BERT

What is BERT?

- BERT = Bidirectional Encoder Representations from Transformers
- BERT is a deep learning model from Google that understands the meaning of text by reading sentences in both directions
- (left → right and right → left).

Why is BERT important

BERT understands:

- ✓ context
  ✓ meaning
  ✓ semantics
  ✓ relationships between words

How BERT Works?

- BERT is trained using **two techniques**:

1. Masked Language Modeling (MLM)

2. Next Sentence Prediction (NSP)

MASKED LANGUAGE MODELING (MLM)

Idea:

Hide (mask) some words in a sentence and ask BERT to **predict the missing word** correctly by understanding the context.

Example

Original sentence:
"The girl is drinking water."

We randomly mask 15% of words:

"The girl is drinking [MASK]."

Task for BERT:

Predict the masked word = **water**

Why this is useful?

- BERT learns meaning from **both left and right context**
  ("drinking " ***and*** " water")
- Helps BERT understand grammar, word meaning, sentence structure.

This is why BERT is **bidirectional** — it sees entire sentence simultaneously.

How BERT Predicts the MASK Word? (MLM)

Step 1: Mask a word

Sentence:
"Riya is drinking water."

Masked sentence:
"Riya is drinking [MASK]."

Step 2: Convert words to vectors

Each word → converted into embeddings
([Riya], [is], [drinking], [MASK])

Step 3: Transformers read the entire sentence

The BERT model reads:

- Words before mask: *Riya is drinking...*
- Words after mask: *...(nothing after water)*

It uses **self-attention** to understand:

- Verb: *drinking*
- Subject: *Riya*
- Common object: *water, juice, tea*

Step 4: Predict the most likely word

BERT calculates probability for **50,000 possible words.**

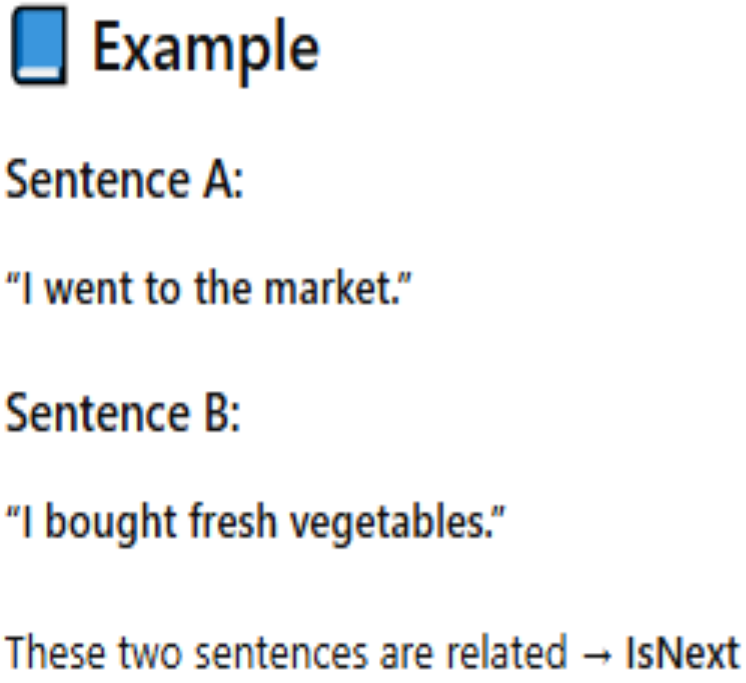
Example probability:

Word	Probability
water	0.92
tea	0.03
milk	0.02
juice	0.01

Highest probability = **water**
So BERT predicts **"water"** at **[MASK].**

How BERT Predicts Relationship Between Sentences? (NSP)

- BERT checks whether Sentence B **logically follows** Sentence A.



Step 1: Combine both sentences

BERT gets input like this:

[CLS] A sentence [SEP] B sentence [SEP]

(First token [CLS] is used for NSP decision)

Step 2: Self-attention across both sentences

BERT checks:

- Do words in A connect to words in B?
- "market" → "bought" (makes sense)
- "vegetables" → "fresh" (semantic connection)

It looks for logical flow, topic similarity, meaning link.

Step 3: Make a decision from [CLS] output

The final vector of [CLS] token tells:

- **0 → NotNext**
- **1 → IsNext**

If sentences are logically connected → predicts **IsNext**

If they are random or unrelated → predicts **NotNext**