Assignment 1

Title: Chat GPT-2 Temperature Analysis

Name: Brennen Cramp

Date: 1/22/2026

Introduction

A Large Language Model (LLM) is an advanced generative AI model trained on massive text datasets to understand, generate, and process human language. LLMs act like a sophisticated generator that will predict the next word in a sequence, creating text for tasks such as answering questions, writing, and summarizing. The goal of this experiment is to observe and develop an understanding of how an LLM's temperature parameter alters the confidence of the model and impacts the logical coherence of its output. For the experimentation of the temperature parameters in an LLM, the LLM_Text_Generator.ipynb code file will be updated to examine the effects of changing the temperature whilst keeping the top k values and tokens generated static.

Temperature Analysis

The temperature value is introduced into the LLM's softmax function which will influence the sampling process. If the temperature is low, the probabilities look more like a max value instead of a "soft-max" value. In contrast, if the temperature is high, the probabilities begin to adopt a more uniform distribution (the sampling process may select any token), making the generation process random and heavily stochastic. For this experiment, other parameters will be the control variables such as top k = 40 and the tokens generated = 20. As seen in *Table 1*, there will be three temperature values tested (0.1, 0.8, 2.0) where the predicted behavior will range due to the increasing temperature values. The prompt that will be leveraged for the experiment is "I want to travel the world. I will go" and the model responses seen in *Table 1* were what was returned based on the varying temperatures along with the coherence of the responses based on if the response follows the rules of English and logic.

Trial	Temperature (T)	Predicted Behavior	Model Response	Model Coherence (1-10)
A	0.1	Conservative	"to the places where I want to go. I will go to places where I want to go."	6
B	0.8	Creative	"to countries that are far from me. And I am going to be there. I will see how"	8
C	2.0	Chaos	"to India to buy coffee at an American retail chain (Bacosh) from a Mexican shop for"	4

Table 1: Varying Softmax Temperatures and the Model's Responses

Looking at the model response where the temperature was 0.1 and was expected to have a conservative behavior in the model response to the prompt, the response was too concerned with not being adventurous that it repeated the "I will go to places where I want to go." phrase. While the English was not bad, the model seemed to lose logic and just made disjunct and unamusing sentences once it believed it found the right answer. The response seemed obvious and matter of fact saying that it would want to travel to the places it would like to travel to. However, the model did produce all real words with no random characters nor punctuation which when the temperature value was low (0.1), it changed the probability distribution to simply pick the highest percentage after converting the raw scores (logits) into percentages of predictability; thus picking the most obvious words. This model coherence was estimated to be a 6 since there were no vocabulary issues but the entire response just repeated the same sentence it first came up with.

The model response where the temperature was 0.8 was expected to have a more creative approach to the prompt which unlike the 0.1 temperature, there were no repeated phrases even though the "And I am going to be there" phrase essentially had the same meaning as the first part of the response. The model's response had good English with a slight disconnect by saying the same sentence but in a different manor but if there were more than 20 tokens, it seems that it may have elaborated on where it would like to travel to. In the same fashion as the more conservative response, this creative response from the model produced all real words with no random characters nor punctuation which having the temperature value be relatively high (0.8), it changed the probability distribution to roughly experiment with words making the model seem more like it wants to have a conversation and elaborate on where it wishes to travel. This model coherence had a higher estimate with an 8 since there were no vocabulary issues and it seemed that if given the opportunity to increase the size of the output, it would explain where and why it would like to travel to the places it wishes.

In contrast to the previous experimentations with temperature, adjusting this value to 2.0 caused the response to instantly include places it would like to travel to and things it wishes to do when there. It plans to "go to India to buy coffee at an American retail chain (Bacosh) from a Mexican shop" which makes sense until the chaos that is built into a temperature value of 2.0 shows since it lists two places it is buying the coffee from, assuming that the American retail chain is also a Mexican shop. In addition to that logic error, the English that the model produced messed up for the first time since "Bacosh" is not a real store nor chain nor even a real word. Having the temperature value be extremely high (2.0) changed the probability distribution to vastly experiment with words (even making one up), making every word equally likely allowing for utter chaos despite mentioning a new place it wishes to travel to, unlike the previous experiments. This model coherence had the lowest estimate with a 4 since while it did mention places it would like to go to (India), the logic as to where it would buy the coffee from is an American retail chain which is also a Mexican store and the response included a non-existent word/store it thought that it was referencing.

In real world usage, if a medical AI was wanted in a hospital to help assist the doctors and nurses on how to give prescriptions or advice about a patient's condition, the temperature that should be utilized should be a range of 0.4-0.7, allowing for a balanced response that is not too chatty and gets more precise and accurate answers. These types of responses are wanted in the medical field since it will help the doctor/nurse to get more fact-based and expected responses to help their patients, especially in a quick fashion to deliver fast and reliable solutions for whatever ailment the patient is under. Values ranging past 0.7 would begin to feel "chattier," making the response longer and expansive instead of quick solutions while values going below 0.4 might make the same recommendation for every patient despite different conditions patients may be experiencing with their ailment which means the model should allow for less conservative answers. In contrast, if an AI was being built to help write a surrealist dream-journal, the author may want more chaotic and physics-altering responses to fit the theme of the novel so the temperature would be around the 2.0 value. This disorder would provide an interesting and entertaining read that could push the reader's minds with phrases and logic they may have never heard before which further cement the surrealism of the world built by the author.

Conclusion

In essence, the temperature is a vital parameter when used in conjunction with the softmax functions used in LLMs. If the temperature is a low value, the creativity of the responses generated would be more matter of fact and obvious, being more conservative and not straying from what the highest predicted next word should be. If the temperature is a high value, the responses seem to be more conversational and creative, acting as if it is an informal chat between two humans. However, if the temperature strays too high and begins to approach the 2.0 value, chaos will ensue in the responses produced, giving more absurd and non-sensical responses which can allow for words that do not exist. Fully committing to either end of the temperature spectrum should be utilized based on the application/environment for which the LLM will reside in and the temperature is always subject to change due to this issue and to accommodate for the desired users.