Assignment 1 Extra Credit

Title: Chat GPT-2 vs DistilGPT2 Temperature Analysis

Name: Brennen Cramp

Date: 1/22/2026

Extra Credit

The prompt is the same as utilized from the first part of the assignment: “I want to travel the world. I will go”.

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 with Chat GPT-2

Trial	**Temperature (T)**	**Predicted Behavior**	**Model Response**	**Model Coherence (1-10)**
A	0.1	Conservative	“to the United States and I will go to the United Kingdom and I will go to the United States”	6
B	0.8	Creative	“to the Olympics in Rio, and I will go to the Olympics in San Francisco. I'll see”	7
C	2.0	Chaos	“about it a day and travel every single day and that alone was exactly, well....I really have”	2

Table 2: Varying Softmax Temperatures and the Model’s Responses with DistilGPT2

After swapping the “gpt2” model being used with “distilgpt2”, the time it took to generate the responses for “distilgpt2” (Chat GPT 2 utilizing the distillation technique) was almost three times quicker than it took for “gpt2” to generate responses (average “gpt2” response time: 17

seconds; average “distilgpt2” response time: 6 seconds). In addition, the responses that “distilgpt2” created mentioned more places however the logic seemed to repeat itself at some times, similarly to the “gpt2” model, at a temperature of 0.1 since it reiterated that it would go to the United States; thus “distilgpt2” receiving the same model coherence of 6 as the “gpt2” model with logic being its downside. When conservative values are utilized for temperature, it illustrates that it will always go for the highest percentage for the next token and its context window seems to be incredibly short for both models. For a temperature value of 0.8, the “distilgpt2” model provided a more concrete answer with a theme of visiting the Olympics in the places it wanted to travel to, however logic was its downside again in that the games in Rio already happened ten years ago and San Francisco will not be holding the next Olympics in 2028, it will be in Los Angeles which is a different city in California. This dropped the model coherence to 7 instead of the “gpt2” model’s 8 model coherence. In contrast, a huge difference was spotted with a temperature of 2.0 in that the “distilgpt2” model’s response was all over the place and felt like a jumble of words, even adding an ellipsis where the time to produce the next token took around 3 seconds as if the model was genuinely confused on how to respond after its previous statement of nothingness. This caused the model coherence to go from the top of poor to just very incoherent like the “distilgpt2” seemed to barely make a thought as to where it would like to go or even what it would like to do when travelling. The “gpt2” model, as expected due to its larger state with 12 layers and 124 million parameters, was still able to produce slightly better responses despite its token limit unlike the “distilgpt2” model’s small size with only 6 layers but also providing a faster response rate than “gpt2”. The lack of layers for the “distilgpt2” model caused the logic in the output to be hindered and limited to not finding better solutions for the next token to be added to the response.