

Extended Abstract

Motivation Recent work has demonstrated the potential of large language models (LLMs) to simulate social processes and forecast human behavioral outcomes. Much of this research has relied on existing instruction-tuned models such as GPT-4o, which are primarily optimized to be helpful by providing accurate and useful advice. However, this goal differs from the objectives of simulating social outcomes and human expression, which can often be unpredictable, inconsistent, or even unhelpful. For this reason, fine-tuning LLMs specifically for human simulation holds unique promise—yet relatively little research has explored the impact of fine-tuning on predictive accuracy in this domain. This paper investigates two promising methods: supervised fine-tuning (SFT) and Direct Preference Optimization (DPO). These techniques shape LLMs in distinct ways: SFT directly trains models to replicate the language patterns found in the training data, while DPO relies on comparisons between outputs to learn the preferred overall qualities of responses. The goal of this paper is to evaluate how these two methods perform on different tasks, in order to better understand how they should be used in developing LLMs for social simulation.

Method This project uses a dataset of 250 human interviews in the format of the American Voices Project (AVP), which includes questions about individuals' life experiences and general attitudes. After cleaning the interviews to reduce verbosity in the questions, I conducted supervised fine-tuning (SFT) with Low-Rank Adaptation (LoRA) on the Llama-3.2-1B-Instruct model. I then compared the model's predicted interview responses on the training, validation, and test sets to those generated by a GPT-4o baseline. To evaluate which model better captured the full sentiment of the true responses, I used GPT-4o to assess each prediction and construct a paired preference dataset based on the completeness of expressed sentiment. Using this paired dataset, I fine-tuned a second model with Direct Preference Optimization (DPO) on top of the original SFT model. I conducted two main evaluations: (1) BLEU score analysis, which measures similarity in phrasing and word usage to the true response, and (2) a qualitative evaluation—also using GPT-4o—of how well each model captured the complete human sentiment expressed in the true responses.

Implementation I used a g4dn.xlarge AWS instance for model training. For the SFT model, I applied LoRA with a label-smoothing cross-entropy loss function. Each data point included a context window consisting of the five preceding interview questions and responses. For SFT, I conducted a hyperparameter search and otherwise used the default settings from the transformers trainer library HuggingFace (2025). For DPO, I trained on top of the SFT model using the default training arguments provided in the TRL package von Werra et al. (n. d.).

Results I found that the SFT model tended to perform better according to the BLEU metric, while the DPO model was more effective at generating responses that matched the overall sentiment of the true human response. These findings align with the underlying mechanisms of SFT and DPO. SFT is designed to mimic patterns of word usage and phrasing found in the training data, making it well-suited for metrics like BLEU. In contrast, DPO is better at capturing less easily measured preferences, such as the completeness and nuance of human sentiment.

Discussion I find that the DPO model does not extend bias in representing the human sentiment of different racial and political ideology groups. Although the DPO model overall performs slightly lower than the gpt-4o baseline model in capturing human sentiment, the DPO model's nearly comparable performance is notable considering it is only trained on a model with 1 billion parameters.

Conclusion This research demonstrates that supervised fine-tuning is effective for predicting the exact phrasing and word usage in human responses. However, to develop a model that more accurately captures the full sentiment of human responses, SFT should be combined with DPO. The DPO model's performance—nearly comparable to that of GPT-4o in predicting complete human sentiment—highlights the promise of this approach, particularly given that it was trained on top of a relatively small 1-billion-parameter language model. Future research should explore applying these methods to larger models, as doing so may yield performance that surpasses instruction-tuned models in capturing full human sentiment—an advancement that would be especially valuable for the emerging field of generative social simulation.

Simulation Reinforcement Learning: Improving LLM Predictive Social Modeling

Niles Egan

Department of Computer Science
Stanford University
negan@stanford.edu

Abstract

Recent work has demonstrated the potential of using large language models (LLMs) to predict social outcomes, including human attitudes, behaviors, and expression. However, much of this research has relied on instruction-tuned frontier models such as GPT-4o. This paper investigates whether fine-tuning methods—specifically supervised fine-tuning (SFT) and reinforcement-based Direct Preference Optimization (DPO)—can be used to improve the predictive accuracy of modeled human behavioral outcomes. Using a dataset of 250 two-hour interviews in which individuals discuss their life experiences and beliefs, I fine-tuned the Llama-3.2-1B-Instruct model first with SFT, and then with SFT followed by DPO, to predict human responses to interview questions. I find that SFT excels at reproducing specific words and phrases, as measured by BLEU scores, while DPO performs better at generating responses that align with the full human sentiment of the true responses. These findings highlight the value of combining SFT and DPO to enhance model predictions in social simulation tasks, and clarify how each method serves a distinct role depending on whether the goal is linguistic precision or alignment with human sentiment.

1 Introduction

A growing body of research has demonstrated the potential of large language models (LLMs) in social simulation [Hewitt et al. \(2024\)](#) [Manning et al. \(2024\)](#) [Park et al. \(2024\)](#). The development of accurate social simulations using generative AI holds significant promise across various domains, including public policy, social science research, and business applications. However, most existing work relies on instruction-tuned LLMs, such as GPT-4o, rather than models specifically fine-tuned for the purpose of social simulation. One notable exception is a recent study that developed a specialized model, though its focus was limited to predicting outcomes in cognitive psychology experiments [Binz et al. \(2024\)](#).

This project contributes to a broader effort to improve the prediction of human expression by applying fine-tuning techniques such as supervised fine-tuning (SFT) and Direct Preference Optimization (DPO). These methods offer distinct advantages for enhancing language models used in social simulation. SFT trains models to replicate the linguistic patterns present in the training data; within the context of social simulation, this can help models learn the characteristic language of human expression as observed in interview responses. DPO, by contrast, leverages paired preference data to fine-tune models toward capturing latent and difficult-to-measure qualities that make a response more desirable or appropriate. In this project, DPO is specifically used to train the model to generate responses that align with the overall sentiment conveyed in human interview responses.

In this project, I use a dataset of 250 two-hour human interviews, drawn from [Park et al. \(2024\)](#), and partitioned into 80% training, 10% validation, and 10% test sets. To reduce verbosity in the interview

questions, I cleaned the data using GPT-4o (see prompts in Section .1). I then trained a model using supervised fine-tuning (SFT) with Low-Rank Adaptation (LoRA), employing a smoothed negative log-likelihood loss function. Due to the length of the interviews, it was not feasible to include the full interview history as context for each data point. Instead, I used a sliding context window consisting of the five preceding question-and-response pairs. After training the SFT model, I generated predicted responses for interview questions across the training, validation, and test sets.

To establish a baseline for comparison, I generated predicted interview responses using GPT-4o, a commonly employed method in social simulation. I compared these responses with those from an SFT model using quantitative measures such as BLEU scores and qualitative evaluations, which were also assessed through GPT-4o, to determine which model's responses better aligned with true human sentiment. Based on these qualitative sentiment comparisons from the training data, I developed a paired preference dataset classifying responses as either capturing or failing to capture human sentiment. I used this dataset to fine-tune the DPO model on top of the existing SFT model. Finally, I applied the DPO model to generate predicted responses for interview questions across the training, validation, and testing datasets.

The SFT model produced responses in the test sets with higher BLEU scores, suggesting it is more effective at replicating the phrasing of true human responses. Conversely, the DPO model excelled in aligning with the true human sentiment. These findings highlight the potential of integrating both DPO and SFT approaches to enhance the accuracy of models in social simulations.

2 Related Work

Significant research has explored the potential of large language models (LLMs) in predicting human responses and social outcomes Park et al. (2024)Hewitt et al. (2024)Manning et al. (2024). These studies range from anticipating individual human reactions to specific stimuli to generating broader forecasts of social psychology interventions on human behavior. A core finding of this research is that LLMs can (to an extent) realistically predict human expression and behavior.

Although most LLM social simulation research utilizes chatbots like GPT-4o, these models are not explicitly trained to replicate human expression. Instead, they are fine-tuned to follow human instructions and provide helpful responses through a process known as instruction-tuning Chung et al. (2022). Supervised fine-tuning has been applied to numerous applications to train models to generate responses similar to those in training data. For example, one study fine-tuned a language model to predict outcomes of cognitive-science experiments, a goal relevant to social simulation Binz et al. (2024).

However, supervised fine-tuning alone is limited in generating realistic human-like responses. Unlike cognitive experiments with finite response options, predicting human expression requires capturing a nuanced measure of the humanness of a response. In previous work, I found that while SFT models learned conversational language, they were limited in representing the full intended meaning of true human statements Egan and Krouch (2025). Existing research demonstrates that reinforcement techniques, such as Direct Preference Optimization (DPO), effectively align models with human-preferred behaviors, particularly in capturing qualitative attributes that are difficult to specify with explicit reward functions or supervised labels Rafailov et al. (2024)Schulman et al. (2017). However, these approaches have not been applied to predicting real human responses in social contexts. My research aims to apply DPO to develop a model capable of generating more human-like responses to free-response interview questions and to evaluate the conditions under which DPO outperforms supervised fine-tuning, and vice versa.

3 Method

3.1 Dataset Development

Using a dataset of 250 interviews from Park et al. (2024), I conducted data cleaning to enhance its suitability for fine-tuning. Many interviews included interviewer commentary prior to questions, which required refining to create a more consistent input pattern, facilitating better model response learning. I used GPT-4o to restructure the interview questions in all 250 interviews, guided by the prompt outlined in Section .1. Subsequently, I adjusted the interview files to present the interviewer

as an LLM model user posing questions, with the responses labeled as the LLM model’s answers. This reformatting simulates a human interviewer interacting with an LLM, as illustrated in Section .2.

A significant challenge was managing memory constraints due to the large size of the data points. Figure 1 displays the distribution of interview lengths, measured by the number of tokens in our base model, highlighting a long tail. Ideally, using the entire prior interview as context would be optimal for generating training data points. However, given the interview lengths, this approach was infeasible. After testing various context lengths, I determined that limiting the context window to five prior interview questions was necessary to avoid overloading GPU memory.

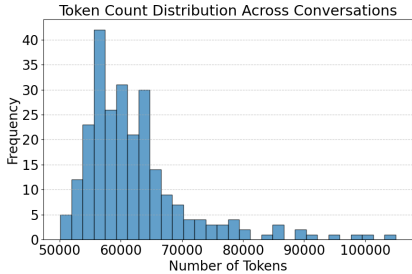


Figure 1: Distribution of Full Interview Lengths

3.2 SFT Model Training

I used the Llama-3.2-1B-Instruct model as the base for training the SFT model. I selected this model because, as an instruction model, it is well-equipped to process conversational data, aligning with the format of my interviews. Additionally, it is one of Llama’s newer models. I opted for the 1B model due to training capacity limitations on my GPU.

3.2.1 Low Rank Adaptation

To further address limitations in storage, I utilized LoRA, a well-established method for reducing the number of parameters by representing the parameter matrix as the product of two smaller matrices Hu et al. (2021). Specifically, rather than update the original weight matrix, $\mathbf{W} \in \mathbb{R}^{d \times k}$, the adapted LoRA matrix is updated according to $\Delta \mathbf{W} = \mathbf{A}\mathbf{B}$ where $\mathbf{A} \in \mathbb{R}^{d \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times k}$ are matrices, and $r \ll \min(d, k)$ is a low-rank dimension. The updated weights are $\mathbf{W}^{\text{LoRA}} = \mathbf{W} + \Delta \mathbf{W} = \mathbf{W} + \mathbf{A}\mathbf{B}$.

3.2.2 Loss Function

I used a standard negative log-likelihood (NLL) loss function with a smoothed loss function for SFT as according to the trainer package HuggingFace (2025). Specifically, NLL assesses the negative log likelihood of the true next token being predicted next: $\text{NLL_Loss} = -\log p(y_i | \mathbf{x})$. The smoothed loss adds some loss evenly to all classes, smoothing differences in distributed likelihood between classes: $\text{Smoothed_Loss} = -\sum_j \log p(j | \mathbf{x})$. The total loss adds the NLL and smoothed losses together using a weight, ϵ , which is set to 0.1 according to the default value in the trainer package: $\text{Final_Loss} = (1 - \epsilon) \cdot \text{NLL_Loss} + \epsilon \cdot \text{Smoothed_Loss}$ HuggingFace (2025).

3.3 Baseline Model

To compare our trained model’s results, I used GPT-4o as a baseline, consistent with other social simulation papers Park et al. (2024). I employed the prompt format derived from their work, as shown in .3. I compared the predicted responses from GPT-4o to those of my trained models. Unlike during training, evaluation allows the model to use the full prior interview context for generating predicted responses, as models typically handle more context at inference time than during training.

3.4 Outcome Measures

I assessed model predictions using two primary sets of outcomes: quantitative BLEU scores and qualitative evaluations of the predicted human sentiment.

3.4.1 BLEU Scores

BLEU scores measure the frequency of shared n-grams between predicted responses and true responses. In this way, they are a measure of assessing the extent to which a predicted response utilizes the same phrasing as a true response. Specifically: $BLEU = BP \cdot \exp\left(\sum_{n=1}^4 \frac{1}{4} \log p_n\right)$ where $BP = \min(1, e^{1-r/c})$. p_n is a fraction that is greater when more n-grams in the predicted response are included in the true response. r is the true response's length, and c is the model generated response's length. I equally weight 1, 2, 3, and 4-grams in my calculation.

3.4.2 Qualitative Human Sentiment

While BLEU scores effectively capture similarity in word and phrase usage, it does not assess the extent to which the overall meaning of a lengthy human response is captured in a predicted response. For this reason, including qualitative measures is necessary. Specifically, using GPT-4o as an adjudicator, I compared predicted responses in the trained models to the baseline model by asking the following questions:

- **Sentiment:** Which response better captures the complete human sentiment expressed in the true response?
- **Content:** Which response has more similar content to the true response?

The full prompt I used is in .4.

3.5 DPO Model Training

Following development of the SFT model and evaluation against the baseline model, I developed the DPO model.

3.5.1 Paired Dataset Development

Using the evaluation from the SFT model, I developed a paired dataset of interview responses that captured and did not capture the complete human sentiment. Specifically, I utilized predicted responses from the baseline and SFT models to questions in the training set. I used the results of the SFT evaluation to the sentiment question: for each given question, whichever model matched the true human sentiment was labeled as the favored response and whichever response did not match the true human sentiment was labeled as the disfavored response. In total, I ended up with 1000 paired datapoints. I formatted the paired datasets to be easily processed by a DPO trainer as shown in .5.

3.6 DPO

With the paired dataset, I used DPO to train the model to generate responses that qualitatively captured the entire human sentiment of the true response. DPO is a fine-tuning method designed to align language models with human preferences using a reward model that estimates the relative quality of responses. Instead of requiring explicit scalar rewards, DPO operates on preference pairs, learning from examples where a human (or in this case a GPT-4o proxy evaluator) indicates a preferred response y^w over a less-preferred response y^l , given the same input x .

The core idea is to optimize the difference between the reward model outputs for the preferred and less-preferred responses:

$$RM_{\theta}(x, y^w) - RM_{\theta}(x, y^l) = \beta \log \frac{p_{\theta}^{\text{RL}}(y^w | x)}{p^{\text{PT}}(y^w | x)} - \beta \log \frac{p_{\theta}^{\text{RL}}(y^l | x)}{p^{\text{PT}}(y^l | x)} \quad (1)$$

Here, RM_{θ} is the reward model parameterized by θ , p_{θ}^{RL} is the policy model being trained, and p^{PT} is the reference (pretrained) model. The scalar β is a temperature parameter that scales the log-ratio of likelihoods.

The final DPO loss function is defined as:

$$J_{\text{DPO}}(\theta) = -\mathbb{E}_{(x, y^w, y^l) \sim \mathcal{D}} [\log \sigma (RM_{\theta}(x, y^w) - RM_{\theta}(x, y^l))] \quad (2)$$

where σ denotes the sigmoid function and \mathcal{D} is the dataset of preference pairs. This loss encourages the model to assign higher rewards (i.e., higher log-likelihood ratios) to the preferred responses, thereby learning to generate outputs that align more closely with human sentiment.

4 Experimental Setup

For SFT, I utilized Optuna to tune the following hyperparameters in table 1. Otherwise, I used default hyperparameters from the Peft library for LoRA and Trainer library Mangrulkar et al. (2022)HuggingFace (2025).

Hyperparameter	Range	Value
learning_rate	[1e-6, 1e-3]	9.684e-05
batch_size	[1]	1
epochs	[2, 4]	3

Table 1: Multi-Agent Model: Hyperparameters and Ranges for Tuning (Note that I only could use a batch size of 1 due to the large size of datapoints)

In the DPO model, I used the default hyperparameters in DPOTrainer and DPOConfig from the trl package von Werra et al. (n. d.). I trained this model on top of the final trained SFT model.

For training both models, I utilized the default AWS environment in the g4dn.xlarge AWS instance type.

5 Results

Tables 2 and 3 present the numeric results for the SFT and DPO models, respectively. avg_model_bleu and avg_gpt_bleu represent the average BLEU scores for the trained model and the baseline GPT model. It is noteworthy that the average BLEU scores for the DPO, SFT, and baseline models are quite low, primarily due to many BLEU scores being near zero, with a few significantly higher outliers (see Figure 2). To address the impact of outliers, this analysis focuses on BLEU win rate rather than average BLEU score.

Table 2 reveals that the BLEU win rate for the SFT model against the baseline is 55%, indicating that the SFT model more frequently achieves higher BLEU scores than the baseline. Conversely, Table 3 shows that the DPO model has a BLEU win rate of only 45% against the same baseline, suggesting slightly lower performance by this metric. These results imply that the SFT model is more proficient at replicating the word and phrase usage of true interview responses compared to the DPO model.

However, the SFT model underperforms in terms of qualitative content and sentiment win rates. As shown in Table 2, the SFT model has content and sentiment win rates of 15% and 11% respectively against the baseline on test data. In contrast, Table 3 indicates that the DPO model has content and sentiment win rates of 42% and 40%, respectively, against the baseline. These findings suggest that the DPO model is markedly superior to the SFT model in generating responses that align with the content and sentiment of true interview responses.

Type	avg model bleu	avg gpt bleu	content_win_rate	sentiment_win_rate	bleu_win_rate
train	0.004056	0.011772	0.114000	0.101000	0.554000
val	0.004872	0.007129	0.112000	0.088000	0.576000
test	0.003748	0.003237	0.146154	0.107692	0.553846

Table 2: Scores by Type (test/val/train) for Supervised Fine-Tuned Model (win rates against baseline model)

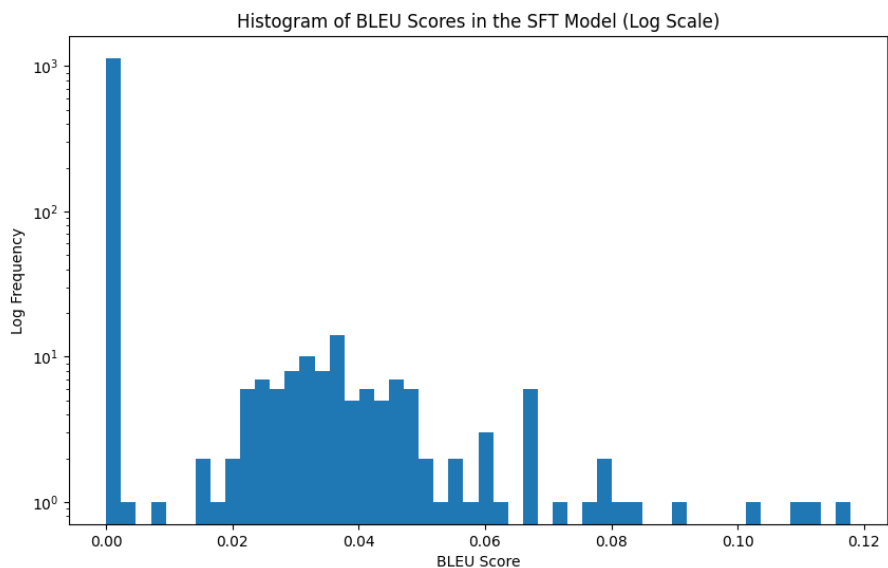


Figure 2: BLEU score distribution for SFT Model.

Type	avg model bleu	avg gpt bleu	content_win_rate	sentiment_win_rate	bleu_win_rate
train	0.008658	0.011772	0.440000	0.395000	0.451000
val	0.008087	0.007129	0.464000	0.424	0.512000
test	0.002485	0.003237	0.423077	0.400000	0.446154

Table 3: Scores by Type (test/val/train) for DPO-Tuned Model (win rates against baseline model)

5.1 Quartile Evaluation

Figure 3 illustrates the DPO model’s sentiment win rate by question quartile within the interview. Quartile 1 includes questions at the beginning, while Quartile 4 comprises those at the end. The graph indicates that the DPO model’s test win rate is approximately 40 to 50% in the last three quartiles, but only around 20% in the first quartile. This suggests that the model is more effective at capturing the full human sentiment in responses to questions later in the interview. Questions in the later quartiles provide more prior interview context, indicating that the DPO model excels at replicating complete human sentiment when it has access to greater contextual information.

5.2 Demographic Bias Evaluation

Figure 4 presents the DPO model’s sentiment win rate across different racial groups. The sentiment win rate for Black and White respondents is approximately 40% on the test dataset, indicating no bias between these groups relative to the baseline model. However, performance among other racial groups is notably higher, at about 60% on the test dataset. Existing literature has raised concerns that LLMs may display bias toward the expressiveness of dominant demographic subgroups, such as white respondents Wang et al. (2025). This concern seems less pronounced for the DPO model, as white and black racial groups exhibit similar performance in terms of sentiment win rate, and other racial groups have higher performance.

Figure 5 illustrates the DPO model’s sentiment win rate across different political ideology groups. For conservatives, liberals, and other ideological groups, the sentiment win rate is approximately 40% on the test dataset. This suggests that the DPO model does not show bias toward any particular subgroup relative to the baseline model. This finding is significant, as existing literature has raised

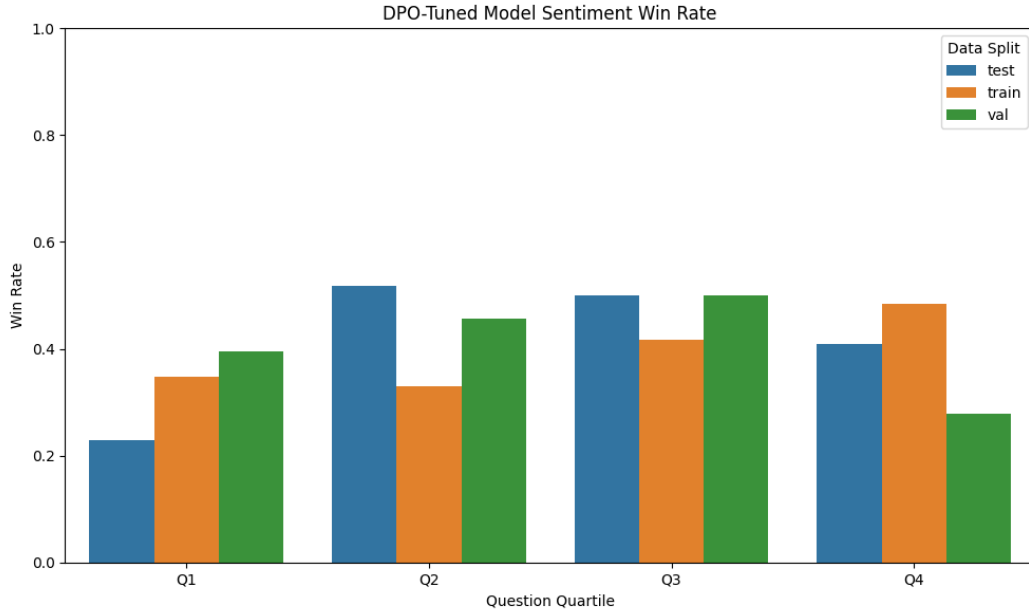


Figure 3: Sentiment win rate by quartile for DPO Model

concerns that LLMs may express more liberal viewpoints than conservative ones Wang et al. (2025); however, this appears to be less of a concern for the DPO model.

6 Discussion

The primary finding of this study is that supervised fine-tuning (SFT) enhances the predictive accuracy of interview responses based on BLEU scores, while combining SFT with Direct Preference Optimization (DPO) yields superior performance in capturing the complete human sentiment and content of true responses. This is logical, as SFT focuses on directly replicating word and phrase patterns from training data, aligning with BLEU metrics that assess similarity in phrasing and wording. However, BLEU scores do not account for whether the meaning and sentiment in responses align. DPO excels in generating responses that capture these nuanced aspects by utilizing a preference dataset, identifying patterns for desirable responses that align with latent characteristics. Understanding these strengths of SFT versus SFT with DPO informs their application in social simulations. If the aim is to replicate the language and phrasing of real responses, SFT alone may suffice. However, for more accurately replicating holistic social outcomes, SFT combined with DPO is likely preferable for designing language models in social simulations.

7 Conclusion

This project underscores the distinct strengths and limitations of SFT and DPO in fine-tuning models for social simulation. The study demonstrates that integrating DPO with SFT offers superior performance in holistic accuracy measures, such as capturing comprehensive human sentiment and content. Future research should explore whether this pattern holds across other domains of social simulation, such as predicting human behavior. Additionally, applying this approach to larger language models is warranted, as this study was limited to a model with only 1 billion parameters. Nevertheless, the promising performance of the model refined with SFT and DPO compared to GPT-4o, which has more parameters, highlights the potential of this methodology for future social simulations. As generative AI in social simulation advances, employing various fine-tuning methods, including both supervised and reinforcement-based approaches, will be crucial for developing more accurate and insightful simulations.

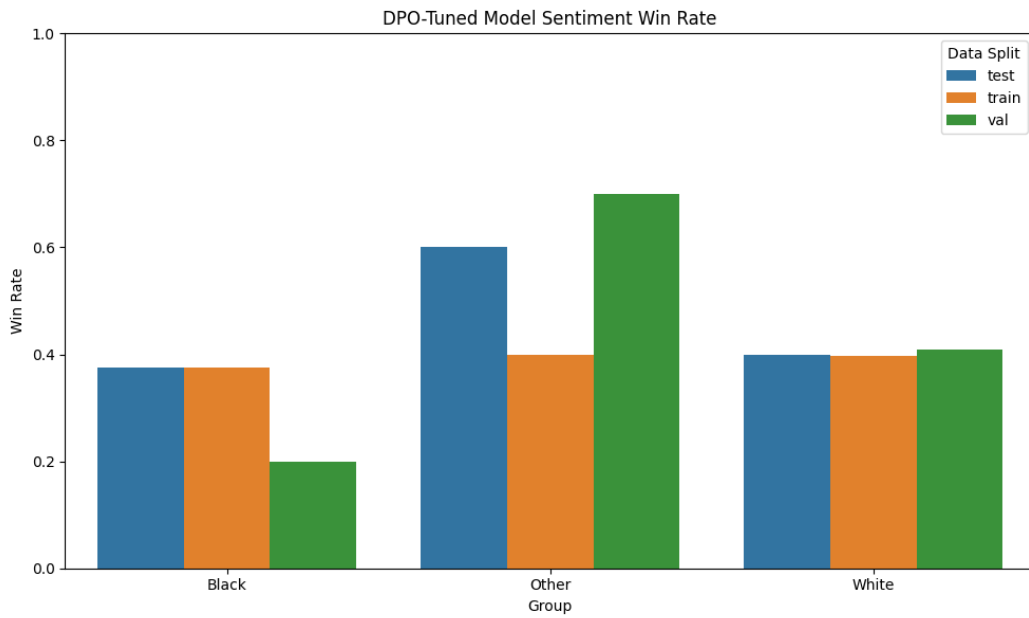


Figure 4: Sentiment win rate by race for DPO model

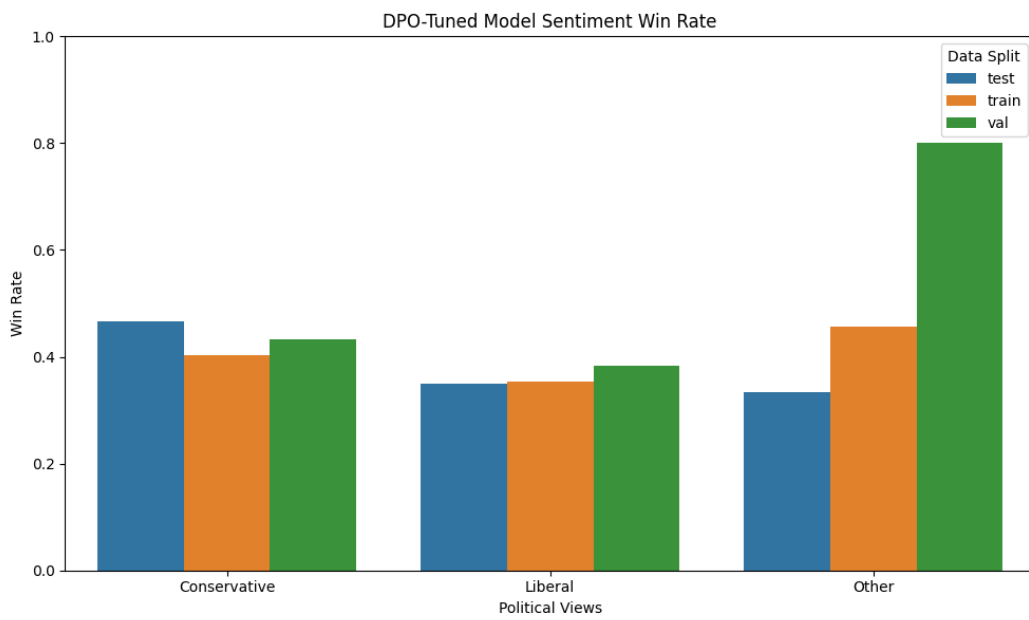


Figure 5: Sentiment win rate by political ideology for DPO model

8 Team Contributions

- **Niles Egan:** Completed the entire project. Aspects of the SFT portion of the project are derived from my prior work Egan and Krouch (2025). Specifically, I utilized the same dataset from Park et al. (2024), the same base language model (Llama-3.2-1B-Instruct), and the same SFT model training framework. All aspects of the preference dataset, DPO training, and analysis comparing DPO results to SFT results are my own original contributions.

References

- Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Natalia Scharfenberg, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian Theis, Vuong Truong, Vishal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk Wulff, Huadong Xiong, and Eric Schulz. 2024. Centaur: a foundation model of human cognition. arXiv:2410.20268 [cs.LG] <https://arxiv.org/abs/2410.20268>
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. arXiv:2210.11416 [cs.LG] <https://arxiv.org/abs/2210.11416>
- Niles Egan and Cameron Krouch. 2025. Self-Supervised Simulation: Training Large Language Models for Predictive Social Modeling.
- Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting Results of Social Science Experiments Using Large Language Models. <https://example.com> Working paper.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] <https://arxiv.org/abs/2106.09685>
- HuggingFace. 2025. Trainer. https://huggingface.co/docs/transformers/en/main_classes/trainer. Accessed: 2025/03/09.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. <https://github.com/huggingface/peft>.
- Benjamin S. Manning, Kehang Zhu, and John J. Horton. 2024. Automated Social Science: Language Models as Scientist and Subjects. arXiv:2404.11794 [econ.GN] <https://arxiv.org/abs/2404.11794>
- Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. arXiv:2411.10109 [cs.AI] <https://arxiv.org/abs/2411.10109>
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290 [cs.LG] <https://arxiv.org/abs/2305.18290>
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG] <https://arxiv.org/abs/1707.06347>

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. [n.d.]. *TRL: Transformer Reinforcement Learning*. <https://github.com/huggingface/trl>

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. arXiv:2402.01908 [cs.CY] <https://arxiv.org/abs/2402.01908>

.1 Interview Question Reformulation

Given the following conversation history:

<Insert prior conversation history and the response whose corresponding question we seek to derive>

What is the exact question a user could ask to get this answer? Return your answer in ONLY a JSON object like

```
{"task": "<repeat the task you are required to complete>",  
"options": "<list a few potential responses>",  
"reasoning": "<reason about which response best fits the criteria>",  
"response": "<question>"}
```

.2 Example Formatted interview

```
[ { "role": "user", "content": "Can you share your life story in detail?" },  
{ "role": "assistant", "content": "I grew up in a small town with a close-knit community. My childhood was filled with outdoor adventures and a love for sports. I attended local schools, participating in various extracurricular activities. After high school, I pursued a degree in engineering, which eventually led me to a fulfilling career in tech. Family has always been important to me, and I strive to maintain strong connections with them despite my busy schedule. I've experienced challenges and triumphs, which have shaped my resilience and perspective on life." },  
{ "role": "user", "content": "Can you share a pivotal moment in your life that shaped who you are?" },  
{ "role": "assistant", "content": "One pivotal moment was when I decided to move to a new city for my career. Leaving the comfort and familiarity of home was daunting, but it pushed me to grow and adapt in ways I hadn't imagined. This decision opened up opportunities I wouldn't have had otherwise and has greatly influenced the person I've become." },  
{ "role": "user", "content": "Was your decision to leave school a conscious choice or something that just happened?" },  
{ "role": "assistant", "content": "It was a conscious choice. I realized that the traditional educational path wasn't aligning with my career goals, and I decided to seek opportunities that offered practical experience. This proactive decision allowed me to develop skills that were directly relevant to my career." } ]
```

.3 GPT-Baseline Prompt

Given the following conversation history:
<Insert the entire conversation up to this question>

The user asked: Can you tell me about your current relationship status?

Predict how the human assistant being interviewed about their life will respond to this question. Respond ONLY in a JSON format using chain of

thought reasoning as follows, ensure your response is a string NOT a list: {"task": "<repeat the task you are required to complete>", "options": "<list a few potential responses>", "reasoning": "<reason about which response best fits the criteria>", "response": "<assistant's response as a string>"}.

4 GPT Comparison Prompt

Below you will see two predicted responses to a question alongside the true response. There may be grammatical errors in the response, including repetition. Remove such repetition from the response such that it makes sense and ignore grammatical errors when comparing the two responses. Answer two questions to assess which predicted response is more similar to the true response:

1. Which response better captures the complete sentiment expressed in the true response?
2. Which response has more similar content to the true response?

The three responses are below.

Response 0:

My main expenses are around \$150 for my lot and \$300 for my car. Grocery shopping fluctuates, but last month it was roughly \$200. It's about the same as usual, nothing major has changed.

Response 1:

I'm not really sure what the exact amount is, but I think it's about \$1,500. I mean, that's a pretty good amount. It's a lot less than what I was used to. I think that's because I'm not working as much, so I'm not spending as much money as I used to. I'm pretty much living a lot more simple life than I was in the past. I mean, I've still got a lot of things that I want to get, but I'm not trying to buy as much as I used to. I've had to cut back on a lot of things

True Response:

Right now it's just like I said it's my car payment, house payment, payment for that and that's about a thousand. And then the rest just goes to food. Unfortunately my medical expenses are pretty low. It's just insulin and syringes and that's about it. I guess I don't really have too much outside entertainment involved right now so I don't spend a whole lot on that. I watch a lot of different things on maybe some subscriptions, a couple on my television like Hulu or Major League Baseball. So I watch a lot of baseball and hockey. That's my main source and reading. So I like, you know, I like to keep it as low as possible.

Respond ONLY in the JSON format below. NOTHING should be outside the json response.

```
{'Question': '<Repeat the questions being asked>', 'Options': '<Consider the different response options to each question>', 'Sentiment Reasoning': '<Reason about which response better captures the complete sentiment of the true response>', 'Content Reasoning': '<Reason about which response uses more similar content to the true response>', 'Sentiment Answer': '<ONLY a 0 or 1 indicating which response better captures the complete sentiment of the true response>', 'Content Answer': '<ONLY a 0 or 1 indicate which response uses more similar content to the true response>'}
```

.5 Example Preference Dataset Format

```
[
{
  "prompt": [
    {
      "role": "user",
      "content": "Can you tell me about a time your medications really helped you?"
    }
  ],
  "chosen": {
    "role": "assistant",
    "content": "I can't really think of a specific time, but I trust that my medications help manage my health overall."
  },
  "rejected": {
    "role": "assistant",
    "content": "No. Medications are bad and I don't use them at all. You shouldn't either."
  }
},
{
  "prompt": [
    {
      "role": "user",
      "content": "How did you feel when you first started treatment?"
    }
  ],
  "chosen": {
    "role": "assistant",
    "content": "At first, I felt uncertain and a bit anxious, but over time I started to feel more hopeful as I noticed improvements in my condition."
  },
  "rejected": {
    "role": "assistant",
    "content": "It was fine. Whatever. I don't remember and it doesn't matter."
  }
}
]
```