**🧠 Step 1: What is Falcon-7B?**

**Falcon-7B** is a **large language model (LLM)** like BERT or ChatGPT, developed by the Technology Innovation Institute (TII). It’s trained on web data and is capable of understanding and generating natural language.

| **Feature** | **Details** |
| --- | --- |
| Type | Decoder-only (like GPT) |
| Use Cases | Text generation, classification, summarization |
| Size | 7 billion parameters |
| Provider | Hugging Face (via tiiuae/falcon-7b) |
| Format | You interact with it using transformers in Python |

**🧩 Difference from BERT?**

| **BERT** | **Falcon-7B** |
| --- | --- |
| Encoder-only (Good for classification) | Decoder-only (Good for text generation) |
| Lightweight | Heavy (needs GPU) |
| Small input support | Can handle longer and complex inputs |
| Already used for your classifier | Now you'll try a new generation-based approach |

**💡 What Can You Do with Falcon-7B?**

Since Falcon is a **decoder model**, you don't "train" it easily on custom labels like BERT. Instead, you can:

1. 🧠 **Ask it**: “Is this a good resume for a Data Science role?”
2. 📋 **Generate feedback** for a resume
3. ✅ **Classify** resumes by prompting it like:

Given this resume: [resume text], which job category is it most suitable for?

This is known as **in-context prompting** (zero-shot or few-shot classification).

**⚙️ Step 2: Setup in Google Colab**

We will now implement Falcon-7B using **inference (no training)** to test your resumes.

Here’s what we’ll do:

1. 🔧 Install necessary libraries
2. 🔑 Load Falcon-7B model from Hugging Face
3. 📝 Feed a resume and ask it for classification
4. ✅ Get a prediction in plain English

**🧠 Step 1: What is Falcon-7B?**

**Falcon-7B** is a **large language model (LLM)** like BERT or ChatGPT, developed by the Technology Innovation Institute (TII). It’s trained on web data and is capable of understanding and generating natural language.

| **Feature** | **Details** |
| --- | --- |
| Type | Decoder-only (like GPT) |
| Use Cases | Text generation, classification, summarization |
| Size | 7 billion parameters |
| Provider | Hugging Face (via tiiuae/falcon-7b) |
| Format | You interact with it using transformers in Python |

**🧩 Difference from BERT?**

| **BERT** | **Falcon-7B** |
| --- | --- |
| Encoder-only (Good for classification) | Decoder-only (Good for text generation) |
| Lightweight | Heavy (needs GPU) |
| Small input support | Can handle longer and complex inputs |
| Already used for your classifier | Now you'll try a new generation-based approach |

**💡 What Can You Do with Falcon-7B?**

Since Falcon is a **decoder model**, you don't "train" it easily on custom labels like BERT. Instead, you can:

1. 🧠 **Ask it**: “Is this a good resume for a Data Science role?”
2. 📋 **Generate feedback** for a resume
3. ✅ **Classify** resumes by prompting it like:

plaintext

CopyEdit

Given this resume: [resume text], which job category is it most suitable for?

This is known as **in-context prompting** (zero-shot or few-shot classification).

**⚙️ Step 2: Setup in Google Colab**

We will now implement Falcon-7B using **inference (no training)** to test your resumes.

Here’s what we’ll do:

1. 🔧 Install necessary libraries
2. 🔑 Load Falcon-7B model from Hugging Face
3. 📝 Feed a resume and ask it for classification
4. ✅ Get a prediction in plain English