Step 2 of ChatGPT Training Demystified: Part 2 of ChatGPT Series

# Introduction

In my [last](https://www.linkedin.com/pulse/chatgpt-how-works-my-notes-part-1-ajay-taneja/) blog on ChatGPT[], I had briefly discussed the steps involved during ChatGPT training. The training of ChatGPT involved 3 fundamental steps – outlined below, again:

1. Fine tuning of the Generative Pre-Trained Transformer (GPT) model where we had human labellers providing the prompt and the response. This step is termed is supervised fine tuning where we had both the questions and the answers available for the training analogous to a classification problem where we have both – the features and a label.
2. In the step 2 of the ChatGPT training, we train a “Rewards” model. Here, we collect data from the Supervised fine-tuned model and then we have human labellers ranking the responses – the ranking (or otherwise called as “reward”) was proportional to the quality of the response. We use the data to train the “Rewards” model.

Thus, the input to the Rewards model will be both the prompt and the response by the model (one or more models may be involved here) and the output will be a scalar which is a measure of the quality of the response.

1. In Step 3 of GPT training, both the supervised fine-tuned model of step 1 and the Rewards model are used.

Here we first pass an unseen prompt to the supervised fine-tuned model (trained in step 1) and then get the response. We then assess the quality of the response using the Rewards model which was trained in Step 2. The rank returned as the output from The Rewards model is used to further fine tune the Supervised Fine-tuned model.

Let us now demystify the step 2 of ChatGPT training as described above in some detail. Thus, firstly let us try and understand why ChatGPT may generate different responses for the same prompt.

# Why ChatGPT may generate responses for the same prompt?

It should be underscored that ChatGPT may not give you the same response for your (same) prompt each time. And the simple reason for that ChatGPT is trained to generate a probabilistic response not deterministic! It has been trained to generate a probabilistic response to achieve a more human like behaviour because humans tend to say words which might not be the most optimal each time they speak.

To understand this better let us throw light on some decoding strategies for response generation. It may be recalled that the generation of a response by a language model is done by generating one token at a time. Following strategies are the most common ones that may be employed whilst generating text:

* **Greedy Approach** – The Greedy Approach uses the most probabilistic response of all based on the training data available but might not always generate the best response. For example, in the prompt:

Context: Try this beer, I have been brewing since last 3 weeks

Response 1: This is ok

Response 2: That tastes great!

In this case, using a greedy response, the model will choose the response: “This is ok” because of the word “is” in the Response 1 which is part of the training data. This is a counter-intuitive based on human behaviour of answering to the query.

* **Temperature sampling**: In case of greedy approach, we look for a stochastic process to avoid the response being a generic or deterministic one. For example, if we are to generate the first token corresponding to a response, the greedy approach will always give the token with the highest probability as explained above whereas an approach such **Random sampling** will generate very random based on all possible worlds/tokens.

We use Random sampling with temperature to increase the probability of probable tokens and reduce the ones which are not – this approach of temperature sampling is “less random” than employing Random sampling and temperature values between 0 <= temperature <= 1 are normally used. A temperature value of 1 indicates no effect at all and the process will work like greedy approach.

* Top-K sampling: In Top K sampling, we eliminate the word s which are less probable and use the top-k words for temperature sampling

This webpage: [Playground - OpenAI API](https://platform.openai.com/playground) is very interesting and allows you to alter the temperature and Top-K values and see the difference between ChatGPT responses!

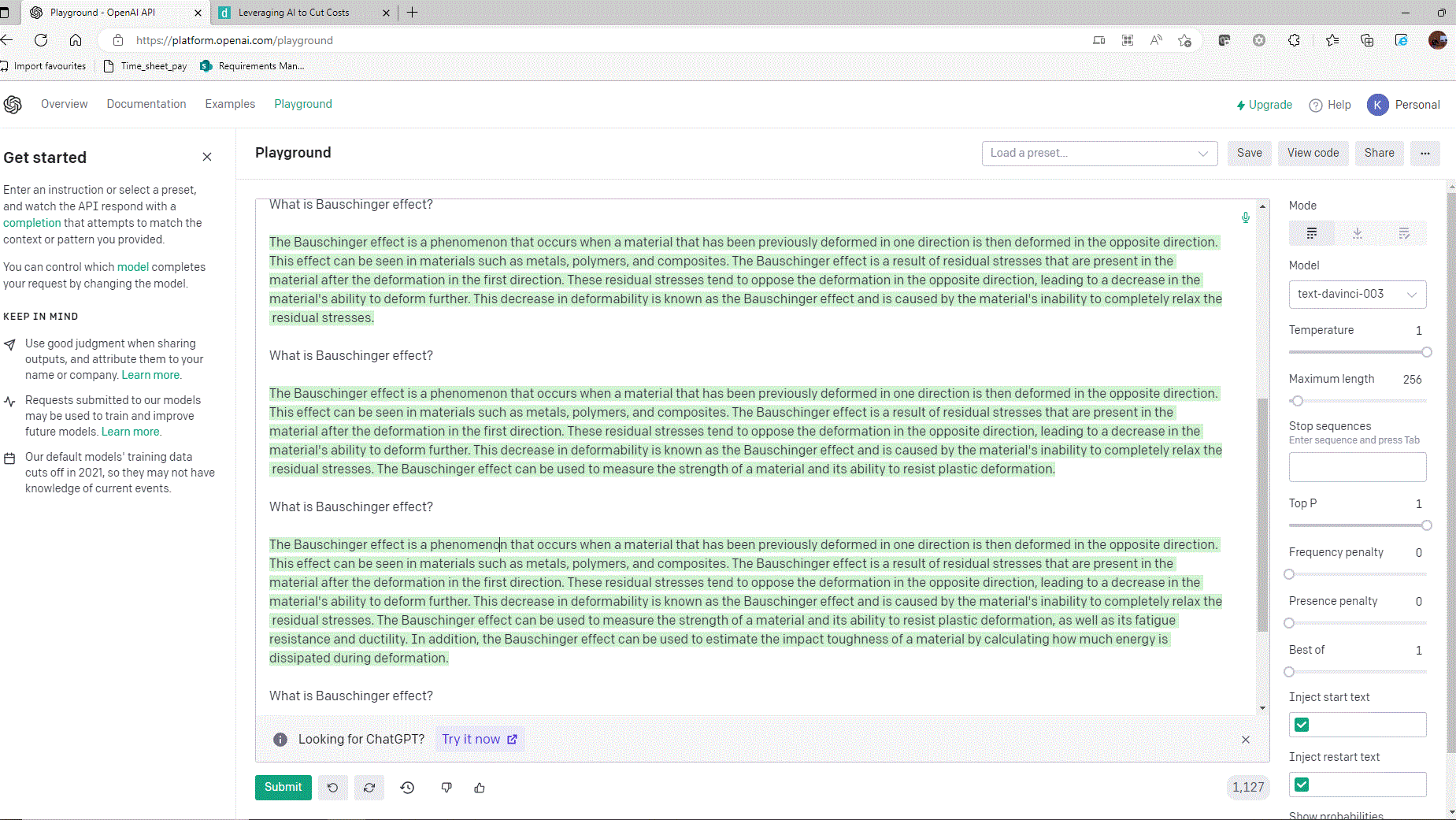


Figure: ChatGPT generating different responses for the same prompt (note the change in the last sentence, more different responses may be obtained when asking ChatGPT to write a code snippet/likewise than a conceptual question)

# How do labellers assess the quality of the responses to train the Rewards model?

Thus, from the above it is clear that how ChatGPT may take the same prompt and yet generate different responses, and this is due to the probabilistic nature of the model meant to emulate human-like behaviour.

In the next part of Step 2, the labellers will have to rank multiple responses. By ranking, the labellers will have to assign a rewards value to the response because this rewards value is going to be used in the step-3 to further fine tune the supervised fine-tuned model of this step (step-2).

The question now needs to be addressed is that – in order for the labellers to rank the responses – how do you quantify the quality of the response?

The labellers assess the quality of the response based on the questionnaire as shown below:

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

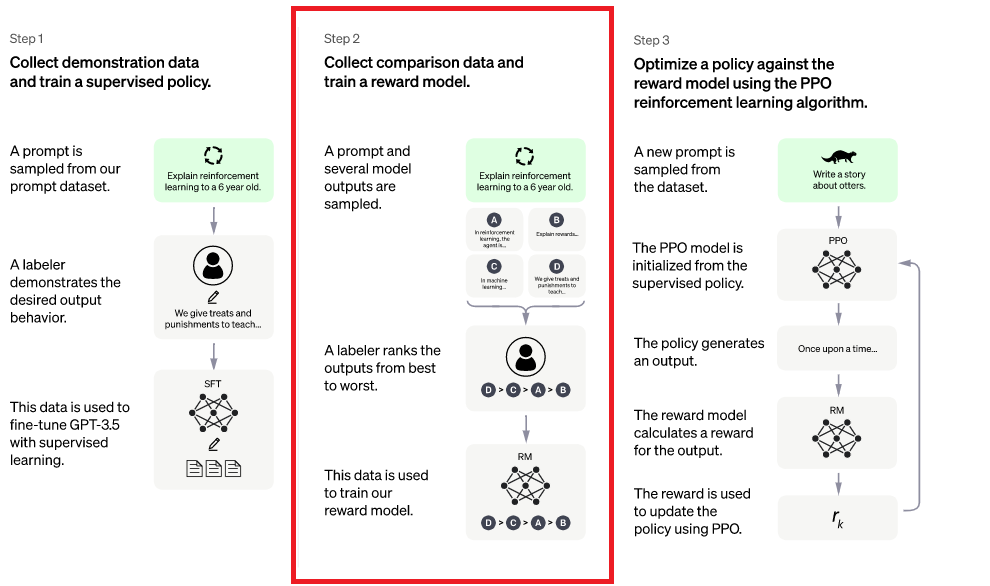
Screenshots of Labelling interface of Step 2 of ChatGPT Training: (a) For each output, labellers give a Likert score for overall quality on a 1-7 scale (b) After evaluating each output individually, labellers rank the output for a given prompt.

The above screenshots are taken from the paper: “Training Language Models to follow instructions with human feedback” by OpenAI and can be found in the below Github Link:

<https://github.com/atanejajlr/language_models>

This is important because we’d like ChatGPT To have an understanding of sensitive topics like: violent content, biased content, or sexually explicit content. All the human labellers will be asked to respond to the questionnaire and the response will then be aggregated. The responses are filled up and only the responses corresponding to the majority are considered. The ratings are used to train the “Rewards Model”

Rewards Model is similar to the supervised fine-tuned model but with a scalar output. I have not gone into more details of the Rewards model, but the details are available in the attached paper on InstructGPT. This Rewards model (after training) is then used to assess the quality of the response and further fine tune the fine-tuned model of step 1 in order to generate more human like responses which are factual and not toxic.



**Figure: Steps in ChatGPT Training**