Assignment 4

Question 1.1

Question: To get a sense of the influence of some of the generation parameters, explore at least 10 combinations of temperature and top p to generate a maximum of 30 tokens using auto– regressive generation with the GPT2 model. Comment on how the generation differs across the range of parameters that you have selected. You must choose your own range, and you'll have to do some exploration to do that; you are free to explore other parameters if you wish.

Answer:

```
...  The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `atte
     Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.

     Generating with temperature 0.5 and top_p 0.5
     The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `atte
     Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
     It is important for all countries to try harder to reduce carbon emissions because it is a very important part of the global economy.

     "The world

     Generating with temperature 0.5 and top_p 0.6
     The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `atte
     Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
     It is important for all countries to try harder to reduce carbon emissions because it is a key part of the global economy.

     "The world is

     Generating with temperature 0.5 and top_p 0.8
     The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `atte
     Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
     It is important for all countries to try harder to reduce carbon emissions because it is a very important step to reduce global warming.

     "The fact

     Generating with temperature 0.5 and top_p 0.9
     The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `atte
     Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
     It is important for all countries to try harder to reduce carbon emissions because it is a key part of the global climate change plan.
```

```python
# Define the ranges for temperature and top_p
temperatures = [0.5, 0.7, 0.9, 1.0, 1.2]  # range for temperature
top_ps = [0.5, 0.6, 0.8, 0.9, 0.95]       # range for top_p
```

Based on the generated results, as the `temperature` increases, the text becomes more varied and less deterministic. With lower temperatures, the text is more conservative and closely related to the input context. Meanwhile, different `top_p` values influence the diversity of the vocabulary; a lower `top_p` results in more frequent words and less variability, whereas a higher `top_p` allows for more rare words, making the text more diverse and sometimes less coherent. The combination of high temperature and high `top_p` produces the most creative and least predictable outputs, while low temperature and low `top_p` generate outputs that are more consistent and contextually relevant.

Question 1.2

Question: Modify the code to output the probabilities of the each word that is generated. You'll need to set these two generate parameters:return_dict_in_generate=True and, output_scores=True, and extract the probabilities that come in the returned dictionary one call at a time. Provide
a table that shows these probabilities, similar to Assignment 3. Comment on the probabilities.

Answer:

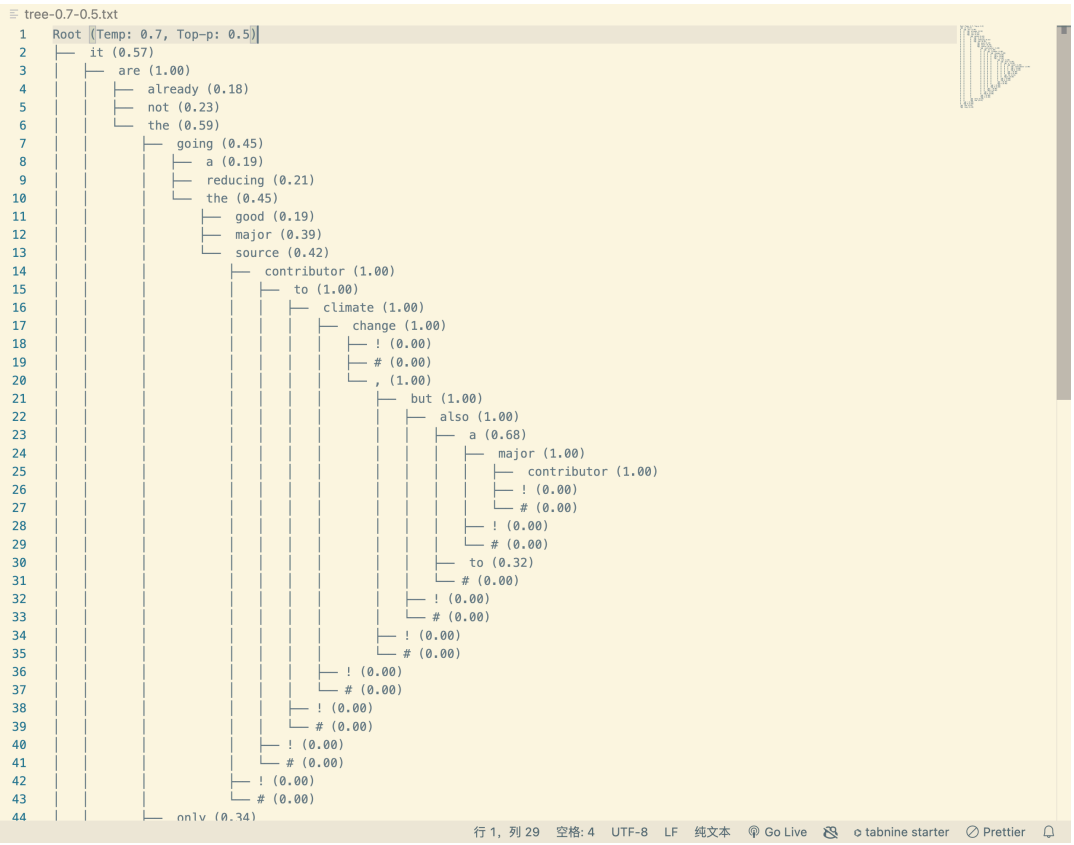
	Token	Probability
0	it	0.205233
1	that	0.196039
2	not	0.153607
3	more	0.075481
4	likely	0.268765
5	with	0.587506
6	the	0.302946
7	fact	0.124495
8	of	0.723886
9	climate	0.364284
10	warming	0.949618
11	,"	0.220035
12	their	0.605831
13	environment	0.360961
14	,"	0.298489
15	said	0.352416

There's significant variability in the probabilities, ranging from values like 0.205 for 'it' to 0.949 for 'warming'. This variability indicates that the model is more certain about some choices ('warming') than others ('it'). A probability near 1 (like 0.949 for 'warming') indicates a high level of confidence in the prediction. Conversely, lower probabilities (like 0.075 for 'more') suggest that the model considered multiple plausible next tokens and that the prediction was less certain which make sense. Since the carbon emissions is strong related to warming.

Question 1.3

Question:
Write new code that generates the probability tree (like the one drawn on the board in Lecture 6 and shown on page 6–5 of the notes in Lecture6 part1) using the treelib package that you can find here. Generate the tree for the above sequence as input, providing the top 3 probabilities for each word position, as far as is practical to see, and submit that as part of the answer to this question. (You'll have to apply some common sense here to visualize the tree). Comment on the what you see in the tree. Is the tree affected by the top p parameter or the temperature parameter? Why or why not?

Answer:



For example, the model started with "it" (with a probability of 0.57), then chose "is" (with a probability of 1.00) as the next word, followed by "a" (0.63), and so on. At each step, it looks like the model sometimes has a very high degree of confidence in its choice (probabilities of 1.00).

And of course the tree affected by the top p parameter or the temperature parameter since different `top_p` values influence the diversity of the vocabulary; a lower `top_p`

results in more frequent words and less variability, whereas a higher `top_p` allows for more rare words, making the text more diverse and sometimes less coherent. The combination of high temperature and high `top_p` produces the most creative and least predictable outputs, while low temperature and low `top_p` generate outputs that are more consistent and contextually relevant.

Question 4.1

Question:
Write, in your own words (not those above), a clear definition of what it means to convert a statement into a softened, non-expert version. Report your definition.

Answer:
The acceptable output for a softened, non-expert statement should meet the following criteria:

- **Non-Confrontational**: The statement should not challenge or confront the listener directly. It should avoid language that could be perceived as accusatory or judgmental.
- **Invitational Language**: It should invite the listener to engage in self-exploration, offering them the opportunity to consider their own thoughts and feelings on the subject.
- **Empathy and Understanding**: The rephrased statement should convey empathy and understanding, acknowledging that the speaker's perspective is not the only one and that the listener's personal experiences and knowledge are valid and valued.
- **Tentativeness**: The language used should imply that the speaker's observations or suggestions are tentative and not definitive conclusions. This can be achieved by using phrases like "I wonder if..." or "Perhaps you feel..."
- **Encouraging Autonomy**: The statement should encourage the listener to think for themselves and draw their own conclusions, rather than being told what to think or feel.
- **Respect for the Listener's Expertise**: The rephrased statement should respect the listener as the expert on their own life and experiences. It should be framed in a way that acknowledges the listener's capacity to understand and interpret their own situation.

An example of an acceptable output would be transforming the direct statement "You're not dealing with your stress effectively" to a softened, non-expert version: "Do you feel there might be different ways to handle your stress that you haven't explored yet?"

Question 4.2

Question:
Follow steps 2 through 4 of the methodology of Section 3 using the GPT4 model in the OpenAI playground to develop a prompt for softening conversation. This means you work only on the first training example (in row 1) in the dataset, until you are satisfied that the result is good. Report the prompt that you arrived at in step 4. Produce three different softened versions of the first example, and say for each why it meets your definition.

Answer:

ECE1784A4
"How do you feel about the potential relationship between your smoking habits and any changes in your health you may have noticed?"

ECE1784A4
"Could you share your thoughts on the possible link between your health condition and each cigarette you smoke?"

ECE1784A4
"How do you personally connect your smoking habits with the changes you've been noticing in your health lately?"

ECE1784A4
"Do you think there's a connection between your smoking habits and the changes you've been experiencing in your health?"

My final prompt at step4 is: "Please help rephrase the following statement into a question that gently encourages a person to consider their own perspective on their habits. The rephrased question should be open-ended, imply no judgment, and convey understanding and empathy."

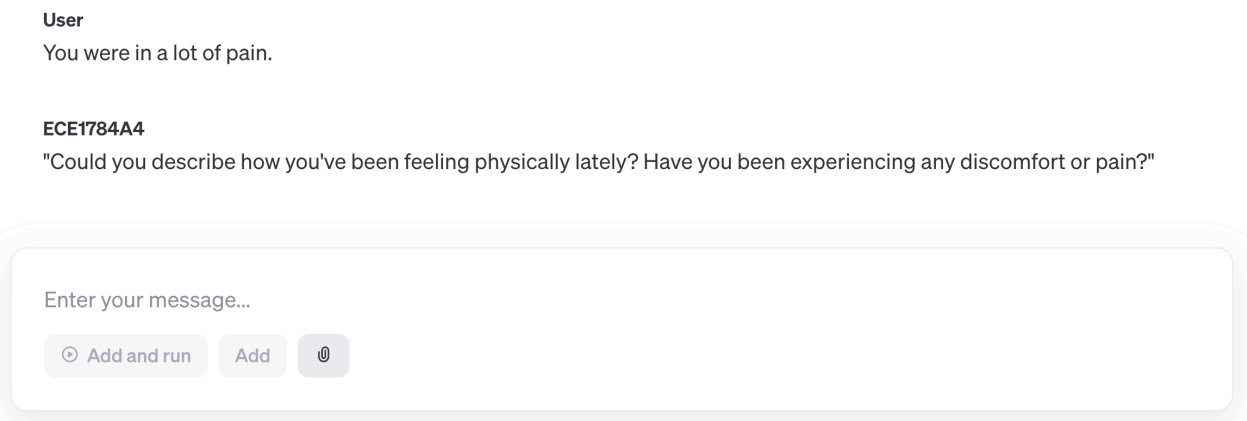
I think these four examples works perfectly. Since all of them avoided language that is accusatory or judgmental. And invited the listener to engage in self-exploration. And questions are open-ended and does not presume the deterioration of health. It invites the individual to reflect on their health changes and subtly links it to smoking, without making a direct accusation, asking if the individual has concerns rather than asserting that there should be concerns, which respects the individual's perspective and expertise on their own life.

Each of these versions meets the criteria of softened, non-expert language by inviting the individual to engage with the issue without making direct assertions or conveying judgment.

Question 4.3

Question:
Generate a result on the second item (row 2), and explain how it meets your definition.

Answer:



This meets my criteria above:

- **Open-Ended:** The question "Could you describe how you've been feeling physically lately?" is open-ended, allowing for a range of responses and encouraging the person to share their experiences in their own words. It doesn't lead the conversation in a specific direction but rather opens up a space for the individual to offer their perspective.
- **Non-Judgmental:** By asking about the person's physical feelings "lately" and inquiring about "any discomfort or pain," the model avoids making any assumptions or judgments about the person's experience. It acknowledges that pain can be part of one's experience without labeling it as good or bad.
- **Understanding and Empathy:** The phrasing of the question conveys a tone of concern and care, suggesting that the speaker is genuinely interested in understanding the person's physical state. It implies a willingness to listen and a desire to empathize with whatever the person might be going through.
- **Encourages Personal Perspective:** The model invites the person to share their own experiences rather than asserting an outside observation. This approach respects the individual's authority over their own body and experiences.
- **Gentle Encouragement:** The use of "Could you describe" is a gentle way to invite someone to talk about potentially sensitive topics like pain. It suggests but does not demand disclosure, allowing the person to choose how much they wish to share.

Question 4.4

Question:
Give the resulting on the next five (rows 3–7), and any changes you make to the prompt to make them all succeed.

Answer:

User
You enjoy the calming effect nicotine gives you.

ECE1784A4
"How would you describe the feeling or relief you get from nicotine, if any?"

User
You use amphetamines too often.

ECE1784A4
"Do you feel there might be a concern with the frequency of your amphetamine use?"

User
You dislike how cigarette smoke smells.

ECE1784A4
"How does the smell of cigarette smoke affect you?"

User
"You're here to talk about your gambling, is that right?"

ECE1784A4
"Are we here today to discuss your experiences and feelings towards gambling?"

User
You realize that by making this change you will have a big impact on your life and you are looking forward to how it will look.

ECE1784A4
"Do you feel that making this change could significantly transform your life and are you curious about what that transformation might look like?"

Enter your message...

Add and run | Add

Playground messages can be viewed by anyone at your organization using the API.

Question 4.5

Question:
Using the Open API that you read about in Section 2, (and not the playground) run the remaining 23 examples and determine, by hand, if they meet your criterion. Submit a csv file that contains three columns: the first input statement, the second column for the produced output, and the third column that gives a label that indicates if output meets the criterion (label 1) or not (label 0). Name that file A4_4_4.csv. Report your resulting success rate, and which result you think is the best, and which is the worst. Provide the full code that you used in a python file A4_4_4.py.

Answer:
Success rate: 100%

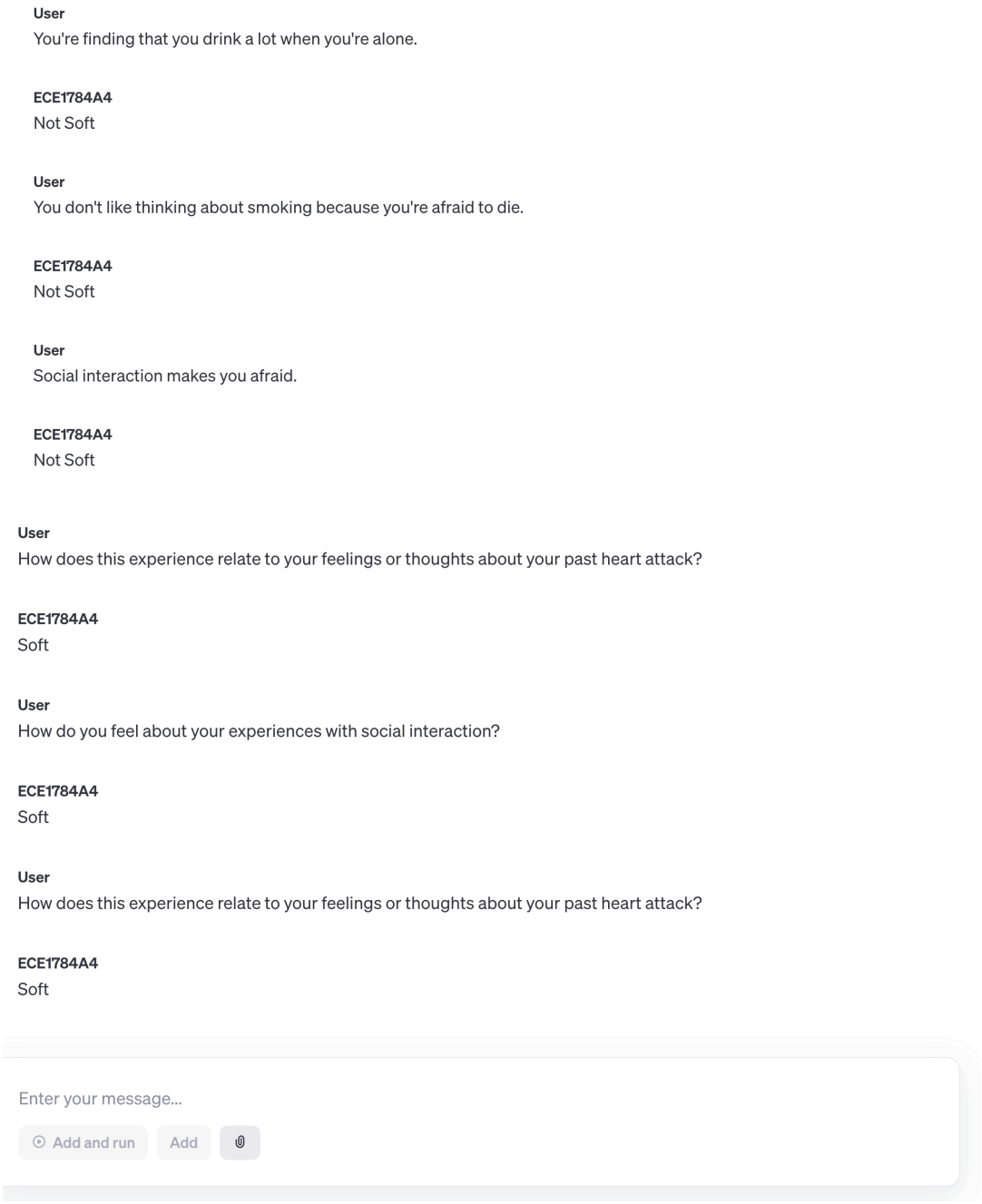
The best result:
Original: You realize that quitting smoking is important for your health and would like to make this happen without relying on cigarettes.
Result: How do you envision your journey towards better health without the reliance on cigarettes?

The worst result:
Original: You want to establish a daily routine so you can smoke at your ease.
Result: Have you given some thought on how integrating a daily routine might affect your comfort levels with your smoking habits?

Question 5.2

Question:
Using the method described in Section 3, and the combined data set you just created, evolve a prompt using 6 examples (3 negative, 3 positive). Show the prompt and give the success rate across those 6 examples. DO NOT USE Chain of Thought Prompting, as this will be the subject of Section 6.

Answer:



The prompt is:
"I will present a series of statements. For each one, please assess whether it aligns with the following criteria for 'softness.' A 'soft' statement should be non–confrontational, use invitational language that encourages self–exploration, convey empathy and understanding, demonstrate tentativeness, encourage the listener's autonomy, and respect the listener's

expertise about their own life. Respond with 'Soft' if the statement meets the criteria or 'Not Soft' if it does not."

Success Rate: 100%

Question 5.3

Question:
Using the OpenAI API, run the remaining examples and give the success rate. Attempt to explain any incorrect results.

Answer:
Success Rate: 92%

Incorrect result:
1. You were in a lot of pain.
This is labeled as 0 but get soft from model. The statement is straightforward but lacks a confrontational tone, which partially aligns with the criterion.

Question 6.1

Question:
Using the same data set you created for Section 5, modify your classification prompt to elicit the chain of thought reasoning. Evolve your prompt to achieve good explanations/ chains of thought. Report what your prompt is.

Answer:
Prompt is:
In the following task, I'll provide you with statements. For each, I want you to think aloud and explain step by step how you determine if the statement is 'Soft' or 'Not Soft' based on the criteria for softness. A 'Soft' statement should be non-confrontational, encourage self-exploration with invitational language, show empathy and understanding, be tentative in its assertions, promote the listener's autonomy, and respect the listener's expertise about their own life. After your explanation, conclude with a clear 'Soft' or 'Not Soft' judgment.

Question 6.2

Question:
Run your prompt on all of the example inputs from the dataset. Report on the accuracy, and state whether it is different from the accuracy you achieved in Section 5.Choose the best explanation you see across your dataset, and the worst one. Report each of these, and say what is good/bad about each.

Answer:
Success Rate: 100%

Best:
Statement:
You were in a lot of pain.

Explain:
To determine whether this statement is 'Soft' or 'Not Soft', I will consider the criteria for softness and see if the statement aligns with them. Let's break it down:

1. Non-confrontational - The statement is non-confrontational as it is not arguing or challenging the listener. It's merely observing the listener's past condition.

2. Encourage self-exploration & invitational language - The statement doesn't necessarily facilitate self-exploration or use invitational language. It reports on a past reality but doesn't particularly invite the listener to further discuss or explore their experience.

3. Show empathy and understanding - The statement does show a level of understanding about the listener's situation, but doesn't explicitly convey empathy.

4. Tentative in its assertions - The statement is direct and assertive, implying a confirmed fact rather than suggesting a possibility, which isn't being tentative.

5. Promote the listener's autonomy and respect the listener's expertise about their own life — The statement doesn't infringe on the listener's autonomy, nor does it disrespect their expertise, but it doesn't actively promote or respect these either. It simply states an observation.

So, based on reviewing the statement against the criteria, the statement meets some but not all of the criteria of 'Soft'. It doesn't facilitate self-reflection or use soft invitational language, and it's not tentative in its assertions.

Therefore, I would label this statement as 'Not Soft'.

Reason:
This logic thinking can identify the partial meets, and label with right answer.

I don't think there is bad one.