```
# === 1) INSTALL DEPENDENCIES ===
!pip install -q transformers accelerate bitsandbytes fastapi uvicor

# === 2) IMPORTS & HF AUTH ===
import os
import threading
import nest_asyncio
import torch
from collections import defaultdict
from fastapi import FastAPI, Request, HTTPException
import uvicorn
from transformers import AutoTokenizer, AutoModelForCausalLM
from huggingface_hub import login

# --- Load HF token from Colab secrets / env vars ---
from google.colab import userdata
userdata.get('HF_TOKEN')


# === 3) LOAD FULL GEMMA-3 12B MODEL (FP16) ===
model_id = "google/gemma-3-12b-it"
tokenizer = AutoTokenizer.from_pretrained(model_id)
model = AutoModelForCausalLM.from_pretrained(
    model_id,
    device_map="auto",
    torch_dtype=torch.float16,
    low_cpu_mem_usage=True
)
model.eval()

# === 4) IN-MEMORY CHAT HISTORY STORE ===
# Keyed by session_id, holds list of {"role","content"}
chat_history = defaultdict(list)

# === 5) FASTAPI APP DEFINITION ===
app = FastAPI()

@app.post("/chat")
async def chat_endpoint(request: Request):
    """
    Expects JSON:
      {
        "session_id": "user123",
        "message": "Hello, who are you?"
      }
    Returns:
      {
        "response": "I am Gemma, your agent..."
      }
    """
    data = await request.json()
    session_id = data.get("session_id")
    user_msg   = data.get("message")
    if not session_id or not user_msg:
        raise HTTPException(status_code=400, detail="`session_id` a

    # Append user message
    chat_history[session_id].append({"role":"user", "content": user
```

Resources

You are subscribed to Colab Pro. Learn more
Available: 36.89 compute units
Usage rate: approximately 7.62 per hour
You have 1 active session.

Manage sessions

Not connected to runtime.

Change runtime type

```
    # Build prompt with simple system header + conversation
    system_prompt = "You are a brilliant, context-aware AI agent."
    prompt_text = system_prompt + "\n"
    for turn in chat_history[session_id]:
        role = turn["role"]
        content = turn["content"]
        prompt_text += f"{role}: {content}\n"
    prompt_text += "assistant:"

    # Tokenize & generate
    inputs = tokenizer(prompt_text, return_tensors="pt").to(model.d
    output_ids = model.generate(**inputs, max_new_tokens=150)
    assistant_reply = tokenizer.decode(output_ids[0], skip_special_
    # Extract only the assistant's new text
    assistant_reply = assistant_reply.split("assistant:")[-1].strip

    # Save assistant reply
    chat_history[session_id].append({"role":"assistant", "content":
    return {"response": assistant_reply}

# === 6) LAUNCH Uvicorn IN BACKGROUND ===
def run_api():
    uvicorn.run(app, host="0.0.0.0", port=8000)

nest_asyncio.apply()
threading.Thread(target=run_api, daemon=True).start()

print("✅ Gemma-3 12B chat API is live at http://localhost:8000/cha
```

```
tokenizer_config.json: 100%  1.16M/1.16M [00:00<00:00, 8.60MB/s]
tokenizer.model: 100%  4.69M/4.69M [00:00<00:00, 51.2MB/s]
tokenizer.json: 100%  33.4M/33.4M [00:00<00:00, 181MB/s]
added_tokens.json: 100%  35.0/35.0 [00:00<00:00, 4.38kB/s]
special_tokens_map.json: 100%  662/662 [00:00<00:00, 91.7kB/s]
config.json: 100%  916/916 [00:00<00:00, 119kB/s]
model.safetensors.index.json: 100%  109k/109k [00:00<00:00, 11.6MB/s]
Fetching 5 files: 100%  5/5 [01:02<00:00, 63.00s/it]
model-00004-of-00005.safetensors: 100%  4.93G/4.93G [01:02<00:00, 203MB/s]
model-00005-of-00005.safetensors: 100%  4.60G/4.60G [01:01<00:00, 37.9MB/s]
model-00002-of-00005.safetensors: 100%  4.93G/4.93G [01:01<00:00, 87.6MB/s]
model-00001-of-00005.safetensors: 100%  4.98G/4.98G [01:02<00:00, 206MB/s]
```

```
from google.colab import files
files.upload()
```

```
Choose Files 556fc405-d...162bea.json
• 556fc405-d362-445a-a5fd-a62d2b162bea.json(application/json) - 175 bytes, last modified: 4/22/2025 - 100% done
Saving 556fc405-d362-445a-a5fd-a62d2b162bea.json to 556fc405-d
{'556fc405-d362-445a-a5fd-a62d2b162bea.json':
b'{"AccountTag":"fcc5c82d1ae27b077b2438eb72d3aa65","TunnelSecr
d362-445a-a5fd-a62d2b162bea","Endpoint":""}'}
```

```
!wget -q https://github.com/cloudflare/cloudflared/releases/latest
!chmod +x cloudflared
```

```
!lsof -i:8000
```

```
COMMAND PID USER   FD   TYPE DEVICE SIZE/OFF NODE NAME
python3 883 root  115u  IPv4 216516      0t0  TCP *:8000 (LIST
```

```
!curl -v http://localhost:8000/chat
```

```
INFO:     127.0.0.1:44288 - "GET /chat HTTP/1.1" 405 Method Nc
*   Trying 127.0.0.1:8000...
* Connected to localhost (127.0.0.1) port 8000 (#0)
> GET /chat HTTP/1.1
> Host: localhost:8000
> User-Agent: curl/7.81.0
> Accept: */*
>
* Mark bundle as not supporting multiuse
< HTTP/1.1 405 Method Not Allowed
< date: Tue, 22 Apr 2025 13:51:42 GMT
< server: uvicorn
< allow: POST
< content-length: 31
< content-type: application/json
<
* Connection #0 to host localhost left intact
{"detail":"Method Not Allowed"}
```

```
!pip install fastapi uvicorn nest_asyncio
import nest_asyncio
from fastapi import FastAPI
from pydantic import BaseModel
import uvicorn

app = FastAPI()

class ChatRequest(BaseModel):
    session_id: str
    message: str
```

```
    persona: str

@app.post("/chat")
def chat(req: ChatRequest):
    return {"echo": req.message, "persona": req.persona}

nest_asyncio.apply()
uvicorn.run(app, host="0.0.0.0", port=8500)
```

```
Requirement already satisfied: fastapi in /usr/local/lib/pythc
Requirement already satisfied: uvicorn in /usr/local/lib/pythc
Requirement already satisfied: nest_asyncio in /usr/local/lib/
Requirement already satisfied: starlette<0.47.0,>=0.40.0 in /u
Requirement already satisfied: pydantic!=1.8,!=1.8.1,!=2.0.0,!
Requirement already satisfied: typing-extensions>=4.8.0 in /us
Requirement already satisfied: click>=7.0 in /usr/local/lib/py
Requirement already satisfied: h11>=0.8 in /usr/local/lib/pyth
Requirement already satisfied: annotated-types>=0.6.0 in /usr/
Requirement already satisfied: pydantic-core==2.33.1 in /usr/l
Requirement already satisfied: typing-inspection>=0.4.0 in /us
Requirement already satisfied: anyio<5,>=3.6.2 in /usr/local/l
Requirement already satisfied: idna>=2.8 in /usr/local/lib/pyt
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/
INFO:     Started server process [883]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
INFO:     Uvicorn running on http://0.0.0.0:8500 (Press CTRL+C
------------------------------------------------------------
--------------
RuntimeError                              Traceback (most
recent call last)
<ipython-input-24-a0db21acc329> in <cell line: 0>()
     17
     18 nest_asyncio.apply()
---> 19 uvicorn.run(app, host="0.0.0.0", port=8500)

                    3 frames
/usr/local/lib/python3.11/dist-packages/nest_asyncio.py in
run_until_complete(self, future)
     94                     break
     95             if not f.done():
---> 96                 raise RuntimeError(
     97                     'Event loop stopped before Future
completed.')
     98             return f.result()
```

Next steps: Explain error

```
!./cloudflared tunnel --url http://localhost:8500
```

```
!ls
```

```
556fc405-d362-445a-a5fd-a62d2b162bea.json  cloudflared  sample
```

```
%%writefile config.yaml
tunnel: ghost-agent-tunnel
```

```
credentials-file: ./556fc405-d362-445a-a5fd-a62d2b162bea.json

ingress:
  - hostname: api.div3rcity.me
    service: http://localhost:8000
  - service: http_status:404
```

Writing config.yaml

```
!./cloudflared tunnel --config config.yaml run
```

```
2025-04-22T13:48:01Z INF Starting tunnel tunnelID=556fc405-d36
2025-04-22T13:48:01Z INF Version 2025.4.0 (Checksum df13e7e0a6
2025-04-22T13:48:01Z INF GOOS: linux, GOVersion: go1.22.10, Gc
2025-04-22T13:48:01Z INF Settings: map[config:config.yaml crec
2025-04-22T13:48:01Z INF cloudflared will not automatically up
2025-04-22T13:48:01Z INF Generated Connector ID: c1b04890-61e6
2025-04-22T13:48:01Z INF Initial protocol quic
2025-04-22T13:48:01Z INF ICMP proxy will use 172.28.0.12 as sc
2025-04-22T13:48:01Z INF ICMP proxy will use :: as source for
2025-04-22T13:48:01Z INF ICMP proxy will use 172.28.0.12 as sc
2025-04-22T13:48:01Z INF ICMP proxy will use :: as source for
2025-04-22T13:48:01Z INF Starting metrics server on 127.0.0.1:
2025-04-22T13:48:01Z INF Using [CurveID(4588) CurveID(25497) C
2025/04/22 13:48:01 failed to sufficiently increase receive bu
2025-04-22T13:48:01Z INF Registered tunnel connection connInde
2025-04-22T13:48:01Z INF Using [CurveID(4588) CurveID(25497) C
2025-04-22T13:48:01Z INF Registered tunnel connection connInde
2025-04-22T13:48:02Z INF Using [CurveID(4588) CurveID(25497) C
2025-04-22T13:48:02Z INF Registered tunnel connection connInde
2025-04-22T13:48:03Z INF Using [CurveID(4588) CurveID(25497) C
2025-04-22T13:48:03Z INF Registered tunnel connection connInde
2025-04-22T13:49:44Z INF Initiating graceful shutdown due to s
2025-04-22T13:49:44Z ERR Failed to serve tunnel connection err
2025-04-22T13:49:44Z ERR Serve tunnel error error="Applicatior
2025-04-22T13:49:44Z INF Retrying connection in up to 1s connI
2025-04-22T13:49:44Z ERR Failed to serve tunnel connection err
2025-04-22T13:49:44Z ERR Serve tunnel error error="Applicatior
2025-04-22T13:49:44Z INF Retrying connection in up to 1s connI
2025-04-22T13:49:44Z ERR Failed to serve tunnel connection err
2025-04-22T13:49:44Z ERR Serve tunnel error error="Applicatior
2025-04-22T13:49:44Z INF Retrying connection in up to 1s connI
2025-04-22T13:49:44Z ERR Failed to serve tunnel connection err
2025-04-22T13:49:44Z ERR Serve tunnel error error="Applicatior
2025-04-22T13:49:44Z INF Retrying connection in up to 1s connI
2025-04-22T13:49:44Z ERR no more connections active and exitir
2025-04-22T13:49:44Z INF Tunnel server stopped
2025-04-22T13:49:44Z ERR icmp router terminated error="context
2025-04-22T13:49:44Z INF Metrics server stopped
```