Software engineering hw 1

# 1. Setup

**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*

# 2. Code

def run\_model(model\_name, params, max\_new\_tokens):

    tokenizer = AutoTokenizer.from\_pretrained(model\_name)

    model = AutoModelForCausalLM.from\_pretrained(model\_name)

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

    inputs = tokenizer(prompt, return\_tensors="pt")

    output = model.generate(\*\*inputs,

      max\_new\_tokens=max\_new\_tokens,

      temperature=params["temperature"],

      top\_k=params["top\_k"],

      top\_p=params["top\_p"],

      repetition\_penalty=params["repetition\_penalty"],

      do\_sample=True

    )

    return tokenizer.decode(output[0], skip\_special\_tokens=True)