After reading the BERT paper, here’s what I understood:

* **BERT** stands for *Bidirectional Encoder Representations from Transformers*. It’s a model built by Google to improve how machines understand language.
* The main idea is that BERT reads text in **both directions** (left and right) at once, which helps it understand meaning more deeply. This is different from earlier models that read only left-to-right like GPT or use two separate directions .
* BERT is trained in two main stages:
  1. **Pre-training** on huge amounts of text (like Wikipedia).
     + One task is **Masked Language Modeling (MLM)** – randomly hide some words in a sentence and train the model to guess them.
     + The other is **Next Sentence Prediction (NSP)** – tell whether one sentence logically follows another.
  2. **Fine-tuning** – once trained, BERT can be adapted to different tasks like question answering, sentence classification, or named entity recognition by just adding a simple layer and training it on labeled data.
* The input to BERT includes special tokens like **[CLS]** (used for classification) and **[SEP]** (used to separate sentences). It uses embeddings for tokens, position, and segments.
* They tested BERT on a bunch of NLP benchmarks like:
  1. **GLUE** (a mix of classification and sentence similarity tasks),
  2. **SQuAD v1.1 & v2.0** (question answering),
  3. **SWAG** (commonsense reasoning),
  4. and **CoNLL-2003** (named entity recognition).
* BERT achieved **state-of-the-art results** on almost all these tasks. It even beat older models.
* There are two versions of the model:
  1. **BERTBASE** (smaller, 12 layers),
  2. **BERTLARGE** (bigger, 24 layers). The large version performed better, especially when there was less training data.
* I also learned that BERT can be used in two ways:
  1. **Fine-tuning**, where the whole model is trained on the new task.
  2. **Feature-based**, where we just use BERT as a feature extractor and train a new model on top.
* One thing they tested was how important each part of BERT is. Removing either the NSP task or the bidirectionality made performance worse, which shows both are key to BERT’s success.

Overall, I understood that BERT was a major breakthrough because it combined bidirectional context, deep transformers, and simple fine-tuning to create a very general and powerful NLP model.