Software engineering hw 1

# 1. Setup & Running

I have used models which are similar in size to gpt2:

* ***distilgpt2: 82M parameters***
* ***EleutherAI/gpt-neo-125M: 125M parameters***
* ***facebook/opt-125m: 125M parameters***

Then I created sets of parameters to test each of them with the 3 models:

* ***temperature=0.3 ; top\_p=0.9 ; top\_k=50 ; repetition\_penalty=1.5***
* ***temperature=0.7 ; top\_p=0.9 ; top\_k=50 ; repetition\_penalty=1.5***
* ***temperature=1.0 ; top\_p=0.9 ; top\_k=50 ; repetition\_penalty=1.5***

The prompt used:

* ***prompt = "Today I learned how to solve differential equations"***

I ran the models through google collab on a personal notebook. The code selects one of the models in the list and loops through the sets of parameters with the same input prompt on each run.

# 2. results

|  |  |  |
| --- | --- | --- |
| **distilgpt2** | **EleutherAI/gpt-neo-125M** | **facebook/opt-125m** |
|  |  |  |

# 3. Interpretations of results

Across all three small models, I saw a pretty consistent pattern with temperature. **distilgpt2** was the steadiest: at **T=0.3** it didn’t stick to the language it mentioned first, **T=0.7** gave a generic sentence without specifying the language, and **T=1.0** started to ramble. **EleutherAI/gpt-neo-125M** felt more “codey”—**0.3** but it was repetitive, **0.7** produced the best mix of concrete details (e.g., plausible class/function names) and coherence, and **1.0** often drifted into half-finished snippets. **facebook/opt-125m** had a chatty vibe; **0.3** read like forum comments, **0.7** was the most useful with step-like guidance, and **1.0** became disjoint quickly. **Overall, the best temperature across models was ~0.7**, which gave the strongest balance of specificity and coherence without the chaos I saw at 1.0 or the repetitiveness and incoherence at 0.3.