```
# 1️⃣ Downgrade Python
!sudo apt-get update -y
!sudo apt-get install python3.10 -y
!sudo update-alternatives --install /usr/bin/python3 python3 /usr/bin/python3.10 1
!sudo update-alternatives --config python3
```

```
Hit:1 https://cli.github.com/packages stable InRelease
Get:2 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease [3,632 B]
Hit:3 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease
Get:4 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Hit:5 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:6 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Get:7 https://r2u.stat.illinois.edu/ubuntu jammy InRelease [6,555 B]
Get:8 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease [18.1 kB]
Get:9 http://security.ubuntu.com/ubuntu jammy-security/restricted amd64 Packages [5,727 kB]
Get:10 https://r2u.stat.illinois.edu/ubuntu jammy/main all Packages [9,336 kB]
Hit:11 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
Hit:12 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:13 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [127 kB]
Get:14 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy/main amd64 Packages [32.8 kB]
Get:15 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1,582 kB]
Get:16 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 Packages [1,275 kB]
Get:17 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages [3,425 kB]
Get:18 http://archive.ubuntu.com/ubuntu jammy-updates/restricted amd64 Packages [5,922 kB]
Get:19 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [3,750 kB]
Get:20 https://r2u.stat.illinois.edu/ubuntu jammy/main amd64 Packages [2,811 kB]
Fetched 34.3 MB in 4s (7,653 kB/s)
Reading package lists... Done
W: Skipping acquire of configured file 'main/source/Sources' as repository 'https://r2u.stat.illinois.edu/ubuntu jammy InReleas
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
python3.10 is already the newest version (3.10.12-1~22.04.11).
python3.10 set to manually installed.
0 upgraded, 0 newly installed, 0 to remove and 41 not upgraded.
There are 2 choices for the alternative python3 (providing /usr/bin/python3).

  Selection    Path                Priority   Status
------------------------------------------------------------
* 0            /usr/bin/python3.12   2         auto mode
  1            /usr/bin/python3.10   1         manual mode
  2            /usr/bin/python3.12   2         manual mode

Press <enter> to keep the current choice[*], or type selection number: 1
update-alternatives: using /usr/bin/python3.10 to provide /usr/bin/python3 (python3) in manual mode
```

```
!pip uninstall -y bitsandbytes
```

```
Traceback (most recent call last):
  File "/usr/local/bin/pip", line 5, in <module>
    from pip._internal.cli.main import main
ModuleNotFoundError: No module named 'pip'
```

```
# Reinstall pip after Python downgrade
!curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py
!python get-pip.py
```

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 2098k  100 2098k    0     0  17.5M      0 --:--:-- --:--:-- --:--:-- 17.6M
Collecting pip
  Using cached pip-25.2-py3-none-any.whl.metadata (4.7 kB)
Collecting setuptools
  Using cached setuptools-80.9.0-py3-none-any.whl.metadata (6.6 kB)
Collecting wheel
  Using cached wheel-0.45.1-py3-none-any.whl.metadata (2.3 kB)
Using cached pip-25.2-py3-none-any.whl (1.8 MB)
Using cached setuptools-80.9.0-py3-none-any.whl (1.2 MB)
Using cached wheel-0.45.1-py3-none-any.whl (72 kB)
Installing collected packages: wheel, setuptools, pip
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 3/3 [pip]
Successfully installed pip-25.2 setuptools-80.9.0 wheel-0.45.1
```

```
from google.colab import drive
drive.mount("/content/drive")

# 📂 Set LoRA checkpoint path
lora_checkpoint_path = "/content/drive/MyDrive/AlpaCare_LoRA_Llama1B_FP16/alpacare-lora-llama1b-fp16/checkpoint-4500"

# 🔄 Load base model and LoRA adapter
import torch
from transformers import AutoModelForCausalLM, AutoTokenizer
```

```
from peft import PeftModel

# Load base model (FP16)
base_model_name = "meta-llama/Llama-3.2-1B-Instruct"
tokenizer = AutoTokenizer.from_pretrained(base_model_name)
tokenizer.pad_token = tokenizer.eos_token

base_model = AutoModelForCausalLM.from_pretrained(
    base_model_name,
    torch_dtype=torch.float16,
    device_map="auto",
    trust_remote_code=True
)

model = PeftModel.from_pretrained(
    base_model,
    lora_checkpoint_path,
    torch_dtype=torch.float16
)

model.eval()
print("✅ Model and LoRA adapter loaded successfully!")
```

```
Mounted at /content/drive
tokenizer_config.json: 100%  54.5k/54.5k [00:00<00:00, 4.32MB/s]
tokenizer.json: 100%  9.09M/9.09M [00:00<00:00, 37.8MB/s]
special_tokens_map.json: 100%  296/296 [00:00<00:00, 15.2kB/s]
config.json: 100%  877/877 [00:00<00:00, 67.9kB/s]
`torch_dtype` is deprecated! Use `dtype` instead!
model.safetensors: 100%  2.47G/2.47G [00:59<00:00, 116MB/s]
generation_config.json: 100%  189/189 [00:00<00:00, 18.1kB/s]
✅ Model and LoRA adapter loaded successfully!
```

```
# ================================================
# 🩺 Model Evaluation / Inference with Disclaimer
# ================================================

DISCLAIMER = (
    "⚠️ Disclaimer: This model is for research and educational purposes only. "
    "Its outputs should not be considered medical advice.\n\n"
)

# Function to generate responses with disclaimer
def generate_response(prompt, max_new_tokens=200, temperature=0.7, top_p=0.9):
    """
    Generate text from a given prompt using the LoRA-adapted model.
    Adds a disclaimer at the start of the output.
    """
    inputs = tokenizer(prompt, return_tensors="pt").to(model.device)

    with torch.no_grad():
        output_ids = model.generate(
            **inputs,
            max_new_tokens=max_new_tokens,
            temperature=temperature,
            top_p=top_p,
            eos_token_id=tokenizer.eos_token_id,
        )

    # Decode generated tokens
    response = tokenizer.decode(output_ids[0], skip_special_tokens=True)

    # Prepend disclaimer
    return DISCLAIMER + response

# Example prompts
prompts = [
    "Instruction: Provide a summary of diabetes management for adults.\nAnswer:",
    "Instruction: Suggest lifestyle modifications for a patient with hypertension.\nAnswer:"
]

# Generate and display responses
for i, prompt in enumerate(prompts, 1):
    print(f"--- Example {i} ---")
    print(generate_response(prompt))
    print("\n")
```

```
Setting `pad_token_id` to `eos_token_id`:128009 for open-end generation.
--- Example 1 ---
Setting `pad_token_id` to `eos_token_id`:128009 for open-end generation.
⚠️ Disclaimer: This model is for research and educational purposes only. Its outputs should not be considered medical advice.

Instruction: Provide a summary of diabetes management for adults.
Answer:Diabetes management for adults involves a combination of lifestyle modifications, medication, and regular monitoring of


--- Example 2 ---
⚠️ Disclaimer: This model is for research and educational purposes only. Its outputs should not be considered medical advice.

Instruction: Suggest lifestyle modifications for a patient with hypertension.
Answer:Lifestyle modifications can play a significant role in managing hypertension. Here are some suggestions: 1. Diet: - Limi
```

```
Start coding or generate with AI.
```