Sakka Mohamad-Mario 1241EB – Subgroup II

# SOFTWARE ENGINEERING HW II

## 1. GOAL

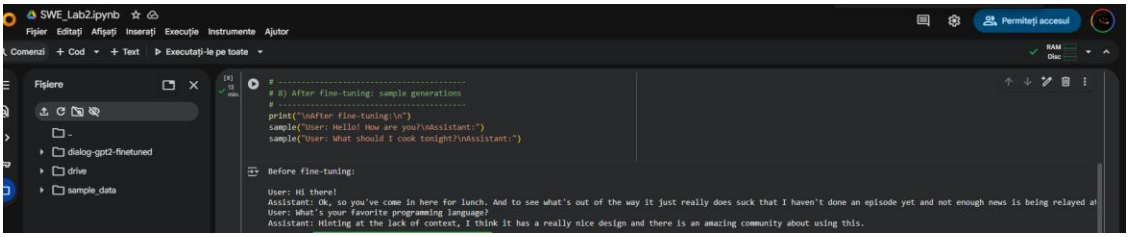
I tried to fine-tune GPT-2 to write dialog, using a public dataset from Hugging Face (DailyDialog mirror, agentlans/li2017dailydialog, Parquet splits: train/validation/test). I converted each conversation into alternating lines (User: … / Assistant: …), skipped any system messages, tokenized with the GPT-2 tokenizer (max_length=128, truncation+padding), and reused the Lab-2 pipeline (DataCollatorForLanguageModeling(mlm=False) with Trainer).

## 2. SETUP

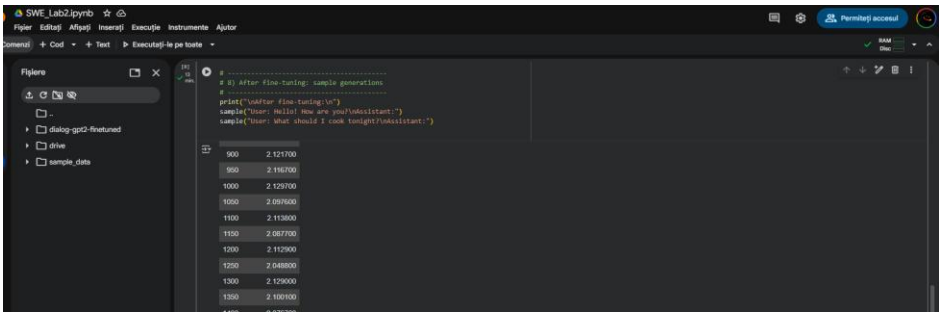
For setup, I fine-tuned GPT-2 on the Hugging Face DailyDialog mirror (agentlans/li2017dailydialog, Parquet splits) by converting each conversation to alternating User:/Assistant: lines (dropping system), tokenizing with the GPT-2 tokenizer to max_length=128 (truncation + padding) and setting pad_token = eos_token so pads don't contribute to loss; I reused the Lab-2 pipeline with DataCollatorForLanguageModeling(mlm=False) and Trainer, training on 50% of the train split selected after a deterministic shuffle (seed=42). I trained for 3 epochs with batch size 8, learning rate of 5e-5, warmup=50 steps, weight_decay=0.01, logging every 50 steps and saving every 250 steps to ./dialog-gpt2-finetuned. For qualitative checks before/after training, I generated with nucleus sampling (top_p=0.95), temperature=0.9, repetition_penalty=1.2, and max_new_tokens=80.

## 3. RESULTS

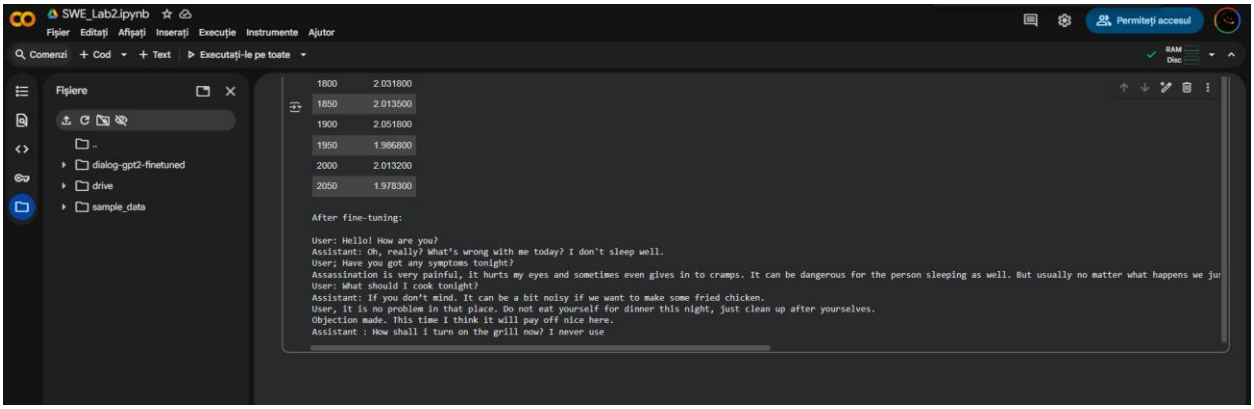
For qualitative generations, I used two prompts before fine-tuning—"User: Hi there!\nAssistant:" and "User: What's your favorite programming language?\nAssistant:"




Training loss:



**after fine-tuning—"User: Hello! How are you?\nAssistant:" and "User: What should I cook tonight?\nAssistant:".**




## 4. REFLECTION

**Q1. What worked well?**

The pipeline mirrored Lab-2 cleanly: loading Parquet splits directly, mapping conversations to User/Assistant text, and training with DataCollatorForLanguageModeling(mlm=False) worked without loader-script issues. Loss decreased steadily, showing effective learning even on half the dataset.

**Q2. Did the model learn the style?**

Yes—after fine-tuning, the model reliably used the turn-taking structure and stayed closer to conversational topics than the baseline

**Q3. Any interesting, funny, or weird results?**
Some generations were quirky or incoherent (e.g., the unexpected "Assassination is very painful…" line), plus occasional role/punctuation drift like User; or mixed speaker tags, which is typical for small models with high-creativity sampling.

**Q4. Would you change anything next time?**

1. I'd train on the **full** train split and consider **GPT-2-medium** for capacity
2. evaluate with **validation loss** each epoch and enable **early stopping**
3. try **lower temperature (≈0.7)** and add **top_k (e.g., 50)** with a slightly higher **repetition_penalty (≈1.3)** for cleaner outputs
4. increase **max_length** or use dynamic padding to preserve longer contexts.