Fine-Tuning GPT-2: An Exploration into the Efficacy And Practicality of Low-Rank Adaptation and Joint Training for Altered Sentence Structure

Jonathon Webster, Nathaniel Alexis, Kaushik Vemparala, Rayan Taghizadeh

Abstract

This paper explores effective fine-tuning of GPT-2 via LoRA to adapt to specific syntactic structures—in this case, talking like Yoda. The hyperparameters swept through this process include LoRA rank, target modules, and utilizing joint training. In addition to the standard loss function, a perplexity score was implemented to measure the quality of generated sentences. We discovered that varying ranks used in LoRA do not make a relatively substantial impact on responses and their perplexity scores, target modules of self-attention and attention-projection are the most influential towards training, and joint training yields good results.

1 Motivation

Language structure is always evolving: modern English, for example, is a culmination of prior phases of differing vernacular and sentence structure. Languages will continue to evolve as time progresses, potentially leading to issues employing LLMs trained on out-of-date language syntax. If somebody in the year 3023 were to make use of an LLM trained today rather than training a new one, there would be a major problem—English will have changed to an extent to which the structure of the words generated by 2023-trained LLM would be hard to understand. Could they instead simply fine-tune the outdated model to better match how they speak in 3023 and continue their project? We answer this question by providing a method for fine-tuning LLMs based on syntactical nuances. This work extends to language translation, where transforming a sentence to the proper structure before translating words is vital.

2 Related Works

Our approach builds on the methodology of "Low-Rank Adaptation for Large Language Models" by Hu et al., available at https://arxiv.org/abs/2106.09685.

3 Objective

In order to advance our understanding and methodology for the anticipated change in language structure, we plan to fine-tune a GPT-2 model to speak like Yoda. Our question: *Can we fine-tune a model to completely change sentence structure rules to one that resembles Yoda? If so, what is the most successful approach?*

4 Approach

We prompted the LLM to give us facts related to random queries and respond like Yoda would. An example of a line of training data is below:

```
"User: Tell me a fact about Aardvarks.
Yoda: Excellent diggers, aardvarks are."
```

We used the Low Rank Adaptation method for fine-tuning and investigated which parameters out of the following options gave us the best results, validated by perplexity score:

- Varying LoRA rank [4, 8, 16, 32, 64]
- Varying target modules: [Token Embeddings, Positional Encodings, Self-Attention, Attention Projection, Self-Attention + Attention Projection]
- LoRA with Joint Training: Instead of just having a "Yoda:" prompt, we train the model concurrently with a "Human:" prompt representing plain English.

4.0.1 Model we Use

We are using the small GPT2 model[9] found on Hugging Face. This model is trained on raw text and has 124 million parameters. The loss function we are using is the default - negative log likelihood.

4.0.2 Evaluation Metric

Our model uses perplexity, a common LLM sequence metric, as the method for evaluating generated responses. Perplexity is defined as

$$\sqrt[n]{\frac{1}{P(\text{Test Sentence})}}$$

where $P(\text{Test Sentence}) = P(w_1) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_1, w_2) \cdots P(w_n \mid w_1, \ldots, w_{n-1})$. A high perplexity score is bad, indicating that the input test sentence was not highly probable, and thus "perplexing" the model. The root n in the metric—the geometric mean—is used to normalize the multiplication of probabilities. After reading through peer reviews, we looked into other NLP evaluation metrics such as BLEU and METEOR. Some downsides of these alternatives are that BLEU does not take word order into account[7], so while this might be a valuable metric to explore in the future, we needed to emphasize syntax and word order in this context. METEOR is another possible option, but as it is more tailored to translation application similar to BLEU, we just decided to stick with perplexity and perhaps explore these alternative evaluation metrics in the future due to time constraints.

5 Dataset Creation

In order to translate a normal sentence to a "Yoda" sentence, we first tried to use an API call, but we were limited to 10 API calls per hour, and this would not be enough to create extensive training data. Instead, we found some baseline code that would translate ourselves[4]. After making changes to the initially faulty code, we utilized it to create training and validation sets. We used our translator on the generics_kb dataset on HuggingFace to synthesize the following datasets:

Dataset Name	Example Line	Purpose
train_data_final.txt	User: Tell me about price. Yoda: Subject to applicable taxes, prices are. <|endoftext|>	Our goal is to train our model with this data in order to be able to prompt the GPT model to talk like Yoda.
val_data_medium.txt	Same formatting as above	This set was used for validation during training.
yoda_human.txt	User: Tell me about wisdom. Yoda: Right understanding, wisdom means. <|endoftext|> User: Tell me about dive site. Human: Dive sites are numerous and visibility is normally very good. <|endoftext|>	This training set is to explore the effects of joint training with both human and yoda prompting on the model.
test_data_final.txt	User: Tell me about dinner.	This dataset is used to first have our model run inference, and then calculate the perplexity of it's response.

Figure 1: Table 2

In response to our review about our data creation method – our code follows the algorithm needed to turn a sentence into a Yoda sentence. The sentences that happened to break our algorithm also broke the algorithms for other Yoda translator websites. In addition, we added code to the already existing algorithm to extend it's abilities and better handle commas as well as words like "and". While some sentences it was trained with were not technically "yoda sentences", the sheer amount of sentences turn these irregularities into noise. I do not think it would be particularly beneficial to extract specific yoda sentences from movies as there are very few and the vast majority of our sentences already follow yoda's sentence structure.

6 Experiments

Experiments were conducted to explore and reason about the impact of varying LoRA parameters and joint training.

6.1 LoRA Rank Parameter Sweep

6.1.1 Context

An advantage to LoRA is the ability to freeze the model weights, W, as opposed to training a ΔW that we can add to the actual weights. The reason this is helpful is as follows: Assume we have a 1000 by 1000 weight matrix W. Training every weight could take a long time, but LoRA allows us to instead train $A = 1000 * R$ and $B = R * 1000$ weight matrices (R being the rank we are sweeping). Matrix multiplication yields $\Delta W = AB$, showing that we only have to train $1000 * 2 * R$ weights, providing speedups assuming R is less than 500. Naturally, a smaller R-value yields fewer trainable weights, and in this section we explore the benefits of different R-values while keeping in mind the accuracy-trainable parameter trade-off.

6.1.2 Experiment

To observe the impact of the LoRA rank on the performance of the tuned model, we swept over rank values of 4, 8, 16, 32, and 64. The metrics we used to evaluate the rank's impact were evaluation loss and test data perplexity. We used a seed of 42 for training, and batch sizes of 4.

6.1.3 Results

Our experiments indicated that the rank had negligible impact on the training evaluation losses (as seen in figures 2 and 3, respectively), and an unclear, albeit slightly larger, effect on the perplexities. We can see in figure 4 that from our runs, using LoRA with a rank of 16 yielded the lowest perplexities, indicating that for a syntactic restructuring task, the finest granularity does not always lead to the best results. Sample responses of the models at different stages of training can be found in figure 11.

6.1.4 Discussion

This concept is illuminated when considering the case of representing an image with PCA–likely only a few principal components capture a large amount of information about the image, while the rest combine for a much smaller portion. In our case, only a few "principal components" add enough value towards fine-tuning our data for syntax restructuring to justify the added computing cost of extra ranks. This idea is also supported in the original LoRA paper where the authors state that LoRA already performs competitively at a small rank, and they even argue that increasing the rank does not cover a more meaningful subspace, suggesting that a low rank adaptation matrix is sufficient.

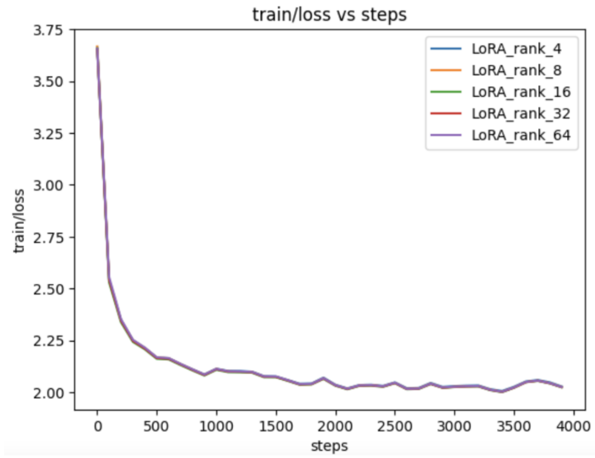

Figure 2: Training losses vs steps when fine-tuning with different ranks.

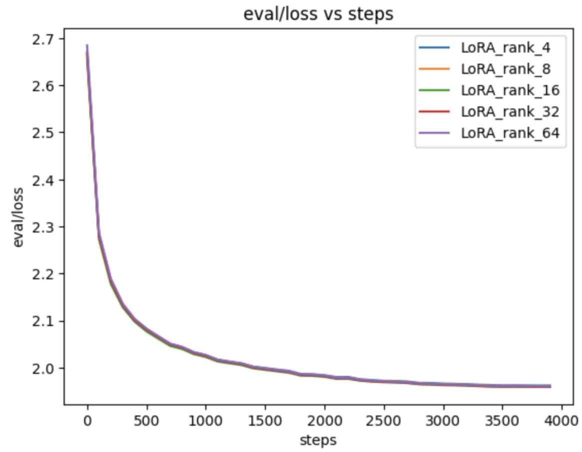

Figure 3: Evaluation losses vs steps when fine-tuning with different ranks.

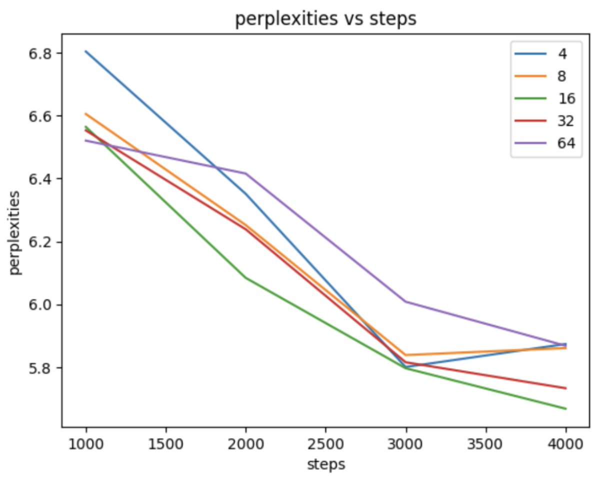

Figure 4: Perplexities vs steps on models fine-tuned with different ranks.

6.2 Target Module Sweep

6.2.1 Key Citation:

We gathered most of our knowledge on target modules from the article: https://jalammar.github.io/illustrated-gpt2/.

6.2.2 Target Modules:

- Token Embeddings (wte): the matrix used to turn a word into a vector.
- Positional Encodings (wpe): a matrix containing an encoding for different positions in the input.
- Self-Attention (c_attn): a matrix of weights that produces the queries, keys, and values for each input of the decoder.
- Attention Projection (c_proj): when doing self-attention in GPT2, it actually has 12 self attention heads. This means that we will have 12 key, query, and value vectors as output from the self-attention matrix multiplied by the input. The attention projection matrix projects the concatenation of the results into something that can be fed as input to the next layer.

6.2.3 Results

Our plots indicated that using LoRA on the token embedding and position encodings yields much worse performance—in fact, judging by the provided examples, they do not offer any added value. We can see this in plots of training and evaluation loss (figures 5 and 6, respectively), and perplexities (figure 7). Sample responses of the models at different stages of training can be found in figure 12.

6.2.4 Discussion

Positional encoding and token embedding both performed poorly in perplexity as well as qualitatively. One interesting thing to note is that at times, they were able to learn the response prompt similar to having "Yoda:", but with completely unrelated names in completely meaningless sentences. These layers are not used in LoRA from all papers/projects we have read – and for good reason. These layers are not repeated in the GPT model like the self attention and projection layers are, so they are not as influential throughout the entire model. The projection and self-attention weights were by far the best at changing syntax structure, with a combination of the two performing the best, as shown by the perplexity of the two being better than the individuals. This does not come as a surprise considering these are both present in each decoder in the gpt model and account for a lot of the weights. In addition, it is common practice to exercise LoRA on the query, value, and key matrices of weights, and these are all included in the attention target module. In the original paper on attention, the authors state that self-attention has been used successfully in textual entailment recognition, which is the relationship between two fragments of text when one implies the other. This is a relationship that is ubiquitous in Yoda's speech, and so textual entailment recognition is something that is very important when transforming sentences in to Yoda's unique structure. All in all, if one is resource-deficient with the goal of altering sentence structure, it is best to pick the smallest, in magnitude, of the projection and self-attention weights to target when fine-tuning.

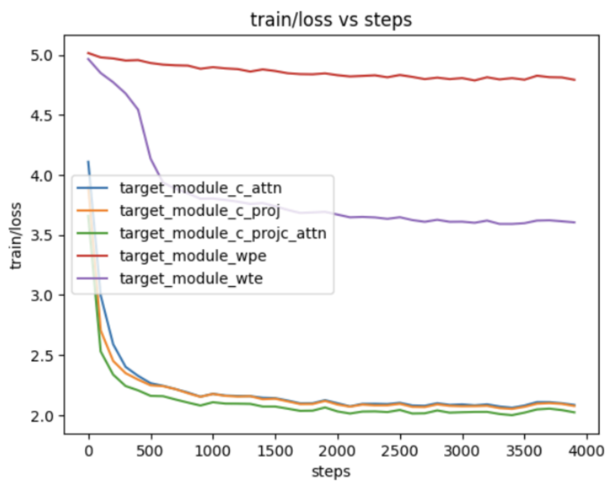

Figure 5: Training losses vs steps when targeting different modules during fine-tuning.

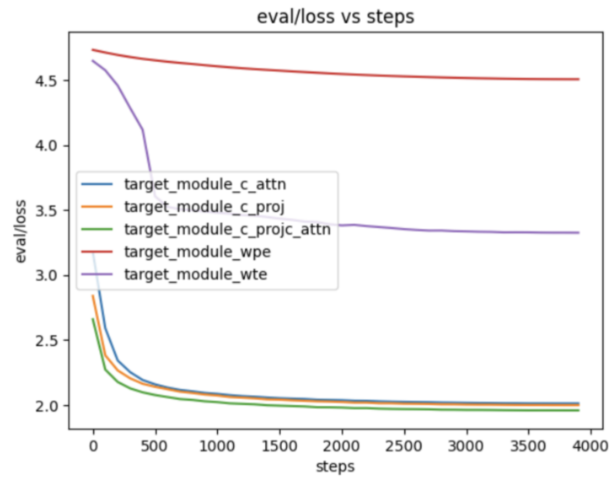

Figure 6: Evaluation losses vs steps when targeting different modules during fine-tuning.

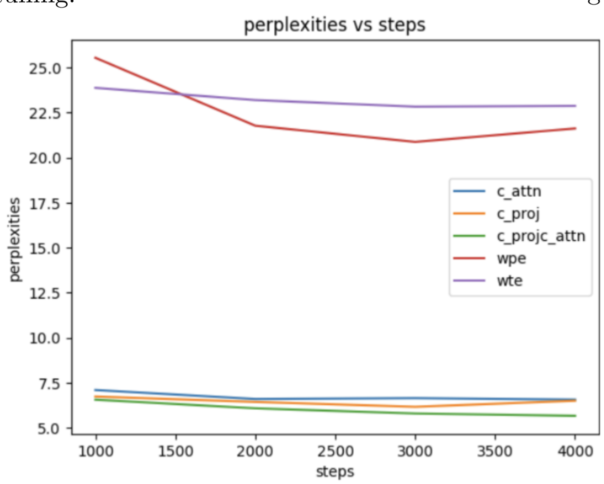

Figure 7: Perplexities vs steps of models fine-tuned with LoRA targeting different modules.

6.3 Joint Training

6.3.1 Context

Joint training is method often used when one wants their model to still "remember" some of the data it was originally tuned with, thus avoiding catastrophic forgetting. Here we explore how our fine-tuning procedure is influenced by interweaving "Human" responses with "Yoda" responses, and how our fine-tuned model is changed.

6.3.2 Results and Discussion

The lower perplexity score for the Yoda data vs. human data came as a surprise, but our hypothesis is that the GPT-2 model already has a clear picture of what human language sounds like. In discussion 12, we learned that joint training should still yield good results for new data as well as for the old data. Since we split 50/50 Yoda and human sentences, we ran for twice as long. While this did take more compute resources (as expected) we still got basically the same Yoda perplexities as our original model. It also performs well for both cases qualitatively. Overall, our findings align with what was taught in discussion. Sample responses of the models at different stages of training can be found in figure 13.

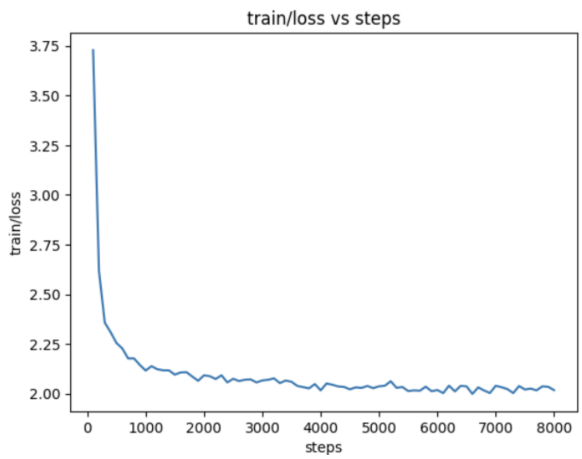

Figure 8: Training losses vs steps during joint training.

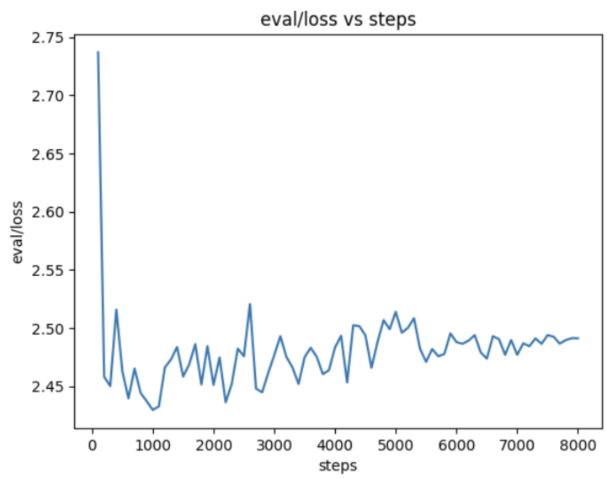

Figure 9: Evaluation losses vs steps during joint training.

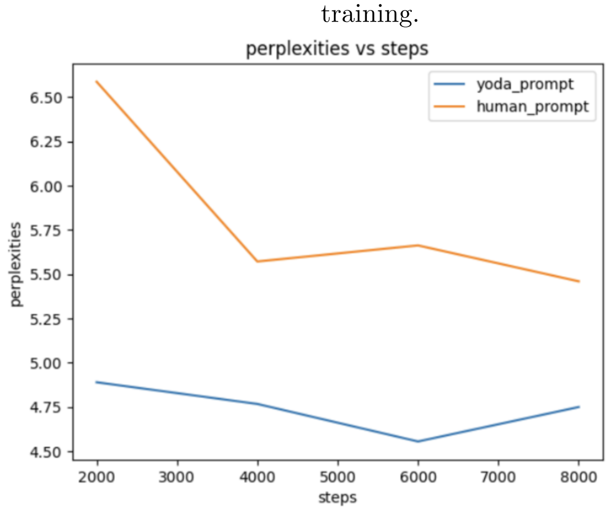

Figure 10: Perplexities vs steps of models that were fine-tuned using joint training.

6.4 Limitations:

Given that the algorithm of translating an English sentence into a Yoda sentence is imperfect, there exist sentences within our large training dataset that are not entirely Yoda sentences. Another limitation is that while we are confident in our results with respect to the candidate sentence structure, Yoda, these results may not generalize to all various types of sentence structure changes.

6.5 Conclusion:

Overall, we found that varying rank when fine-tuning with LoRA did not have much of an effect on perplexity and overall outcome, so it therefore best to use a lower rank for tasks involving syntax changing for maximal compute efficiency. We also found that targeting the positional encoding and token embedding modules does not provide any benefit, while targeting both the self-attention and projection modules gives the best results. Joint training yielded good results, teaching us that the model has the capacity to know two syntax structures at once without performance degradation.

7 Code:

Github: `https://github.com/webs8328/YodaNLP`

7.1 Directory Map

- gpt_code
    - logs: Directory containing all the TFEvent files we used for plotting.
    - LoRATuner.py: this file contains the LoRATuner class which handles the fine- tuning operations and evaluation.
    - LoRA.ipynb: This notebook is where the models are fine-tuned and the hyperparameter sweeps are done.
    - PlottingUtils.py: Contains Plotter class which handles all the plotting of the informatin in the TFEvents files.
    - Plotting.ipynb: This notebook is where all the outcomes of the experiments are plotted.
- data
    - create_train_val_test.ipynb: This notebook creates the training, validation, and test data and handles shuffling of data points, writing data to text files, and interaction with data_funcs.py.
    - data_func.py: This contains the functions for turning a sentence into a Yoda sentence as well as formatting all of the data we need.
    - train_data_final.txt
    - val_data_medium.txt
    - yoda_human.txt
    - test_data_final.txt

8 Citations

1. getting target modules
2. fine tuning method
3. Sentence → Yoda Sentence Code
4. generics_kb dataset
5. LoRA resource
6. Info
7. BLEU Info
8. Perplexity Article
9. Model Description

Sentence	Rank	100 Steps Response	2000 Steps Response	4000 Steps Response
"User: Tell me about dinner." [Using seed 2384]	4	Yup. At a barbecue, I'd ask about your favorite flavor. This is a personal relationship. It's something I do with my wife. I have an entire day to go	Yoda: A place to be, dinner has.	Yoda: A food exchange of food while you are attending university in the summer, dinner is.
"User: Tell me about dinner." [Using seed 2384]	16	Yup. At a barbecue, I'd ask about your favorite flavor. Yup. Just a little, like, the kind where it's like a chili chili, or the kind where it	Yoda: A place to be, dinner has.	Yoda: A food exchange of food while you are attending university in the summer, dinner is.
"User: Tell me about dinner." [Using seed 2384]	64	Yup. At a barbecue, people tend to order burgers in pairs. This is a common practice for the barbecue. They get into the grill and cook bacon, the bacon being on	Yoda: A place to be, dinner has.	Yoda: A food exchange of food while you are attending university in the summer, dinner is.
"User: Tell me about Darth Vader." [Using seed 4567]	4	S: Don't say so Darth. I'm not interested in his, I just want to know if he's capable. He was so awesome as a kid. He was like, oh, he	Yoda: With a powerful aura, most evil is.	Yoda: With a powerful aura that invigorates them to fight, Darth Vader is.
"User: Tell me about Darth Vader." [Using seed 4567]	16	S: Don't say so Darth. I'm not interested in his, I just want to know if he's capable. He was so awesome as a kid. He was like a Jedi, he	Yoda: With a powerful hammer-like weapon, most villains rely.	Yoda: With a powerful hammer- blade, Darth Vader can put.
"User: Tell me about Darth Vader." [Using seed 4567]	64	S: Don't say so Darth. I'm not interested in his, I just want to know if he's capable. He was so awesome as a kid. He was like a Jedi, he	Yoda: With a powerful aura, most evil is.	Yoda: With a powerful hammer- blade, Darth Vader can.

Figure 11: Generated responses with different LoRA Ranks.

Sentence Input		100 Steps Response	2000 Steps Response	4000 Steps Response
"User: Tell me about dinner." [Using seed 2384]	Project and Self Attention	Yup. At a barbecue, I'd ask about your favorite flavor. Yup. Just a little, like, the kind where it's like a chili chili, or the kind where it	Yoda: A place to be, dinner has.	Yoda: A food exchange of food while you are attending university in the summer, dinner is.
"User: Tell me about dinner." [Using seed 2384]	Positional Encoding	[Entertainment journalist/producer] The most important thing now. Trey: I think I read somewhere that you're telling me about dinner. Trey: Are you going to	[He grunts as he tries to talk but is ignored by the maid] KARL: Ok. Alright. PURSAS: What? KARL:	[He grunts as if to be stopped by a child dressed in a t-shirt. She is a blonde with curly brown hair above a long green earring, brown eyes, and red lipstick.
"User: Tell me about dinner." [Using seed 2384]	Token Embedding	[Entertainment journalist, Janae Janae?] Janae: Do you think I'm bad, or am I not?" [What? Is that really true?]	[He opens a tray and tries to taste it. He gets too drunk to speak, takes the tray and tries to shake it.] [Psssst.....wow] Now, his wife	[He opens a tray and a tray of foods. He takes a glass of juice out. She picks it up and heads out of the room. She says to her companion, "Now, you
"User: Tell me about dinner." [Using seed 2384]	Projection	Draco: My date. They are very close. Draco: Well, do you know what they were planning tonight so far. Draco: Well, I wouldn't say they	Yoda: A feast, dinner has.	Yoda: Serves like an annual dinner, dinner includes.
"User: Tell me about dinner." [Using seed 2384]	Self Attention	[A small smile and voice sounds. A lot of girls are getting ready to go to school, so they leave their plates for a nap and go, they've probably been going	Yoda: A delicious meal of deliciousness and the comfort of a relaxed night, dinner is.	Yoda: A delicious meal of deliciousness and the comfort of a relaxed night, dinner is.

Figure 12: Responses Generated from Various Target Modules.

Sentence Input	Type	100 Steps Response	4000 Steps Response	8000 Steps Response
"User: Tell me about dinner.\nHuman:" [Using seed 2384]	Human	Human: Yes, we had dinner. The first time we talked about dinner, there were about 20 other guests here. Human: A lot of diners. What was your favorite dish they served it to	Human: Dinner is served by the same person with the same number of guests per room.	Human: Dinner is one of the most boring meals you've ever eaten.
"User: Tell me about dinner.\nYoda:" [Using seed 2384]	Yoda	Yoda: Dinner is always good for you. [Yoda nods and walks. Yoda turns around at him.] [Yoda walks out to the balcony where, as you can see,	Yoda: Dinner, all events take place.	Yoda: Dinner is always good for you.
"User: Tell me about Darth Vader.\nHuman:" [Using seed 4567]	Human	Human: Yeah, of course. Human: I mean, yeah. They're a big name in terms of being able to put in there a lot of firepower. Human: Yeah, right.	Human: Darth Vader is one of the most impressive people I have ever met.	Human: Darth Vader is one of the most impressive people I have ever met.
"User: Tell me about Darth Vader.\nYoda:" [Using seed 4567]	Yoda	Yoda: Yes, I am A/N: Please tell me about Darth Vader. Yoda: I'm the son of the Lord of the Sith, and I spent my whole life	Yoda: All-female characters, Darth Vader becomes.	Yoda: All-female characters, Darth Vader becomes.

Figure 13: Unlike the other prompts where we waited for Yoda to respond, in this section we purposefully gave the model a "Human" or "Yoda" prompt as well and received the response that way.