Dynamic probability adjustment in Simulation-based Off-Policy Evaluation

Dan Amler
Technion – Israel Institute of Technology
dan_amler@campus.technion.ac.il

Abstract

This project aims to enhance the accuracy of predictions of the model presented in "Human Choice Prediction in Language-based Persuasion Games: Simulation-based Off-Policy Evaluation." And check an alternative method to the choice of strategy. While the original model improved accuracy using simulated data and off-policy evaluation, it relied on randomly selected strategies each round and an oracle strategy that knows the right action. The oracle simulates the learning of the agent, by increasing this strategy each round it simulated the learning curve of a real player. This approach did not fully capture the dynamic interactions between human players and travel agents. Our work dynamically adjusts strategy probabilities based on previous rounds, reflecting the agent's perceived reliability of the travel agent. This method aims to provide more accurate predictions of human behavior in language-based persuasion games.

1 Introduction

Recent advancements in Large Language Models (LLMs) have paved the way for developing agents capable of interacting with both human and artificial counterparts in various tasks. Among these tasks, predicting human decisions in language-based persuasion games has gained significant interest. These games involve agents using verbal communication to influence their partners' decisions, making the understanding of human choice in these interactions crucial.

The original model presented in "Human Choice Prediction in Language-based Persuasion Games: Simulation-based Off-Policy Evaluation" demonstrated that incorporating simulated data based on off-policy evaluation significantly improved the accuracy of predicting human players actions. This model relied on a strategy selection process where each round's strategy was chosen randomly, with the learning component modeled as an oracle strategy. While effective, this approach did not fully capture the dynamic interactions a human player might have with different travel agents. It also used the same decision process for all agents, limiting its ability to adapt to the unique characteristics of each interaction with different travel agents.

Our project aims to refine this model by dynamically adjusting the probabilities of using specific strategies based on previous rounds outcomes. Instead of relying on an oracle, we implemented a system where the probability of selecting a strategy increases if it was successful and decreases if it was not. We aim for this method to more accurately reflects the learning process a human player undergoes, adapting to the perceived reliability of the travel agent over time.

In our approach, we utilized some of the original strategy combinations but focused on modifying the probability rate changes based on the outcomes of previous games. This allowed the agent to learn and adapt its strategies dynamically, trying to improve the simulation's resemblance to real human interactions.

By making the strategy selection process to reflect past interactions, our enhanced model aims to create simulations that better mirror real human behavior. This dynamic approach is expected to yield more accurate predictions of human decisions in language-based persuasion games, providing deeper insights into human choice and

improving the performance of LLM based agents in these tasks.

2 Methodology

To improve the model in "Human Choice Prediction in Language-based Persuasion Games: Simulation-based Off-Policy Evaluation," we introduced a dynamic probability adjustment mechanism. This adjusts strategy probabilities based on their success in previous rounds, aiming to better emulate human learning and adaptation.

2.1 Strategy Selection Process

The original model selected strategies randomly, guided by an oracle. Our approach dynamically adjusts these probabilities: increasing for successful strategies and decreasing for unsuccessful ones, reflecting real human decision-making more accurately.

2.2 Evaluate strategy success

The update_proba_dynamic function adjusts strategy probabilities based on previous rounds outcomes. The strategy evaluated as successful if a hotel with a mean score greater than 8 is chosen or a hotel with a mean score smaller than 8 is rejected.

2.3 Probability adjustment

We added a new parameter to the original model, θ (theta in the code), which is the rate of change to the strategy each round.

$$new\ prob = \max\{0, \min\{1, prob + \theta\}\}$$

We used the max and min functions to keep the probability in range [0,1]. Note that theta can be negative in case the strategy was unsuccessful, to decrease the probability. Then we determine the change in the probability, the value can be different from θ because of the max and min functions

$$\delta = new\ prob - prob$$

Now we distribute the change over all other strategies (n – number of strategies)

$$Adjustment = -\frac{\delta}{n-1}$$

For each strategy we adjust the probability:

$$prob = \max\{0, \min\{1, prob + adjustment\}\}$$

Ensure that the total probability is 1:

$$Total = \sum_{k=1}^{n} prob[k]$$

If Total is not equal to 1 we normalize the probability vector:

$$prob[k] = \frac{prob[k]}{Total}\ for\ all\ k$$

By following these steps, the model dynamically adjusts the strategy probabilities based on the previous rounds while keeping the probability vector valid.

3 Experiments

3.1 Dataset

We used the same dataset from "Human Choice Prediction in Language-based Persuasion Games: Simulation-based Off-Policy Evaluation," consisting of 87,000 decisions made by 245 human players interacting with 12 different artificial agents. Each player engaged with six agents, each using distinct strategies, in a multi stage game aimed at persuading the players to select certain hotels based on reviews. This provided a consistent source of data for comparing the original and our model.

3.2 Parameters

We experimented with different sets of the following parameters, all the other parameters that were used are the default from the original work.

Theta: Different values of theta (θ) which are 0.02, 0.05, 0.07, 0.1, that adjusts the probabilities of selecting strategies based on their performance.

Feature: We tested different LLM representations, EFs, GPT-4, and BERT, to capture various aspects of the textual data to check if one of them may work better with dynamic probability adjustment.

Basic Nature Vectors: We tried different basic nature vectors, which are the initial probabilities of the strategies we were testing.

Online simulation factor: We ran the tests with online_simulation_factor 0 and 4 to test the effectiveness of the simulations in light of the changes we made to the way the probability is updated each round.

Random Seeds: We used different random seeds, to ensure the robustness of our results.

3.3 Execution

To evaluate the performance of our model, we conducted extensive testing using Weights and Biases (wandb) to log and track the performance metrics of the experiment and later to visualize the results. Eight different agents were used simultaneously to employ distinct combinations of parameters to efficiently run the tests. A total of 108 different combinations of the previously mentioned parameters were tested.

4 Results

In our testing, we aimed to evaluate the performance of our dynamic probability adjustment method compared to the original oracle based approach. Here are the key findings from our experiments:

Accuracy: The overall accuracy of our modified model was similar to that of the original model. The highest accuracy we got in our testing was 84.3% which is good as the accuracy in the original paper. While the overall accuracy did not improve much, our experiments demonstrated the effectiveness of using the dynamic probability adjustment as a viable alternative to the oracle approach.

Impact of theta value: Different values of theta were tested to determine their impact on the model's performance. The results showed that the model's accuracy remained consistent across various theta values, different combination of parameters resulted in different best theta values but most of time theta = 0.1 gave the best results, lower theta values gave worse results, meaning the change in probability was too small to learn the best strategy.

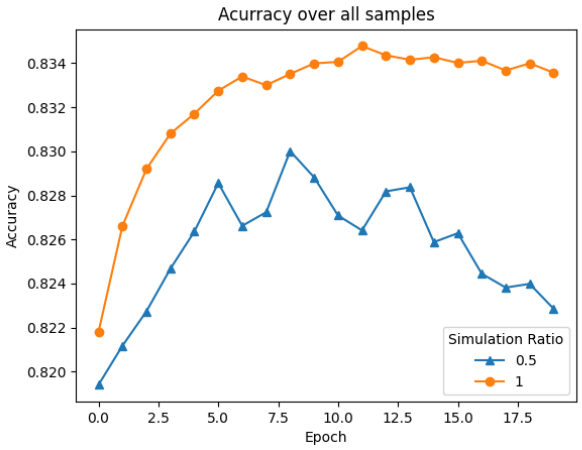
Feature method: As in the original work we got the best result with the representation ‘EFs’

Online simulation factor: The results for online simulation factor equals to 4 were much better, showing the importance of the simulated players which improve the accuracy of predictions on real human players.

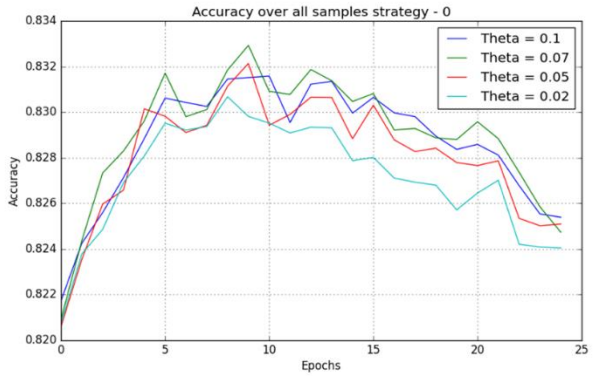
5 Visualization of the results

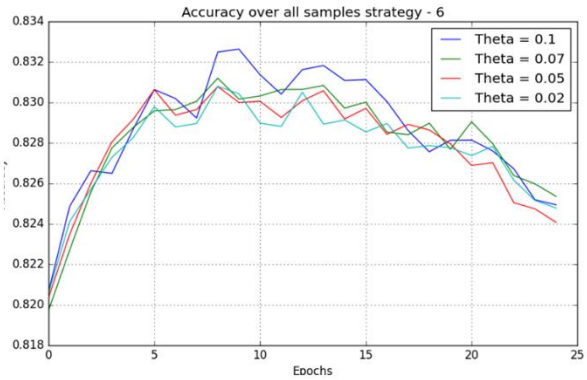
We visualized the results of our tests to better illustrates the performance of the model

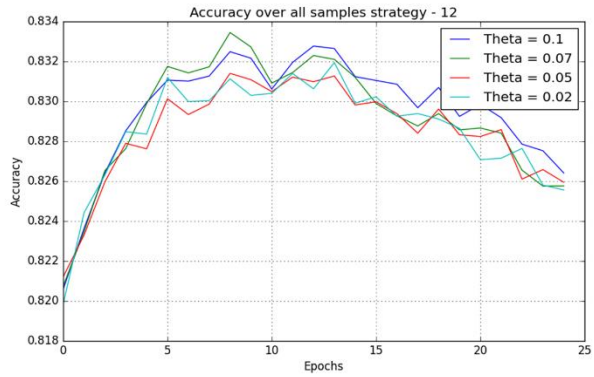
1) Average accuracy over all tests per epochs illustrates the importance of the simulations

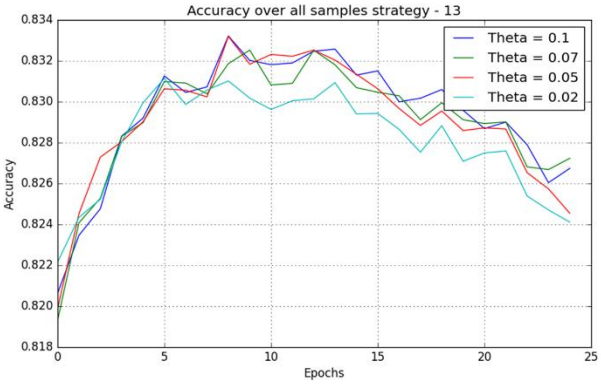


2) Average accuracy over all samples of a strategy with different theta values

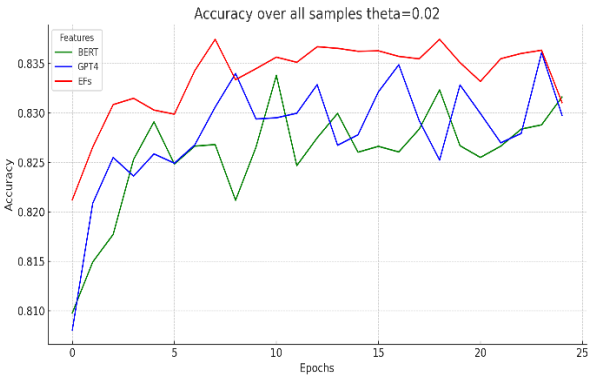


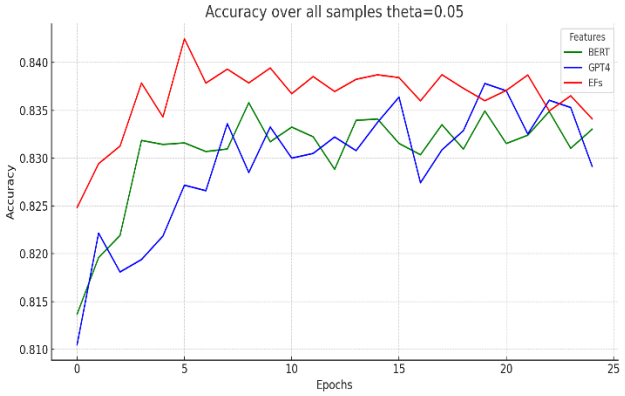


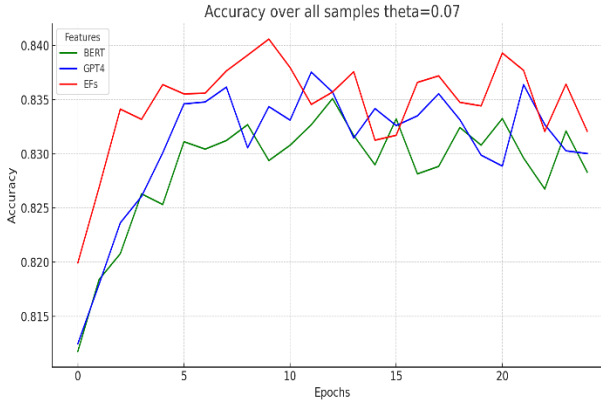




3) Average accuracy over all samples of a theta value with different feature representation







6 Conclusion

We didn’t manage to improve the accuracy results of the paper "Human Choice Prediction in Language-based Persuasion Games: Simulation-based Off-Policy Evaluation" but this work demonstrated that the probability adjustment method that was used is a valid alternative to the oracle based approach from the original work to simulate the learning aspect of a real human player. Our tests showed that the accuracy of this method is as good as the oracle based method. This validates the effectiveness of using dynamic probability adjustments for strategy selection.