BERT (Bidirectional Encoder Representations from Transformers)

> BERT is a pre-trained language model by Google that understands language deeply using Transformers.

- Unlike traditional models that read text **left-to-right** or **right-to-left**, **BERT reads both directions at once (bidirectional)** — which helps it understand context **better than older models**.

🧠 Why Was BERT Revolutionary?

Before BERT:

- Models could only look at words **before** the target word (e.g., GPT), or **after**.
- They couldn't get the **full meaning** of a word based on both sides of the sentence.

With BERT:

- It looks **both left and right** at the same time using **Transformers** (specifically only the **Encoder** part of Transformers).
- That's why it understands word meaning **in context**.

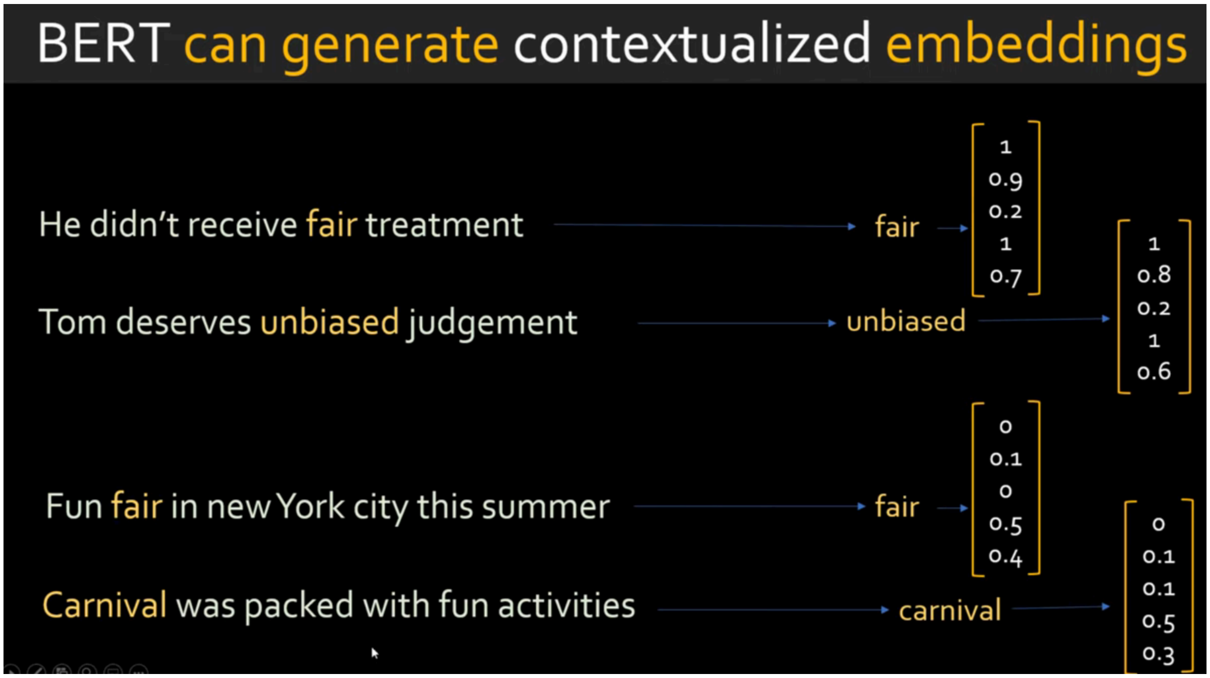
📌 Example

Consider the word "**bank**":

- "He sat by the river **bank**."

- "He went to the **bank** to withdraw money."

> **BERT will understand that “bank” in the first case is about nature, and in the second, it’s about money, because it sees both sides of the sentence.**



BERT can generate contextualized embeddings

He didn’t receive fair treatment

Tom deserves unbiased judgement

Fun fair in new York city this summer

Carnival was packed with fun activities

🔧 How is BERT Trained?

BERT uses **unsupervised pretraining**, then **finetuning**.

🧪 1. Pretraining Tasks

BERT is trained using two self-supervised tasks:

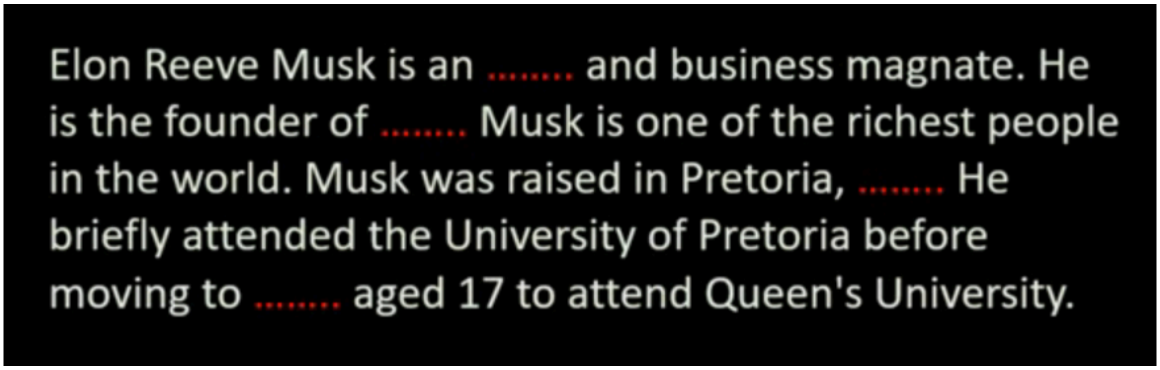
✅ A. Masked Language Modeling (MLM)

> Randomly masks some words in the sentence and asks the model to predict them.

🧱 Example:

```
Input:  "The cat sat on the [MASK]."
Target: "The cat sat on the mat."
```

- 15% of tokens are replaced with `[MASK]`
- BERT learns to **fill in the blanks** using full sentence context



Elon Reeve Musk is an .... and business magnate. He is the founder of .... Musk is one of the richest people in the world. Musk was raised in Pretoria, ... He briefly attended the University of Pretoria before moving to .... aged 17 to attend Queen's University.

✅ B. Next Sentence Prediction (NSP)

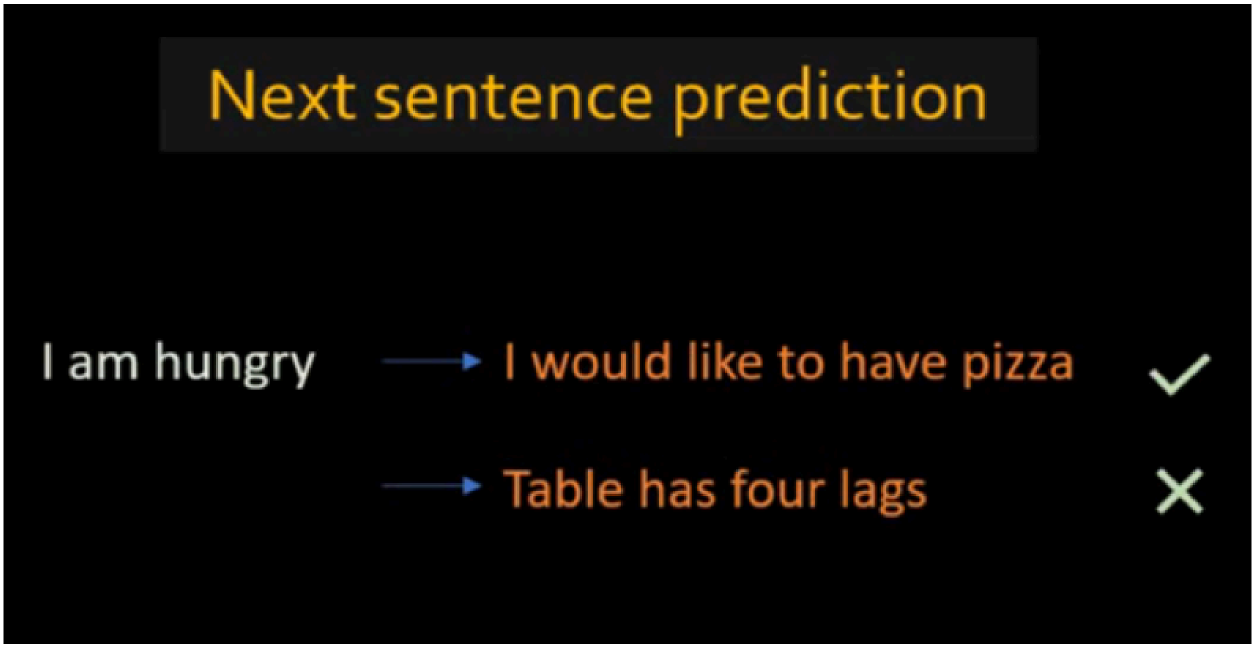
> BERT is shown pairs of sentences and asked if the second follows the first.

🧱 Example:

```
Sentence A: "The man went to the store."
Sentence B: "He bought a gallon of milk."

Label: IsNext → ✅
```

This helps BERT learn **relationships between sentences**, useful for tasks like QA and text inference.

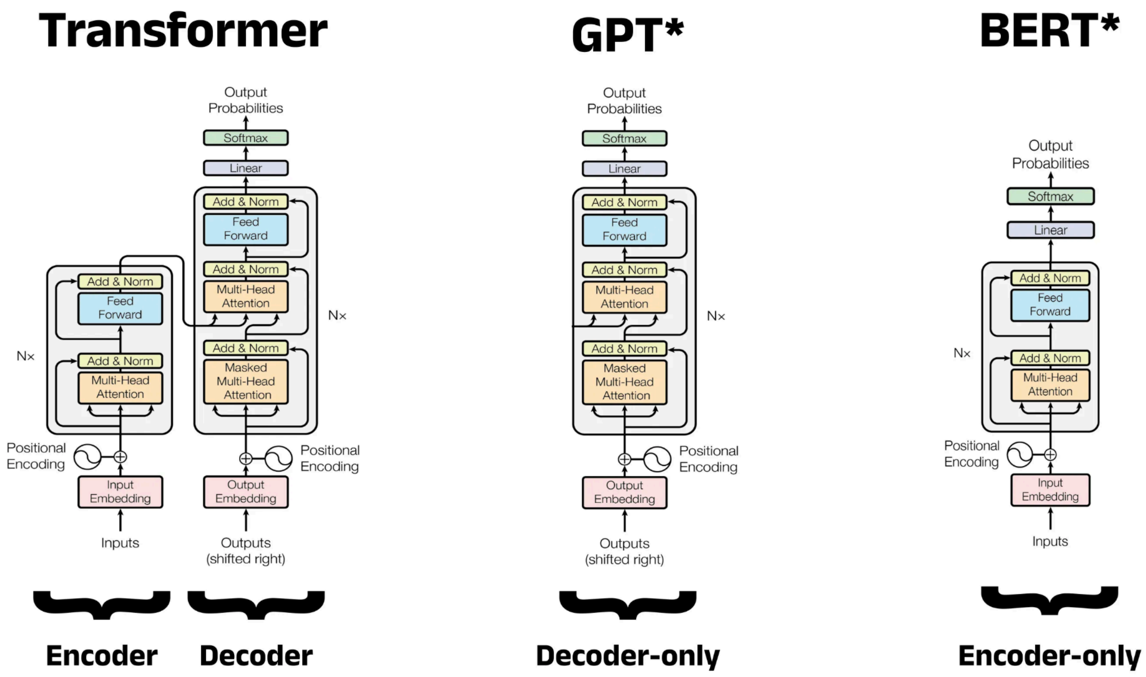


Next sentence prediction

I am hungry → I would like to have pizza ✓

→ Table has four lags ✗

📦 Architecture of BERT



- BERT uses **only the encoder** part of the Transformer.
- Has multiple **layers (blocks)** of self-attention and feedforward layers.

Common Variants:

Model	Layers	Hidden Size	Parameters
BERT-Base	12	768	110M
BERT-Large	24	1024	340M

💼 How to Use BERT in Real Tasks?

After pretraining, BERT is **fine-tuned** on specific tasks like:

- Sentiment classification
- Question answering
- Named Entity Recognition (NER)
- Text classification

Just **add a small output layer** on top and train on your dataset.

🎯 Real-World Use Cases

Task	How BERT Helps
Chatbots	Understanding questions
Search Engines	Semantic search
Customer Support	Classify issues from text
Medical NLP	Understand clinical notes
Legal Document Analysis	Extract facts and entities

🧠 Trivia

- Developed by **Google AI in 2018**
- Pretrained on **Wikipedia + BooksCorpus**
- BERT = **Bidirectional** + **Transformer Encoder**
- Spawned many variants: RoBERTa, DistilBERT, ALBERT, TinyBERT, etc.

Limitations

Issue	Workaround
Fixed context length (512 tokens)	Use Longformer or chunk inputs
Computationally expensive	Use smaller variants (e.g., TinyBERT)
No generative capability	Use BART or GPT for text generation