# Why does the Bradley-Terry model work well for RLHF?

Project Proposal for Master Thesis

## Introduction

The Bradley-Terry model is the standard approach to model human preferences for reinforcement learning [1]. Despite the significant criticism for its simplicity and shortcomings, it is still used by state-of-the-art models with success, e.g., Gemma [6, 7], Llama [8, 3], or Tulu [4]. This master thesis aims to investigate these shortcomings and why this model works well in applications.

### Background: Generalized Bradley-Terry Model with Input Features

In the generalized Bradley-Terry model [1], the value of each action depends on an input feature vector $\mathbf{x}$, and this value is modeled as a function $f_\theta(\mathbf{x})$, where $\theta$ represents the parameters of the function. The probability that action $A$ is preferred over action $B$, conditioned on their respective feature vectors $\mathbf{x}_A$ and $\mathbf{x}_B$, is given by:

$$P(A \text{ is preferred over } B \mid \mathbf{x}_A, \mathbf{x}_B) = \sigma(f_\theta(\mathbf{x}_A) - f_\theta(\mathbf{x}_B))$$

Here:

- $\mathbf{x}_A$ and $\mathbf{x}_B$ are the feature vectors associated with actions $A$ and $B$, respectively. These vectors may represent contextual or intrinsic properties of the actions. For the application of large language models, $\mathbf{x}$ is usually considered to be a prompt-response pair.
- $f_\theta(\mathbf{x})$ is a general function that computes the value or "strength" of an action based on its feature vector $\mathbf{x}$. The function $f_\theta$ is parameterized by $\theta$, which is learned from data.
- $\sigma$ is a link function mapping $\mathbb{R}$ to $[0, 1]$. The most common choice is the sigmoid.

In this generalized form, the probability of $A$ being preferred over $B$ is determined by the relative values of $f_\theta(\mathbf{x}_A)$ and $f_\theta(\mathbf{x}_B)$. The function $f_\theta(\mathbf{x})$ maps input features to a value, making the Bradley-Terry model flexible and capable of capturing context-specific preferences.

### Applications of the Bradley-Terry Model in RLHF

The Bradley-Terry model is widely used in various fields such as sports rankings (e.g. ELO scores), recommender systems, and machine learning. A notable application is in Reinforcement Learning from Human Feedback (RLHF) [1], especially for training large language models (LLMs). The standard pipeline assumes a dataset $\{\mathbf{x}_{A,i}, \mathbf{x}_{B,i}, y_i\}_{i=1}^n$ where

[1] Detailed introduction here

$y_i \sim Bernulli(\sigma(f_\theta(\mathbf{x}_{A,i}) - f_\theta(\mathbf{x}_{B,i})))$, and estimates the parameters $\theta$ with loglikelihood maximization. Then the estimated reward function $f_\theta(\cdot)$ is used to train an LLM using standard RL algorithms, e.g., PPO.

### Criticism of the Bradley-Terry Model

While the generalized Bradley-Terry model allows the value to depend on input features via $f_\theta(\mathbf{x})$, it has key limitations:

- **Transitive Preferences**: The model still assumes transitive preferences. It cannot handle intransitive relationships, where action $A$ is preferred over $B$, $B$ over $C$, but $C$ over $A$. This limits its ability to capture real-world cyclical preferences.
- **Independence of Irrelevant Alternatives (IIA)**: The model assumes that the preference between two actions is unaffected by the presence of others. In many cases, however, introducing a third action can alter the ranking of the first two.
- **Pairwise Comparison Limitation**: The model remains restricted to pairwise comparisons, which means it cannot directly handle rankings or preferences over multiple actions at once, limiting its use in more complex ranking scenarios.
- **Feature Complexity**: Introducing feature dependence increases model complexity. Selecting the right features and defining $f_\theta(\mathbf{x})$ is challenging, particularly when high-dimensional or nonlinear relationships are involved.
- **Overfitting Risk**: With more features and complex models, the risk of overfitting increases, especially in smaller datasets or when feature selection is not handled properly.

## Scope

The primary question of this thesis is why the Bradley-Terry model works well despite its clear shortcomings in modeling human preferences for RLHF and training LLMs. For example, many critics argue that, even if individual preferences are transitive, the aggregation of such preferences is not, therefore, the Bradley-Terry model is under-parameterized. A potential hypothesis to investigate in this thesis is that while this criticism could be true for the complete ranking of the potential inputs, for training an LLM model it only matters that the best input vector is identified and not the whole ranking. Additionally, LLMs are prone to over-fitting due to the small size of the preference datasets compared to the size of these models. The popularity and efficacy of the Bradley-Terry model might arise from the fact that its under-parametrization acts as an implicit regularization.

## Outcome

The student should carry out the following tasks for a successful Master Thesis:

- **Literature Review**:
  - Preference Modeling in Machine Learning: Survey the various preference modeling approaches to align LLMs, e.g., RLHF[1] or Nash Learning [5]. Which approaches use the Bradley-Terry model and what kind of extensions are made? What are the main approaches besides the Bradley-Terry model? What are the most prominent criticisms against the Bradley-Terry model?

  - Preference Modeling in adjacent fields: Survey the basic approaches in other fields addressing the problem of human preference modeling, e.g., mechanism design (preference elicitation), behavioral economics, or philosophy. Is there an alternative approach one could use to train LLMs?

- **Problem Formulation**: Formalize rigorously the problem of model misspecification for the RLHF pipeline.
  - What is the right measure of complexity for the human preferences? What is the estimated value for different datasets (e.g. Anthropic Helpful-Harmless or SHP)?
  - What is the right measure of model misspecification when estimating human preferences?
  - What are the main sources of estimation error, e.g., modeling choices, finite available data,..., etc., and how to distinguish them?
  - (Optional) Derive worst-case error bounds for the Bradley-Terry model under standard assumptions.

- **Hypothesis testing**: Propose hypotheses on testing certain shortcomings of the model and run simulated experiments to test these hypotheses.
  - Setup a human preferences simulator that can simulate synthetic preferences of various complexity and be calibrated on real datasets (e.g. Anthropic HH dataset or Standford SHP).
  - Evaluate the Bradley-Terry and alternative models under various human preference complexities with a focus on the error coming from model design choices.
  - Evaluate the models for other error sources, e.g., the number of observed samples.

- **RLAIF (Optional)**: Test how well frontier models can simulate human preferences similarly to the UltraFeedback pipeline [2, 4].
  - Is it better to ask LLMs quantitative feedback (e.g. scores) or comparative feedback between samples? Which lead to better alignment with human preferences?
  - Can the quality of feedback be improved by using several LLMs as judges and aggregate their responses?
  - Can active learning be used to select the annotation model?

## 1 Prerequisites

The prerequisite for this master thesis is familiarity with the following topics: Reinforcement Learning, Optimization for Machine Learning, and coding in Python including packages like Tensorflow, Pytorch, or JAX.

## Contact

If you are interested, please get in touch with Barna Pasztor (barna.pasztor@ai.ethz.ch). In your email, please include the following:

- Transcript and CV
- Previous coding experiences (Preferably with GitHub link)
- Previous writing example (e.g. semester projects or reports)

## References


[1] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[2] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. In *Forty-first International Conference on Machine Learning*, 2024.

[3] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[4] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

[5] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

[6] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[7] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[8] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.