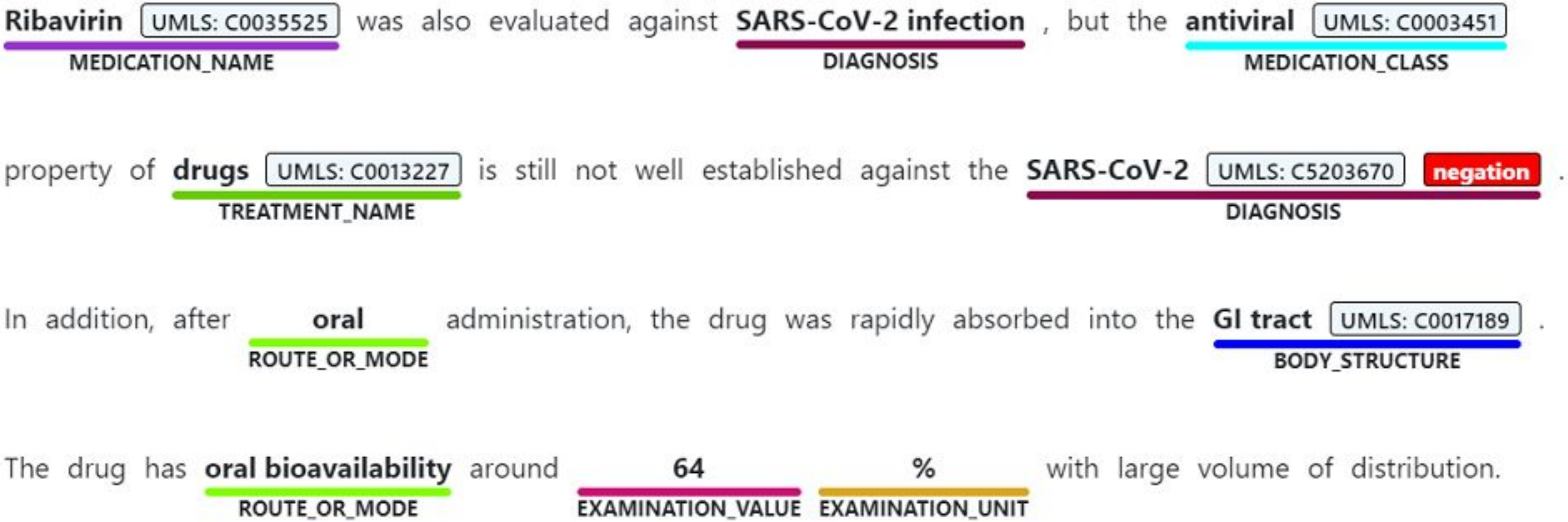
job category: seasonal

job type: stock associate

Seasonal stock associate jobs in Atlanta GA

query type: jobs

location: Atlanta



Ribavirin was also evaluated against SARS-CoV-2 infection , but the antiviral

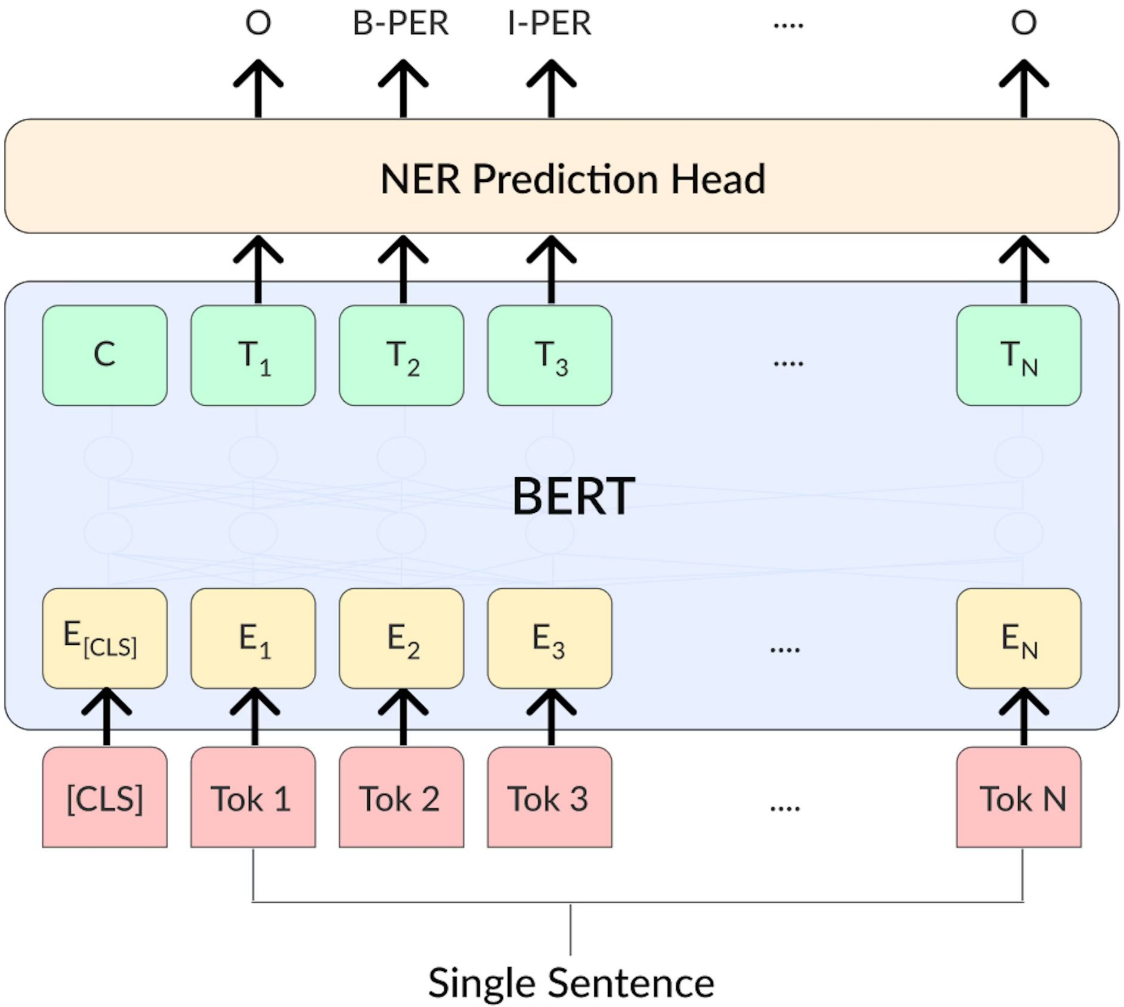
property of drugs is still not well established against the SARS-CoV-2 .

In addition, after oral administration, the drug was rapidly absorbed into the GI tract .

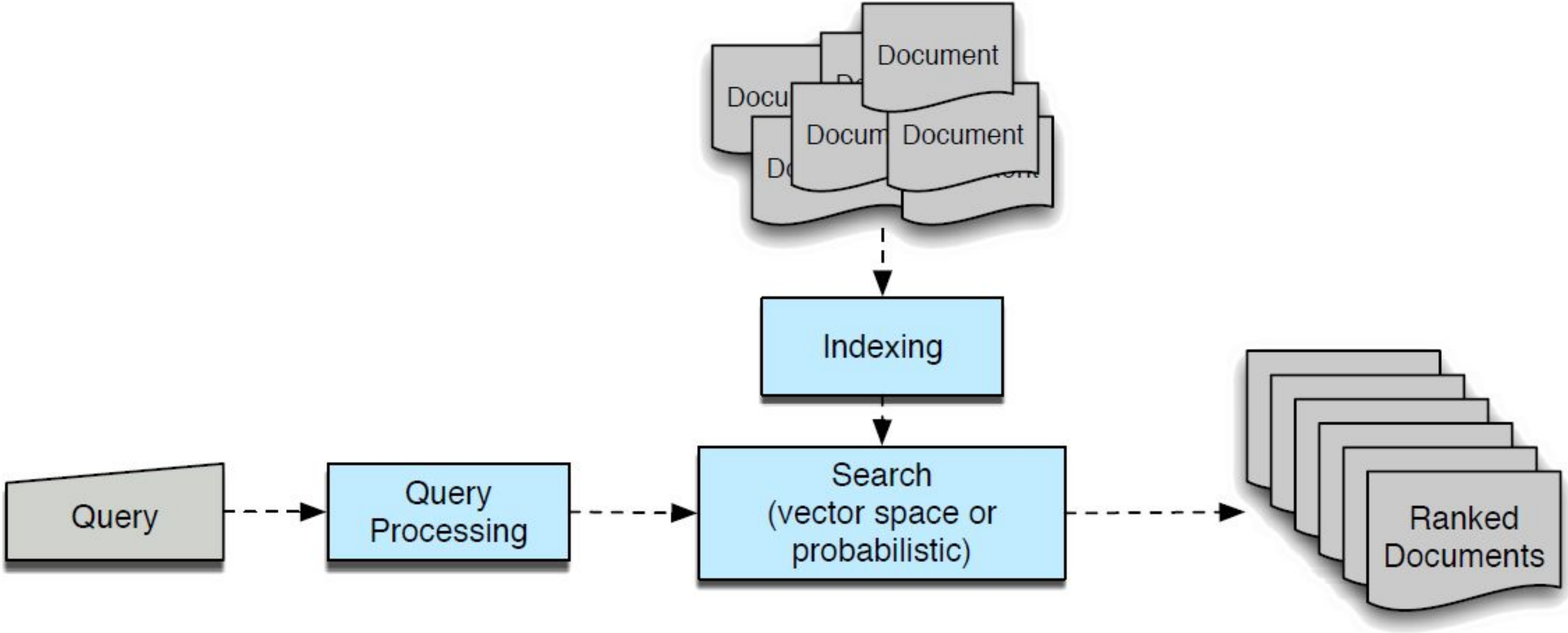
The drug has oral bioavailability around 64 % with large volume of distribution.

BERT NER : The B-I-O Notation

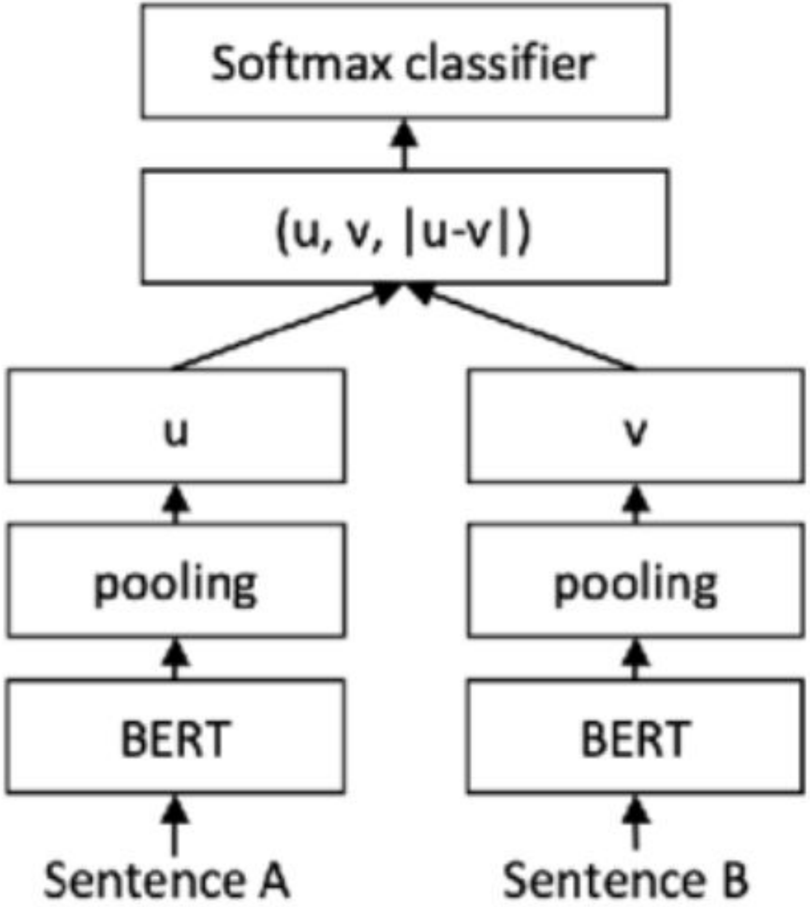
Yesterday	,	Rohan	Sharma	traveled	to	Mumbai	.
O	O	B-PER	I-PER	O	O	B-LOC	O

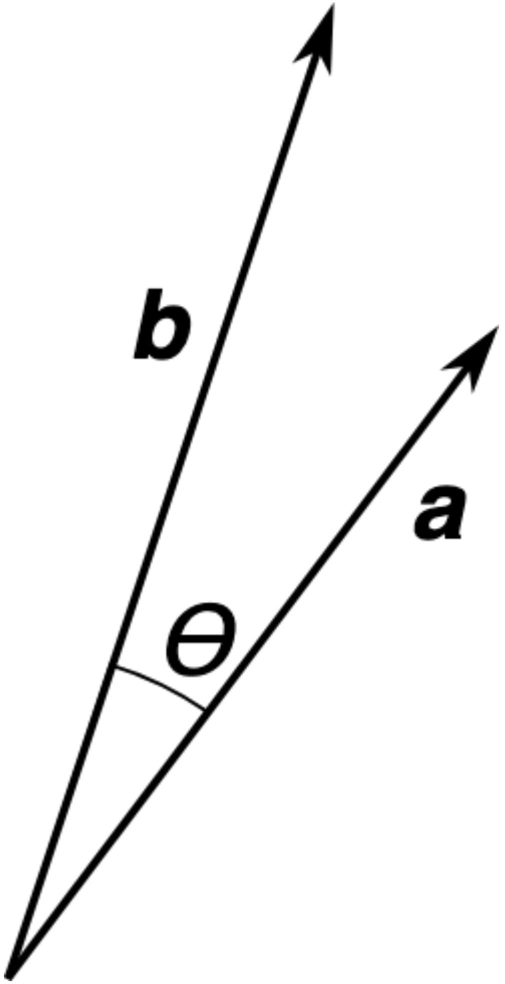

O
B-PER
I-PER
....
O
NER Prediction Head
C
T_1
T_2
T_3
....
T_N
BERT
$E_{[CLS]}$
E_1
E_2
E_3
....
E_N
[CLS]
Tok 1
Tok 2
Tok 3
....
Tok N
Single Sentence

INFORMATION RETRIEVAL


Document
Docu
Do
Docum
Document
Do
Indexing
Query
Query Processing
Search (vector space or probabilistic)
Ranked Documents

SBERT Fine-Tuning





- The query has a vector representation using embeddings

- Documents in the database stored as embeddings

- **Brute Force Approach:**
  Do a dot product of the query vector with the embeddings of all the documents, and choose the one that gives the closest match

- **Hierarchical Navigable Small World (HNSW):**
  Create a layered graph structure of the document embedding vectors so that the search process is made much faster

QUESTION ANSWERING

Passage Sentence

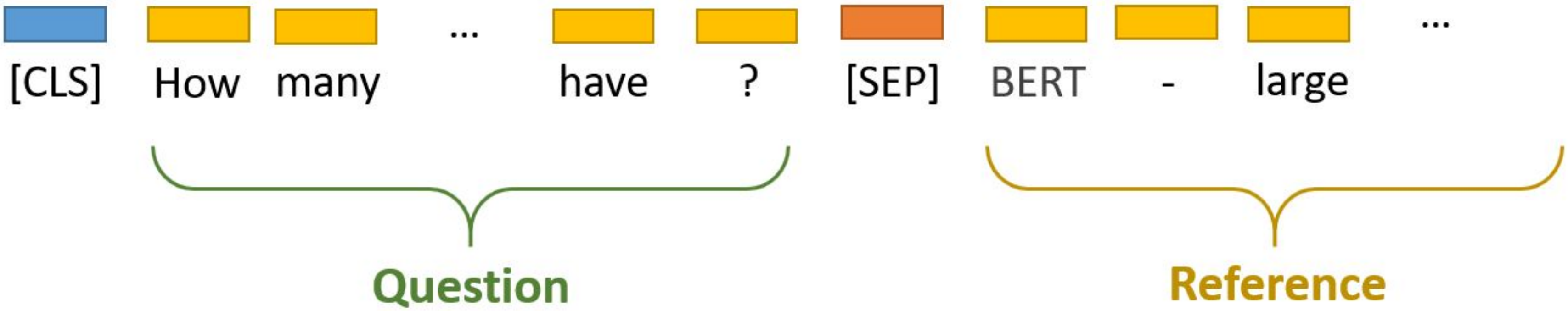
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

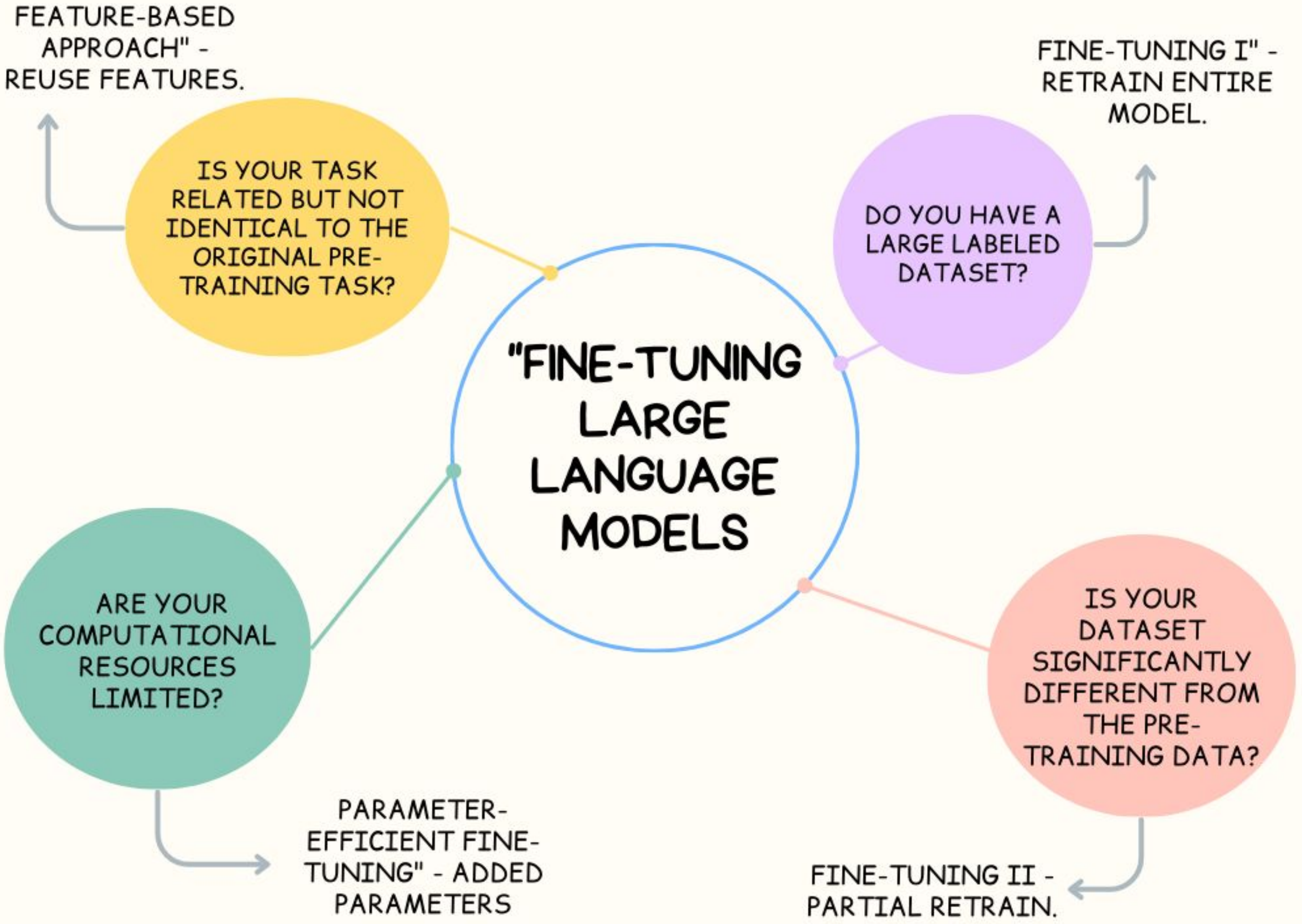
Answer Candidate

Gravity



Question: `How many parameters does BERT-large have?`

Reference Text: `BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.`



"FINE-TUNING LARGE LANGUAGE MODELS

IS YOUR TASK RELATED BUT NOT IDENTICAL TO THE ORIGINAL PRE-TRAINING TASK?

FEATURE-BASED APPROACH" - REUSE FEATURES.

DO YOU HAVE A LARGE LABELED DATASET?

FINE-TUNING I" - RETRAIN ENTIRE MODEL.

ARE YOUR COMPUTATIONAL RESOURCES LIMITED?

PARAMETER-EFFICIENT FINE-TUNING" - ADDED PARAMETERS

IS YOUR DATASET SIGNIFICANTLY DIFFERENT FROM THE PRE-TRAINING DATA?

FINE-TUNING II - PARTIAL RETRAIN.

How to fine-tune BIG models?

Quantization

- LLMs require a large amount of expensive GPU memory
  - Large number of parameters
  - High precision of the floating point numbers

Model	Original Size	Quantized Size (4-bit)
LLaMA2 7B	13 GB	3.9 GB
LLaMA2 13B	24 GB	7.8 GB
LLaMA2 30B	60 GB	19.5 GB
LLaMA2 65B	120 GB	38.5 GB

NVIDIA A100 has 80 GB memory and costs around INR 12-15 lakhs

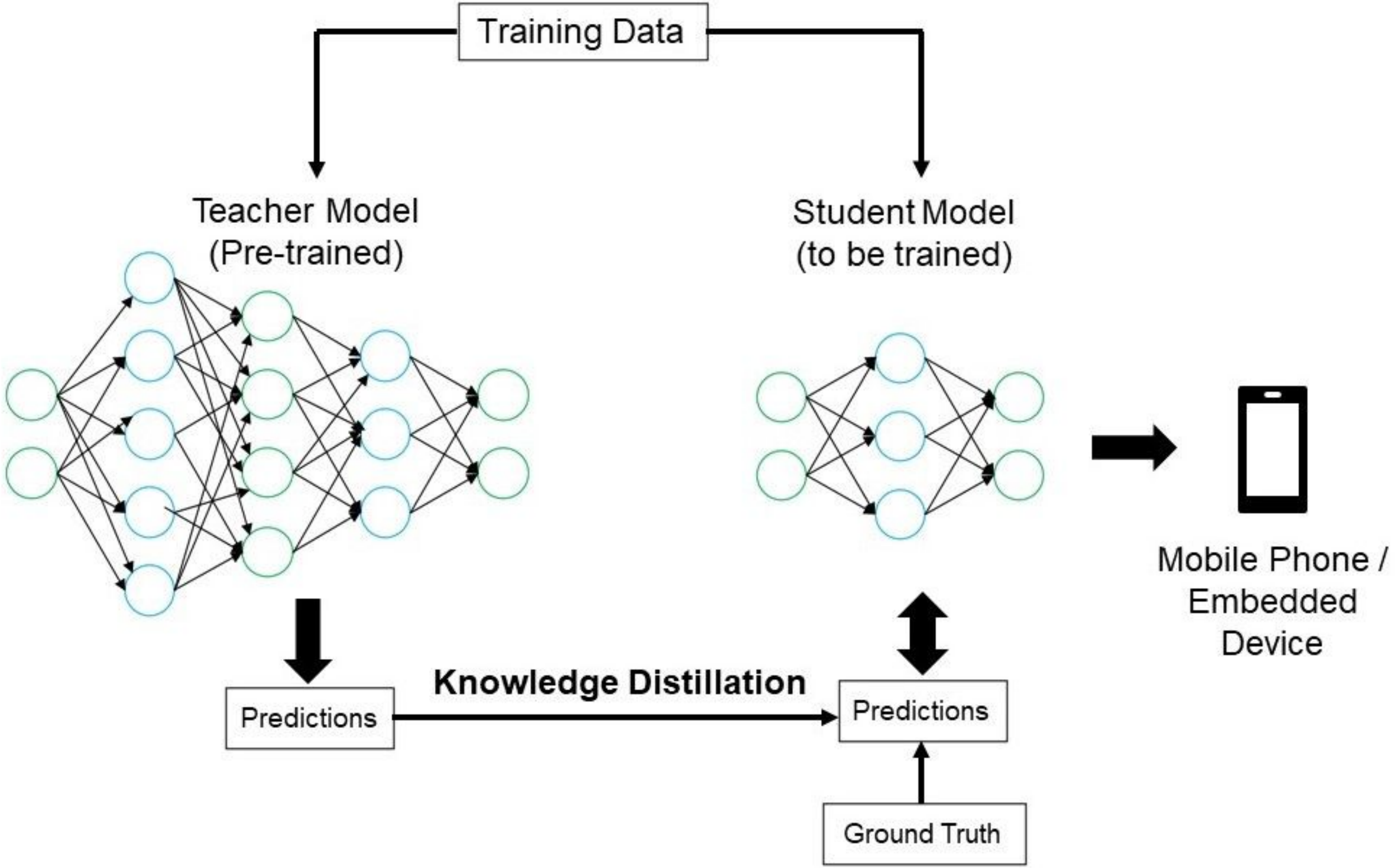
Distillation

- Transfer of knowledge from larger “teacher” model
  to a smaller “student” model

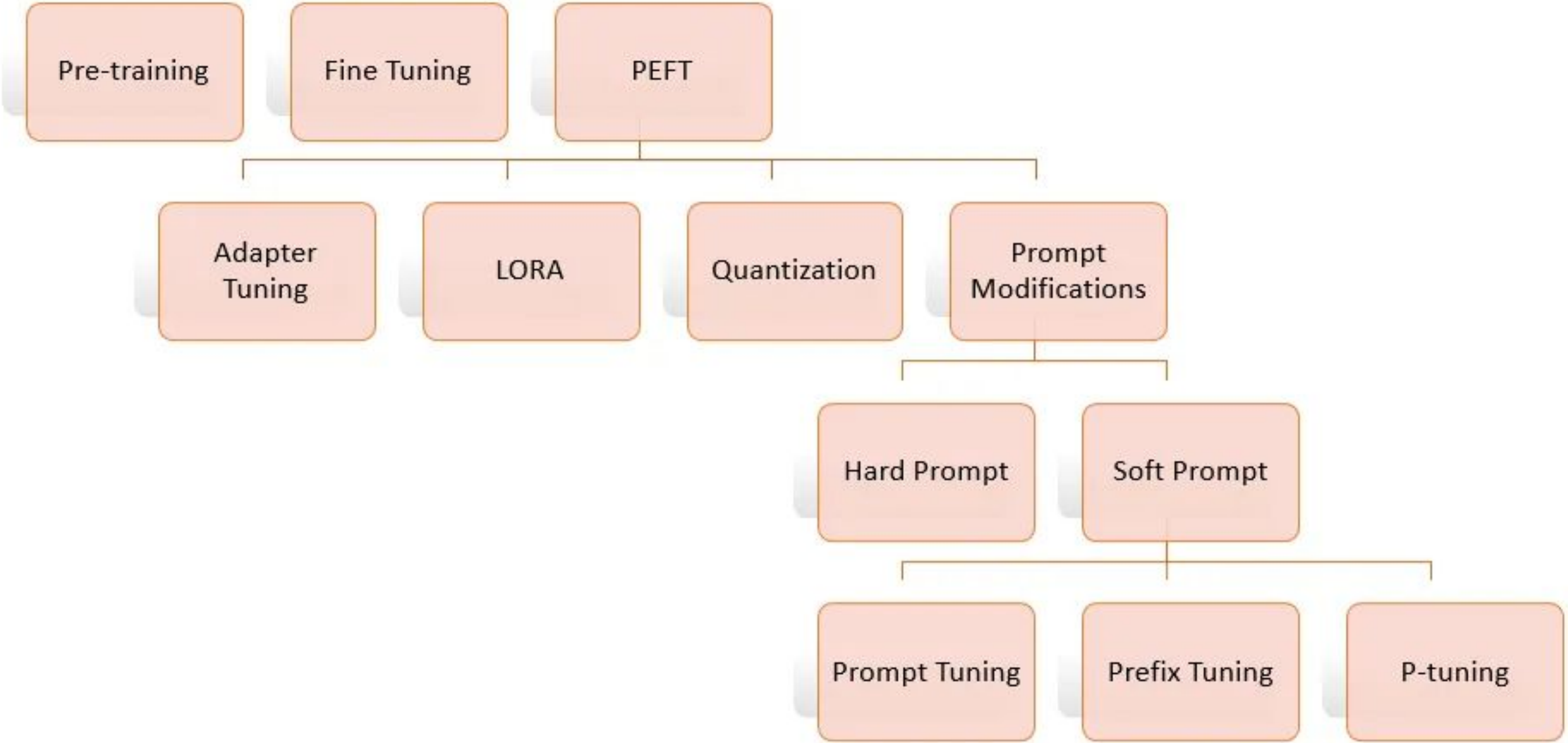
- Smaller model represents the bigger model for specific tasks

- Larger model learns the distribution from the data

- Smaller model learns the distribution from the larger model


Training Data
Teacher Model
(Pre-trained)
Student Model
(to be trained)
Mobile Phone /
Embedded
Device
Predictions
Knowledge Distillation
Predictions
Ground Truth

	BERT	RoBERT	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base:** 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	5% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling


Pre-training
Fine Tuning
PEFT
Adapter Tuning
LORA
Quantization
Prompt Modifications
Hard Prompt
Soft Prompt
Prompt Tuning
Prefix Tuning
P-tuning