Schedule

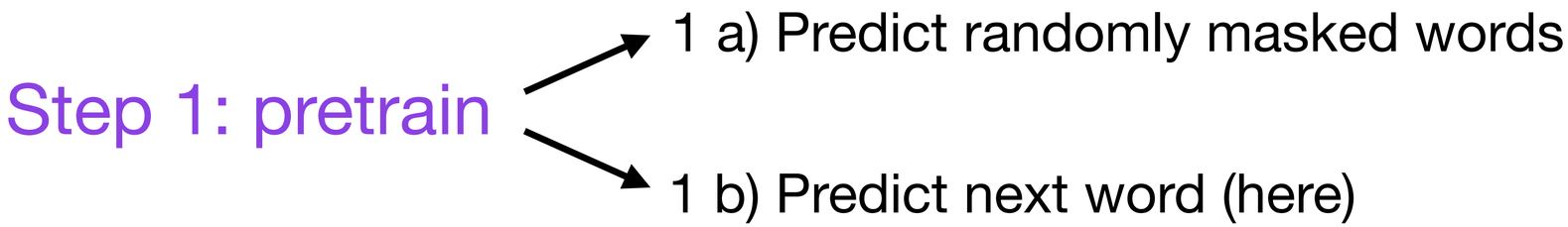
Block 4 (3:30 - 5:00 pm)

(8) Accelerating PyTorch model training

(9) Finetuning large language models

(10) Conclusion

Self-supervised pretraining



Step 2: fine-tune (

	Date
GPT 1	2018
GPT 2	2019
GPT 3	2020
InstructGPT & ChatGPT	2022

	Date	Size
GPT 1	2018	110 million
GPT 2	2019	1.5 billion
GPT 3	2020	175 billion
InstructGPT & ChatGPT	2022	GPT 3-based

	Date	Size	Paper
GPT 1	2018	110 million	Improving Language Understanding by Generative Pre-Training
GPT 2	2019	1.5 billion	Language Models are Unsupervised Multitask Learners
GPT 3	2020	175 billion	Language Models are Few-Shot Learners
InstructGPT & ChatGPT	2022	**GPT 3-based**	Aligning Language Models to Follow Instructions

GPT models are generative models (as opposed to classifiers)

GPT models are generative models (as opposed to classifiers)

How to make a banana smoothie?

To make a banana smoothie, you will need the following ingredients:

- 1 ripe banana
- 1 cup of milk (dairy or non-dairy)
- 1/2 cup of plain yogurt (dairy or non-dairy)
- 1 tablespoon of honey (optional)

Instructions:

1. Peel the banana and cut it into small chunks.
2. Add the banana chunks, milk, yogurt, and honey (if using) to a blender.
3. Blend the ingredients on high speed until smooth.
4. Taste and adjust sweetness if necessary.
5. Pour the smoothie into a glass and enjoy!
   You can also add some ice cubes, or some other fruits of your choice.

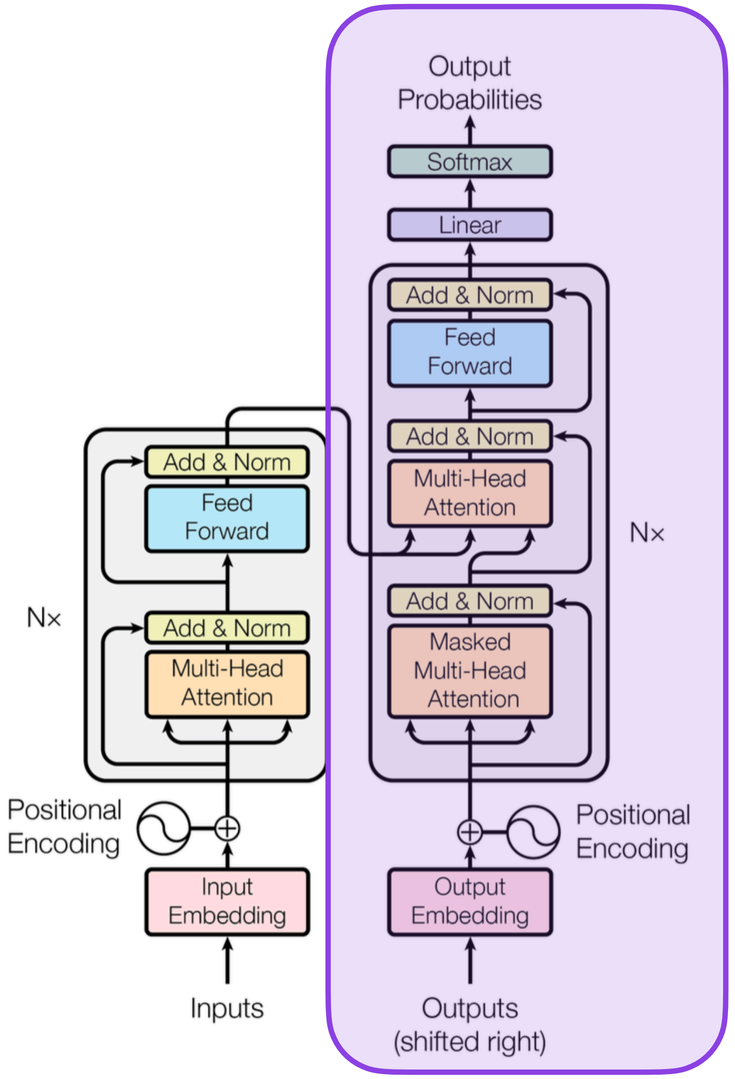
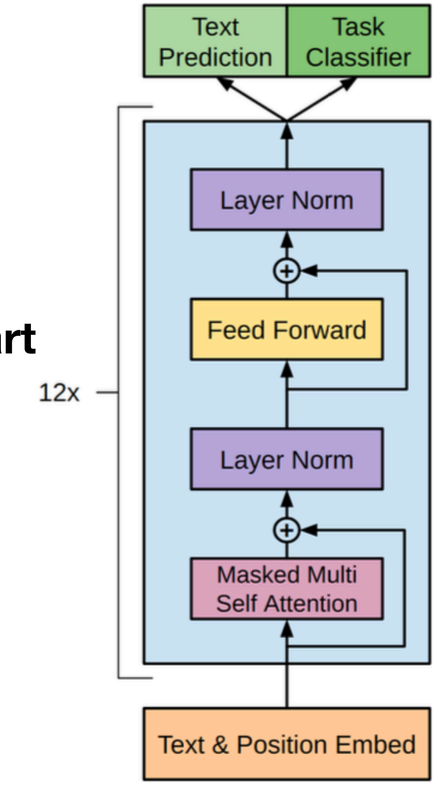


Figure 1: The Transformer - model architecture.

GPT is essentially the decoder part of the original transformer



https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

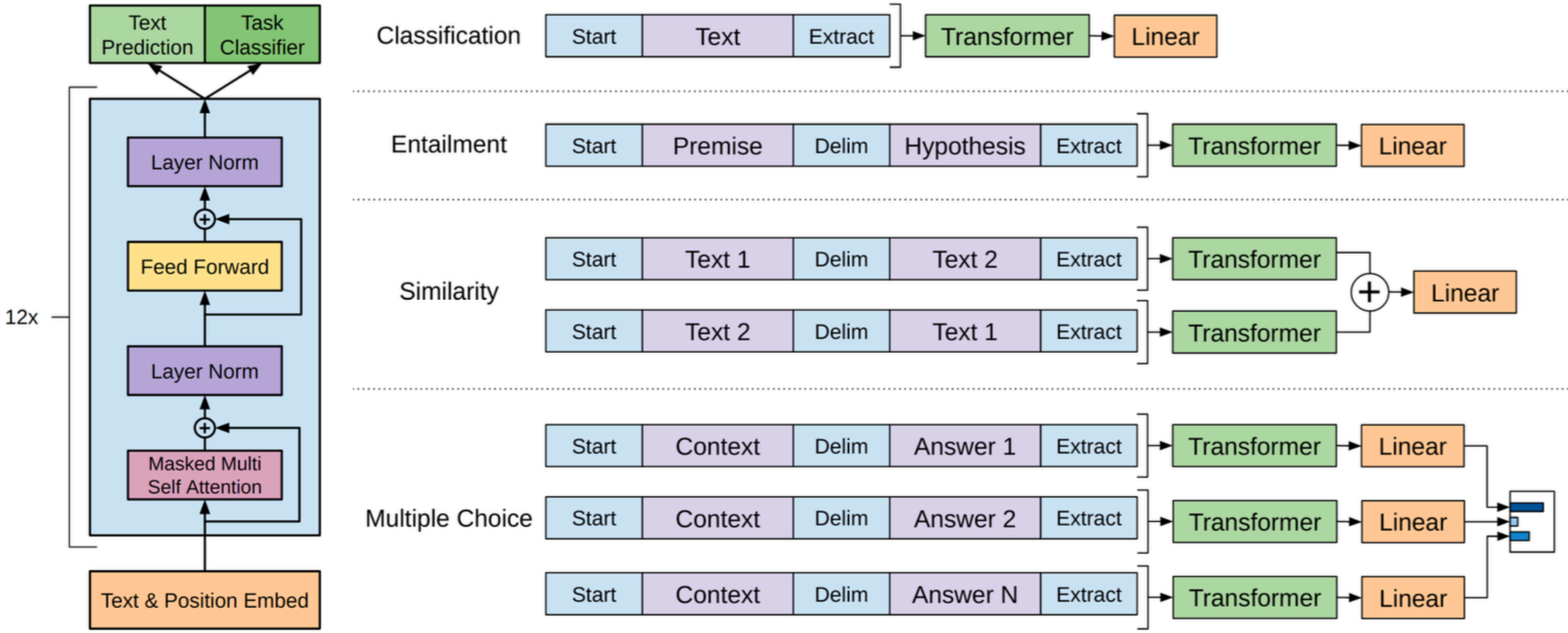
Feed model text from left to right, and it learns to predict the next word.

Self-supervised pretraining

Step 1: pretrain ⟶ **Predict next word** **t**

Step 2: fine-tune

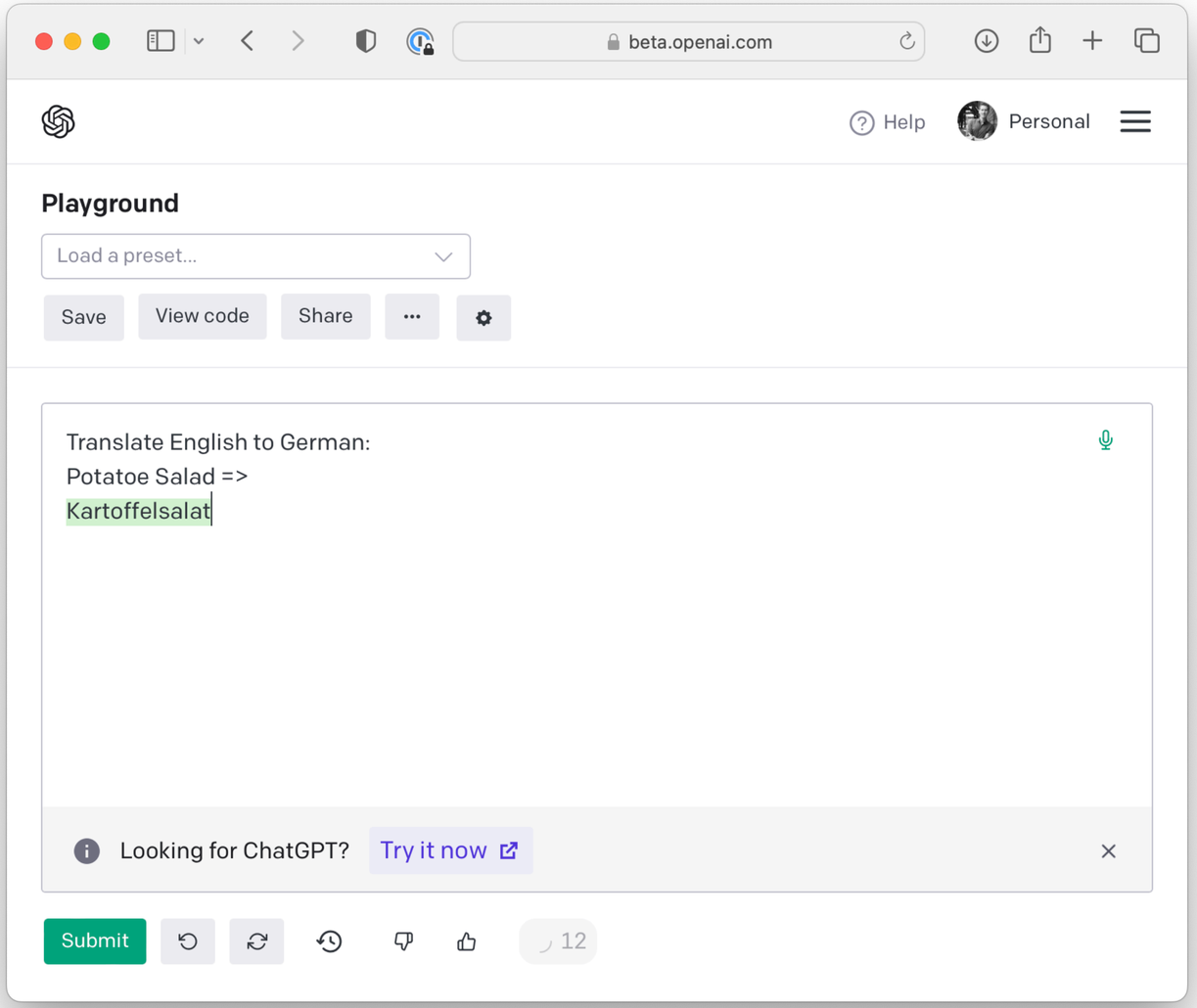
Fine-tune for target task



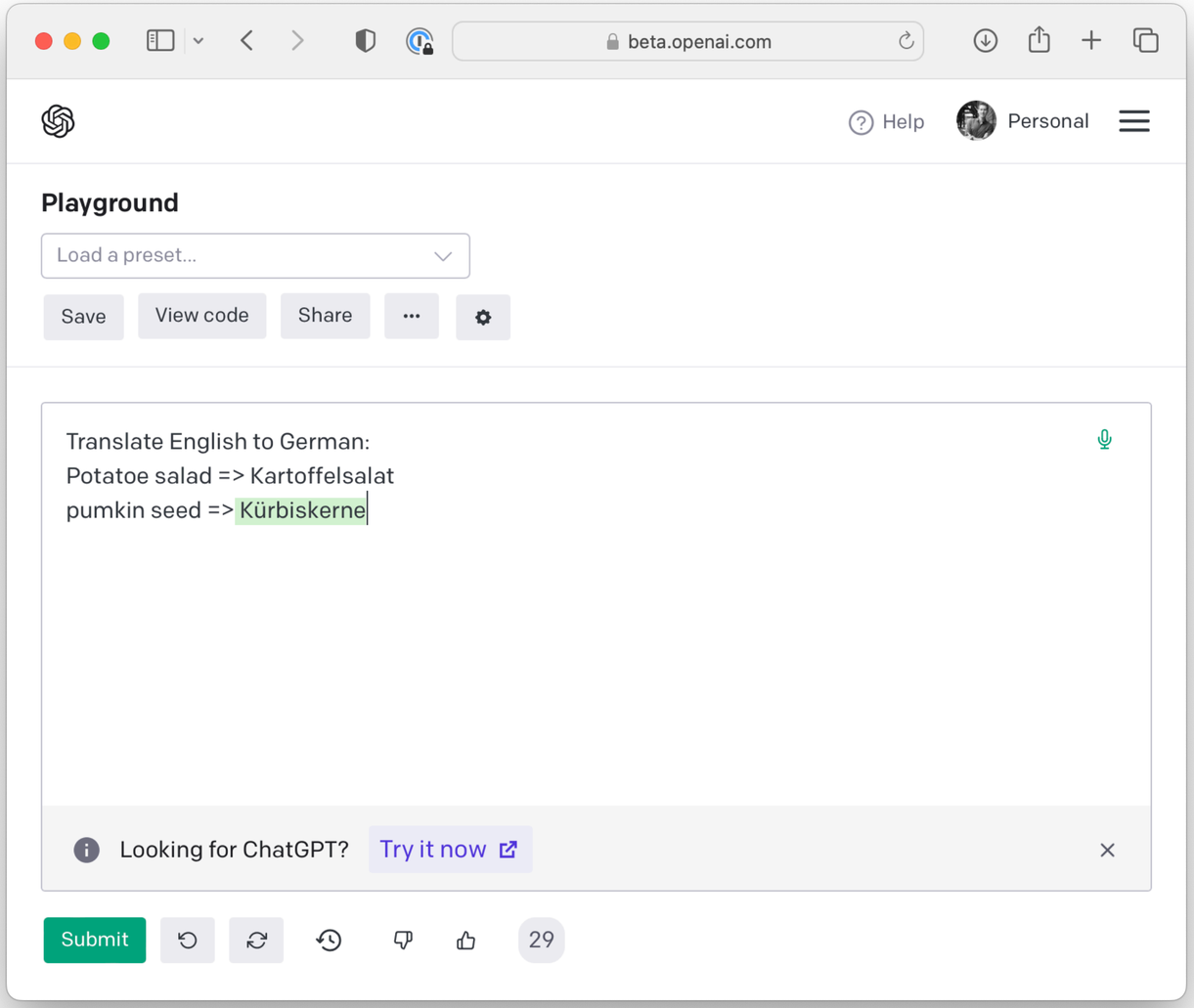
GPT 2 and 3 focused on zero- and few-shot learning

via in-context learning

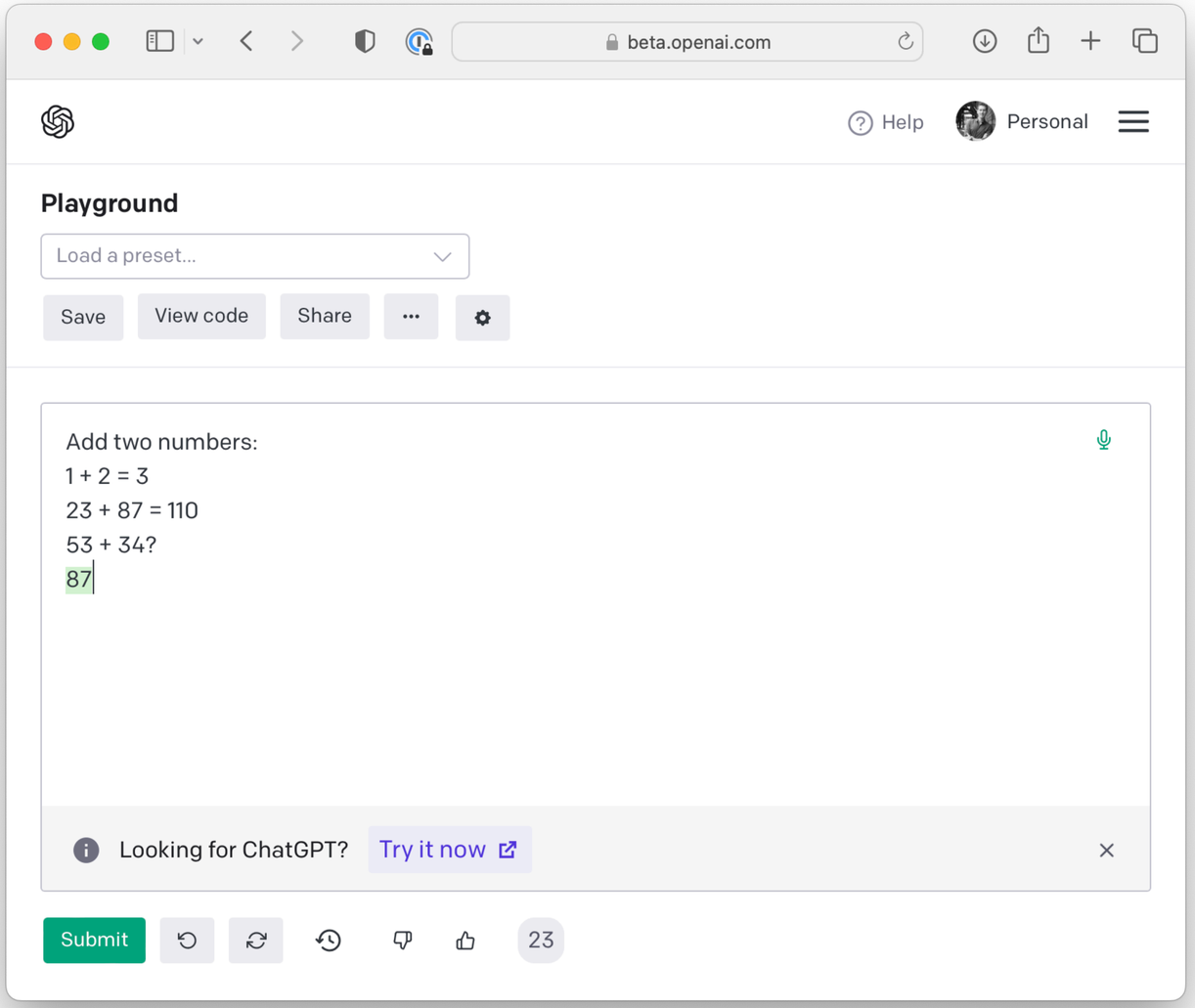
Zero-shot


beta.openai.com
Help
Personal
Playground
Load a preset...
Save
View code
Share
Translate English to German:
Potatoe Salad =>
Kartoffelsalat
Looking for ChatGPT?
Try it now
Submit
12

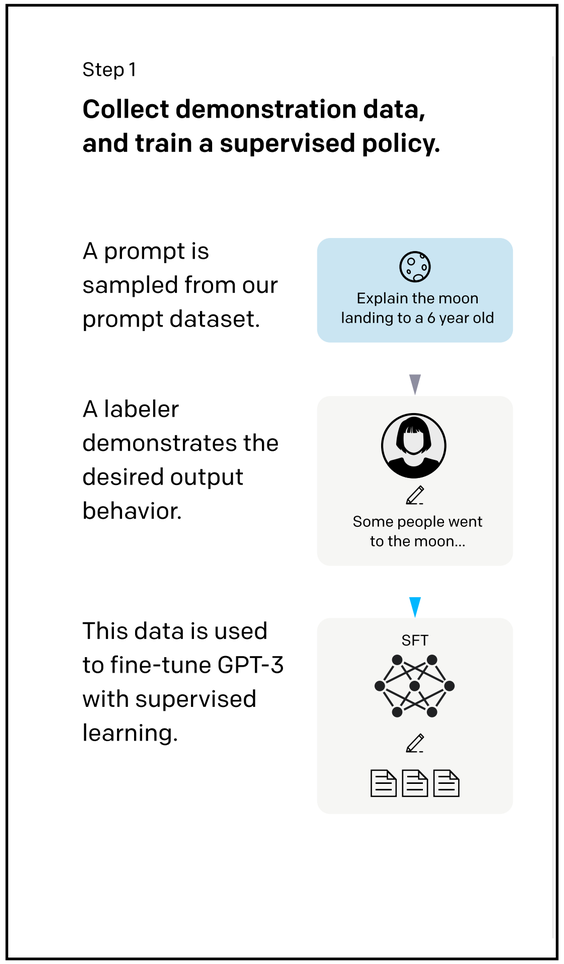
One-shot


beta.openai.com
Help
Personal
Playground
Load a preset...
Save
View code
Share
Translate English to German:
Potatoe salad => Kartoffelsalat
pumkin seed => Kürbiskerne
Looking for ChatGPT?
Try it now
Submit
29

Few-shot

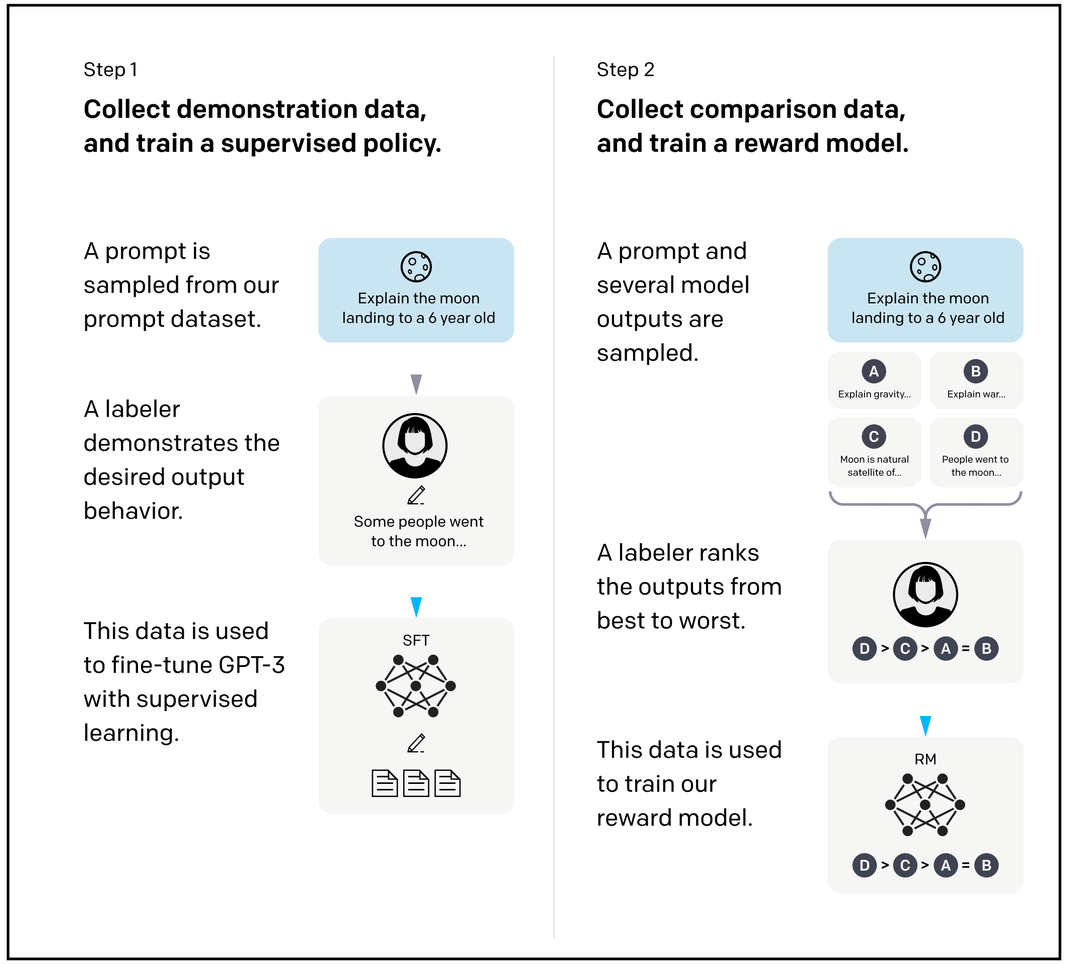

beta.openai.com
Help
Personal
Playground
Load a preset...
Save
View code
Share
Add two numbers:
1 + 2 = 3
23 + 87 = 110
53 + 34?
87
Looking for ChatGPT?
Try it now
Submit
23

InstructGPT and ChatGPT are Additionally Trained on Human Feedback



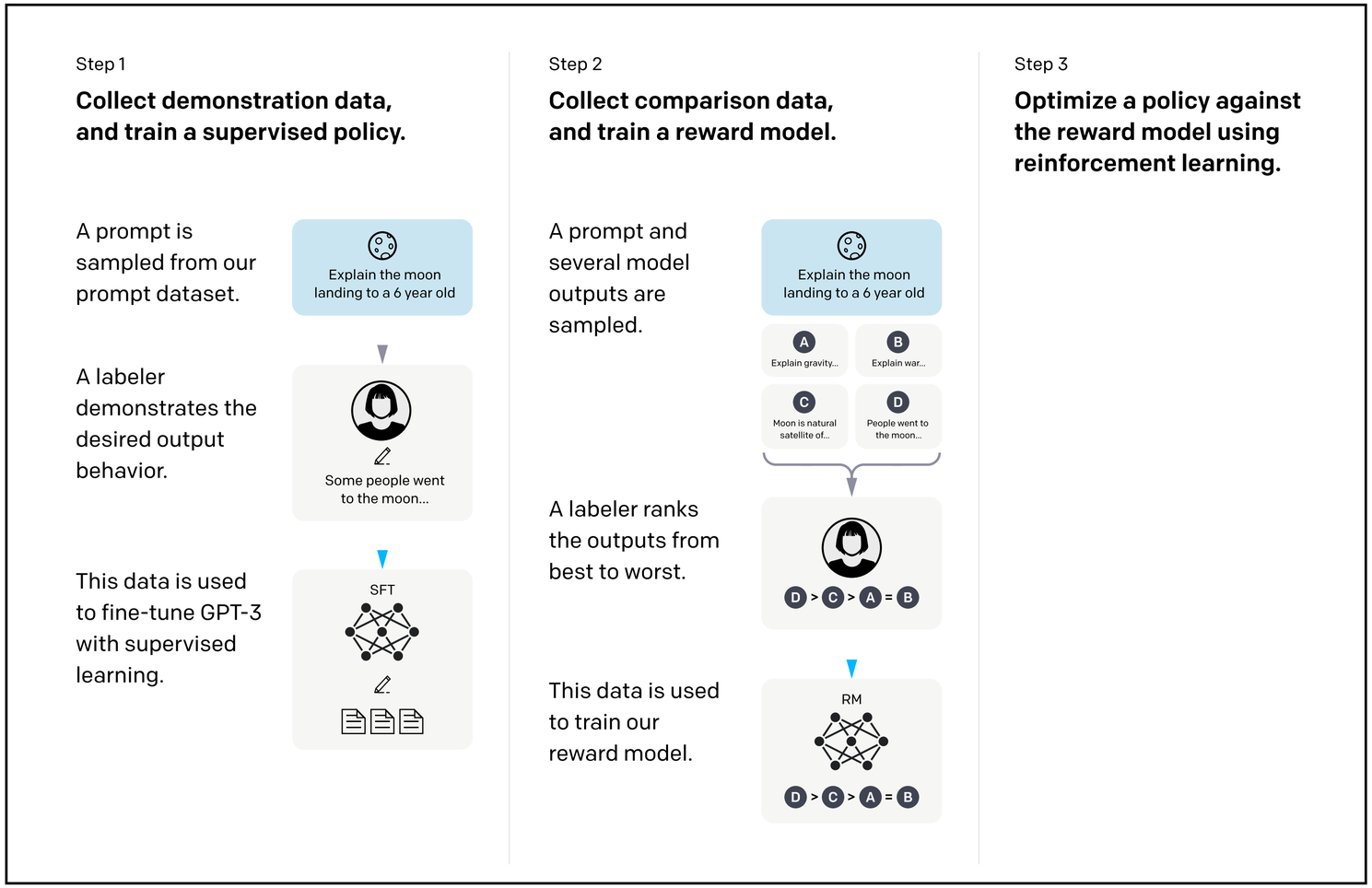
Training language models to follow instructions with human feedback, https://arxiv.org/abs/2203.02155

InstructGPT and ChatGPT are Additionally Trained on Human Feedback



Training language models to follow instructions with human feedback, https://arxiv.org/abs/2203.02155

InstructGPT and ChatGPT are Additionally Trained on Human Feedback



Training language models to follow instructions with human feedback, https://arxiv.org/abs/2203.02155

Today, transformers (large language models) are also used for …

- Classification (e.g., BERT)
- Various text summarization and generation tasks (e.g., GPT)
- Conversational chatbots (e.g., ChatGPT)
- Protein structure prediction from sequence data (e.g., AlphaFold 2)

GPT Recap

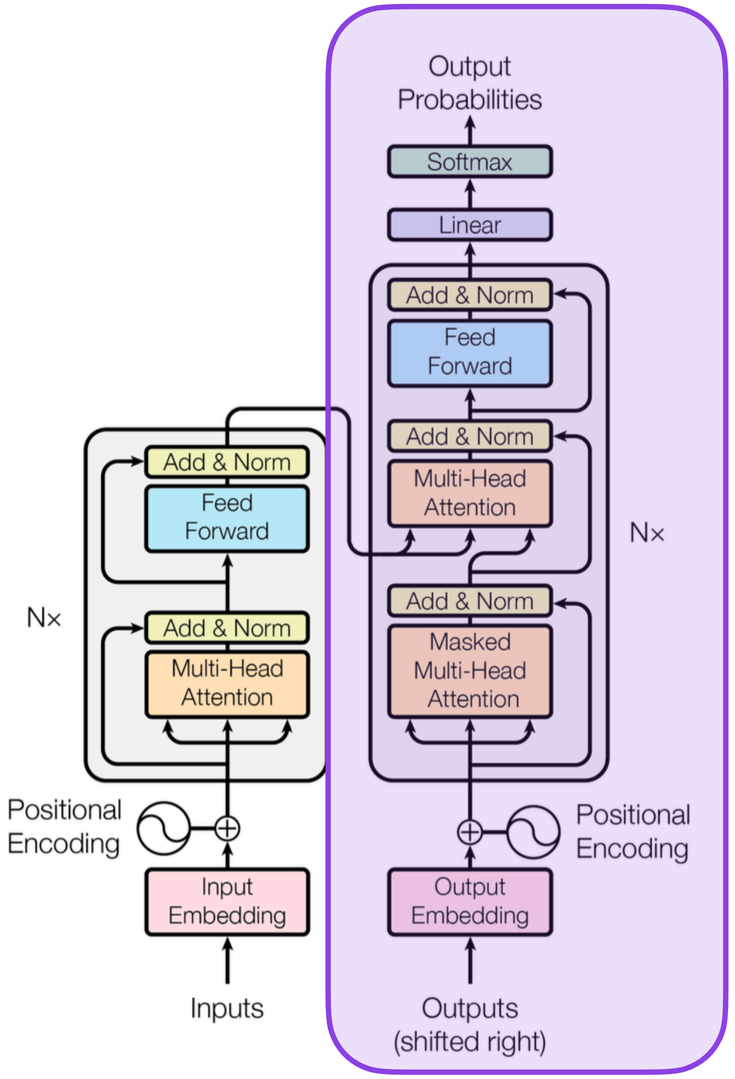
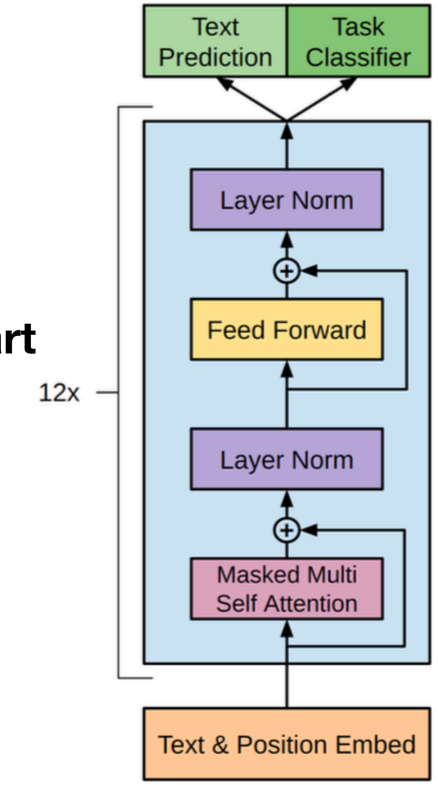


Figure 1: The Transformer - model architecture.

GPT is essentially the decoder part of the original transformer



https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

GPT Recap

Self-supervised pretraining

Step 1: pretrain ⟶ Predict next word (unidirectional self-attention)

Step 2: fine-tune

BERT

Self-supervised pretraining

Step 1: pretrain → Predict next word (~~unidirectional~~ self-attention)

a) Predict randomly masked words (bidirectional / nondirectional)

b) Sentence-order prediction

Step 2: fine-tune

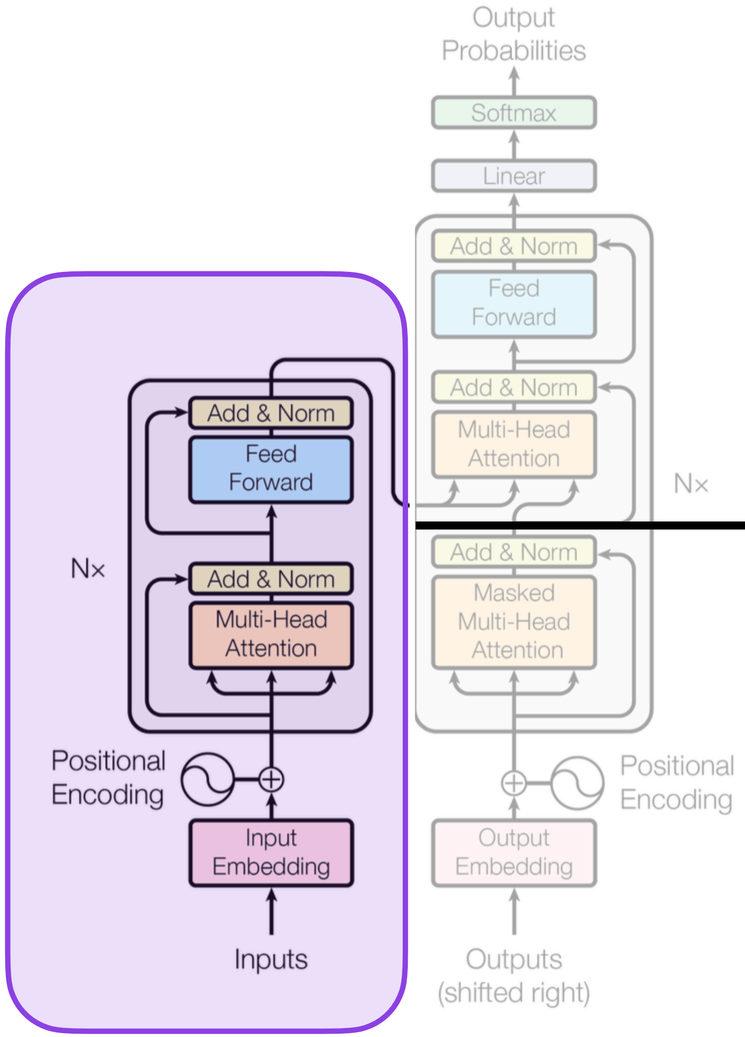
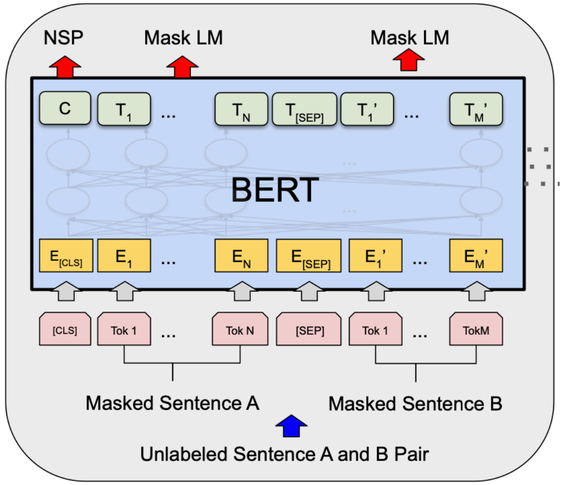


Figure 1: The Transformer - model architecture.

BERT is essentially the **encoder** part of the original transformer



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

Step 1: pretrain on large unlabeled dataset (learn a general language model)

a) Predict input sentence given randomly masked words

Input sentence: *The curious kitten deftly climbed the bookshelf*

Pick 15% of the words randomly

The curious kitten deftly climbed the bookshelf

Input sentence: *The curious kitten deftly climbed the bookshelf*

Pick 15% of the words randomly

The curious kitten deftly climbed the bookshelf

- **80% of the time, replace with [MASK] token**
- **10% of the time, replace with random token (e.g. ate)**
- **10% of the time, keep unchanged**

Step 1: pretrain on large unlabeled dataset (learn a general language model)

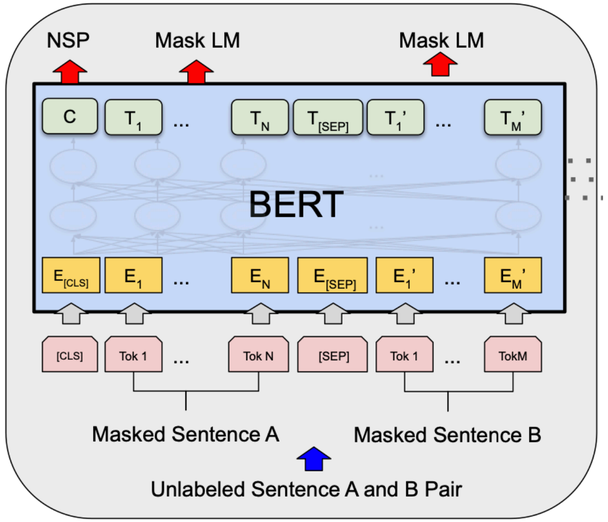
a) Predict input sentence given randomly masked words

b) Predict sentence order

b) Predict sentence order

[CLS] Sentence A [SEP] Sentence B

Placeholder for the IsNext=True / False label in the decoder output



b) Predict sentence order

[CLS] Toast is a simple yet delicious food [SEP] It's often served with butter, jam, or honey.

IsNext = True

[CLS] It's often served with butter, jam, or honey. [SEP] Toast is a simple yet delicious food.

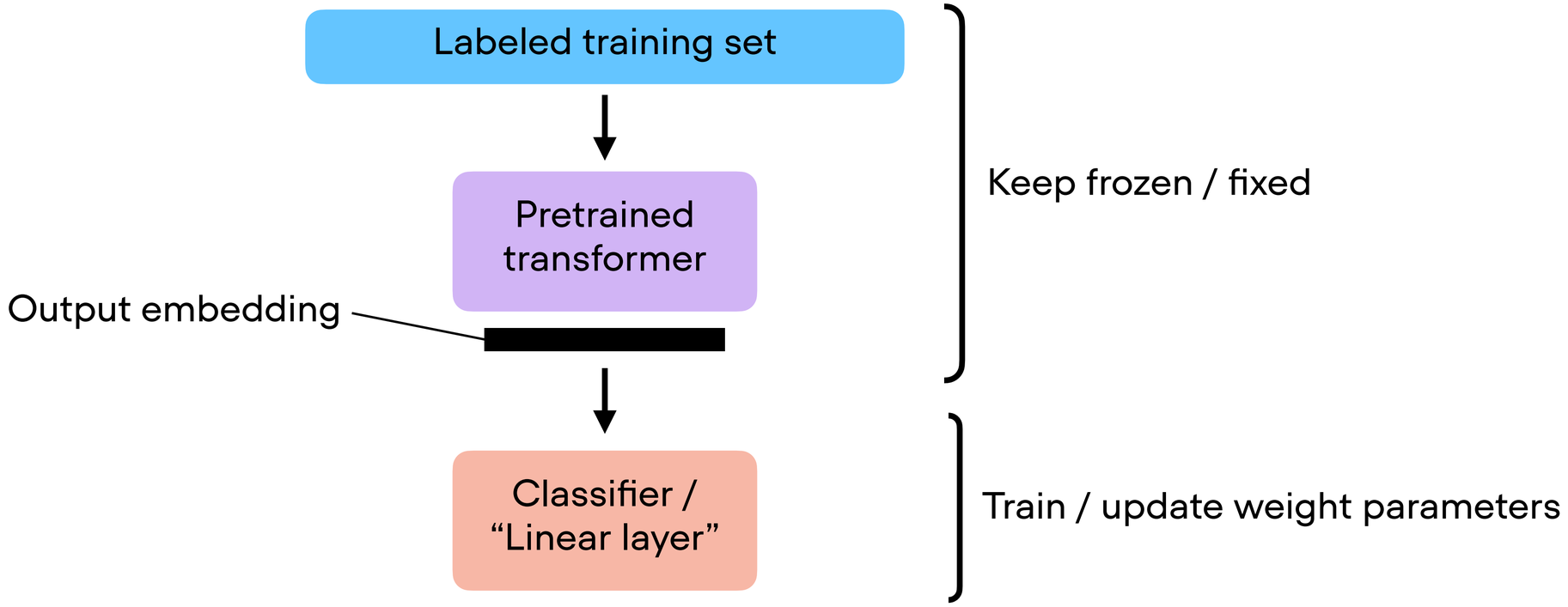
IsNext = False

2 ways of adopting a pretrained transformer for classification

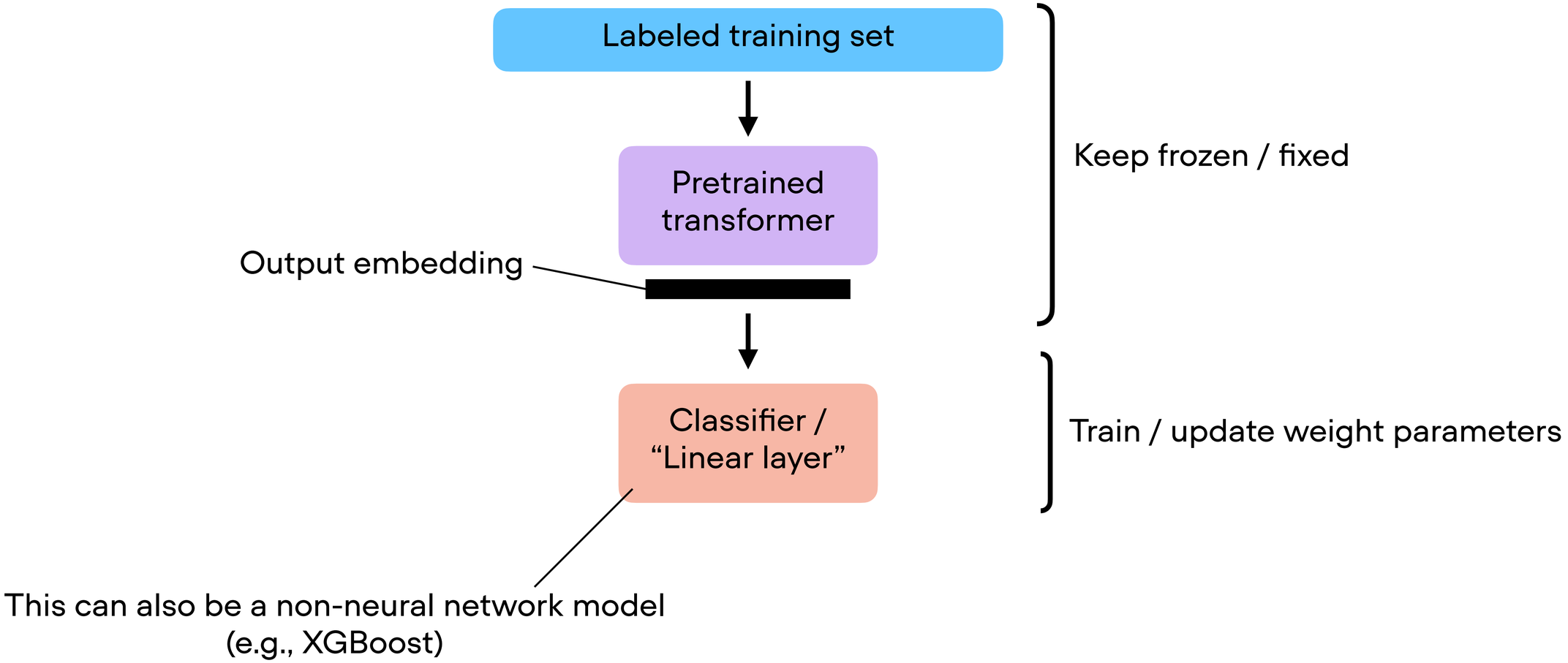
1) Feature-based approach

2) Fine-tuning approach

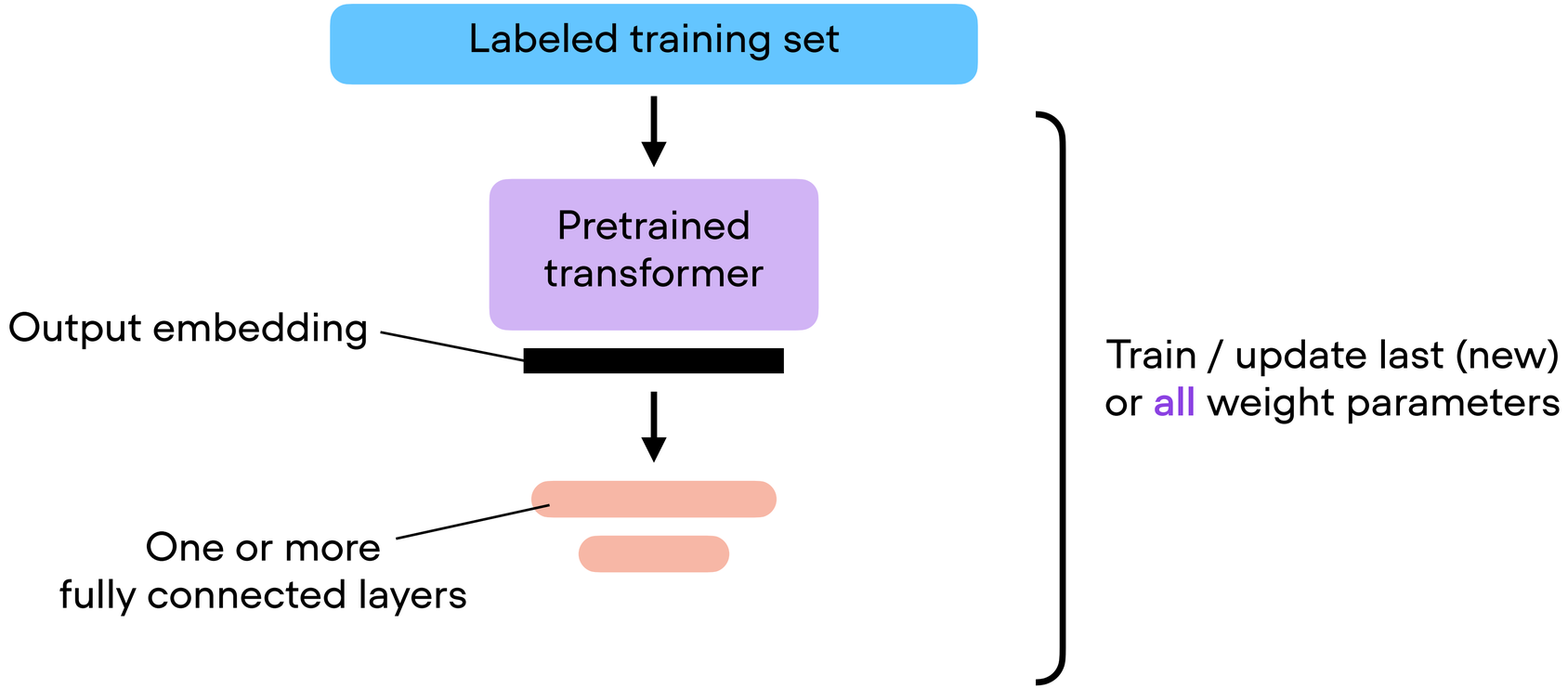
1) Feature-based approach



1) Feature-based approach

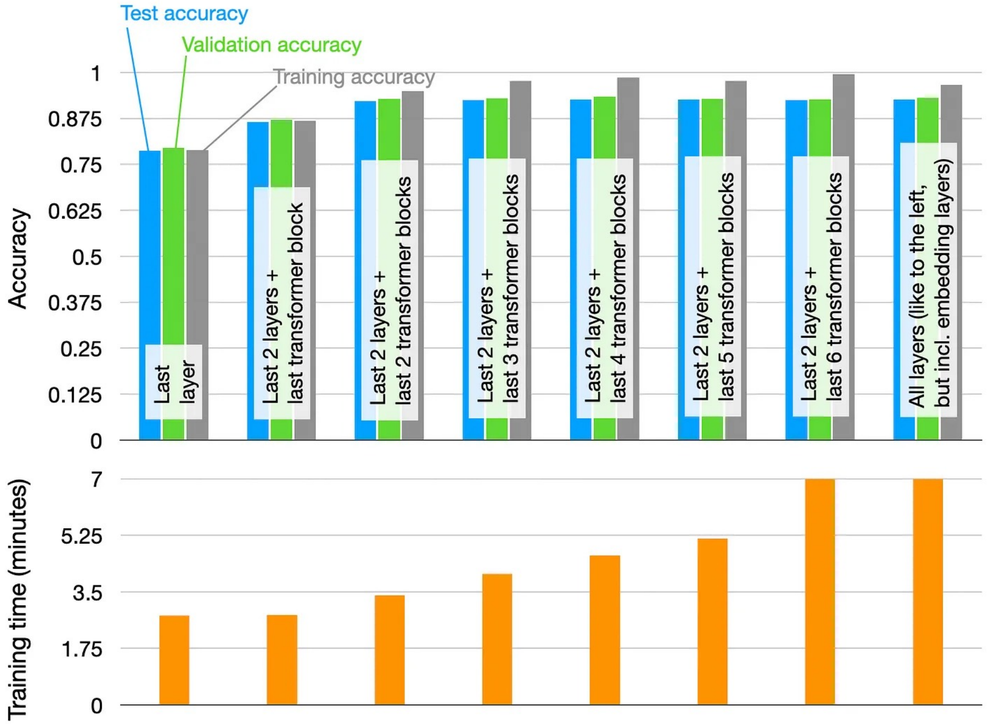


1) Fine-tuning approach



Exercise

Toggle between full finetuning and finetuning individual layers



(On 1 GPU)