ChatBot Arena Human Preference Prediction

DTSA 5511

December 7, 2024

Problem Statement

To better understand human preferences for chatbot (ChatGPT, Gemini, and etc.) responses, Chiang et al. introduced **Chatbot Arena**, an open-source platform designed to evaluate AI through human preference.

This project aims to develop a predictive model of user preferences using data from the leading AI chatbots with head-to-head battles conducted from Chatbot Arena platform.

The data challenge is available: *https://www.kaggle.com/competitions/lmsys-chatbot-arena/*

My Github is available: *https://github.com/lzheng01/LMSYS*

Analysis Plan

Data Exploration

Visualizations: Use histograms and bar chart to visualize distributions and relationships between features.

Correlations: Compute correlations between features (e.g., winning ratio) using heatmap

Predictive Modelling

Neutral Language Processing (NLP) Model : Decoding-enhanced BERT with disentangle attention (DeBERTa)

- Disentangled Attention Mechanism
- Enhanced Mask Decoder
- Training Efficiency

Evaluation Metrics

The validation performance is primarily measured using log loss, which is appropriate for evaluating the probability outputs of the model.

Data Overview

Training Dataset: 57,477 rows

(Data processing: 14 Duplicates from the analysis dataset; No Missing Values)

Test Dataset: 3 rows

Data Structure

Each row in the dataset represents a user interaction. The columns include:

- **id:** A unique identifier for each interaction.
- **model[a/b]:** Identifiers for the two models involved in the interaction.
- **prompt:** The input prompt given to both models.
- **response[a/b]:** The responses generated by model_a and model_b respectively.
- **winnermodel[a/b/tie]:** Indicates which model's response was chosen as the winner by the judge.

Data Snapshot

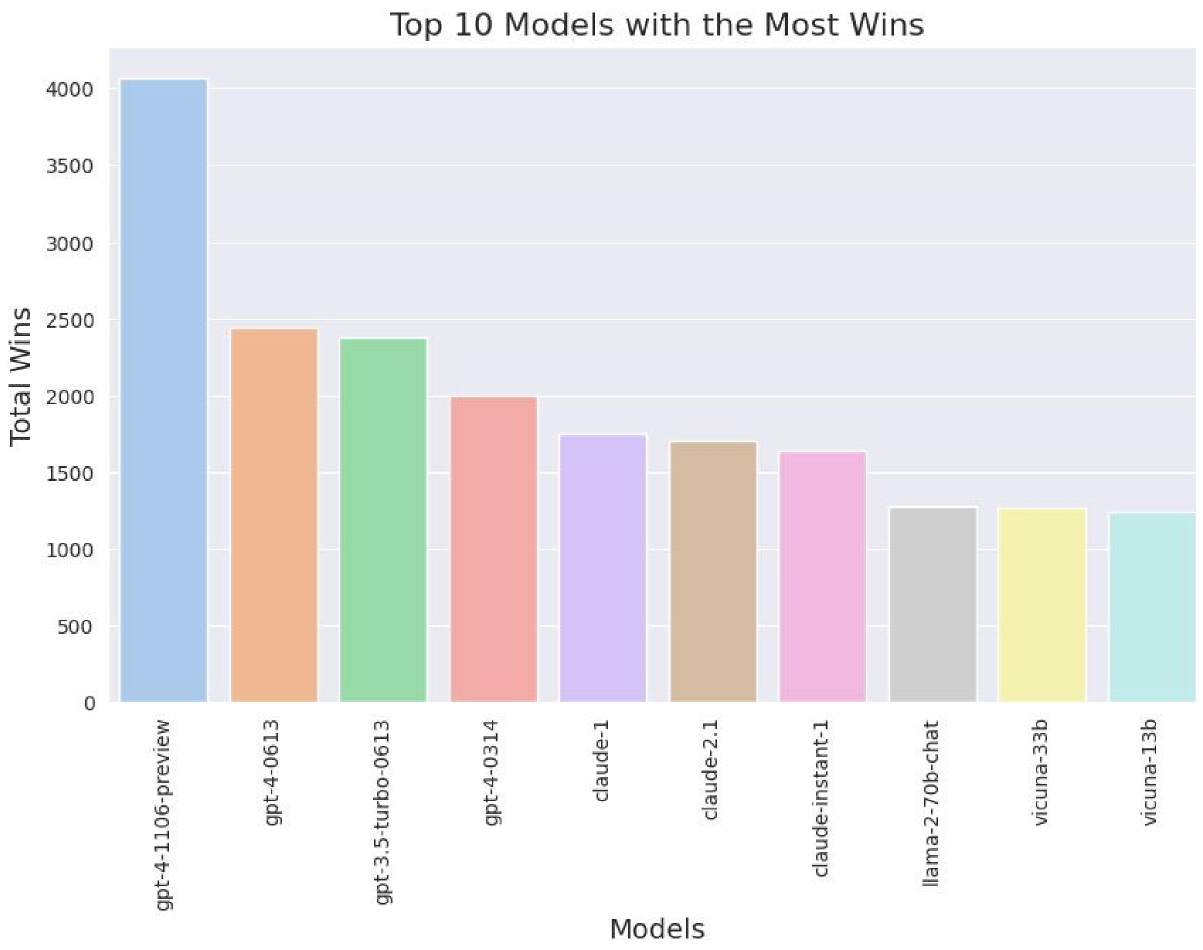
Training Data

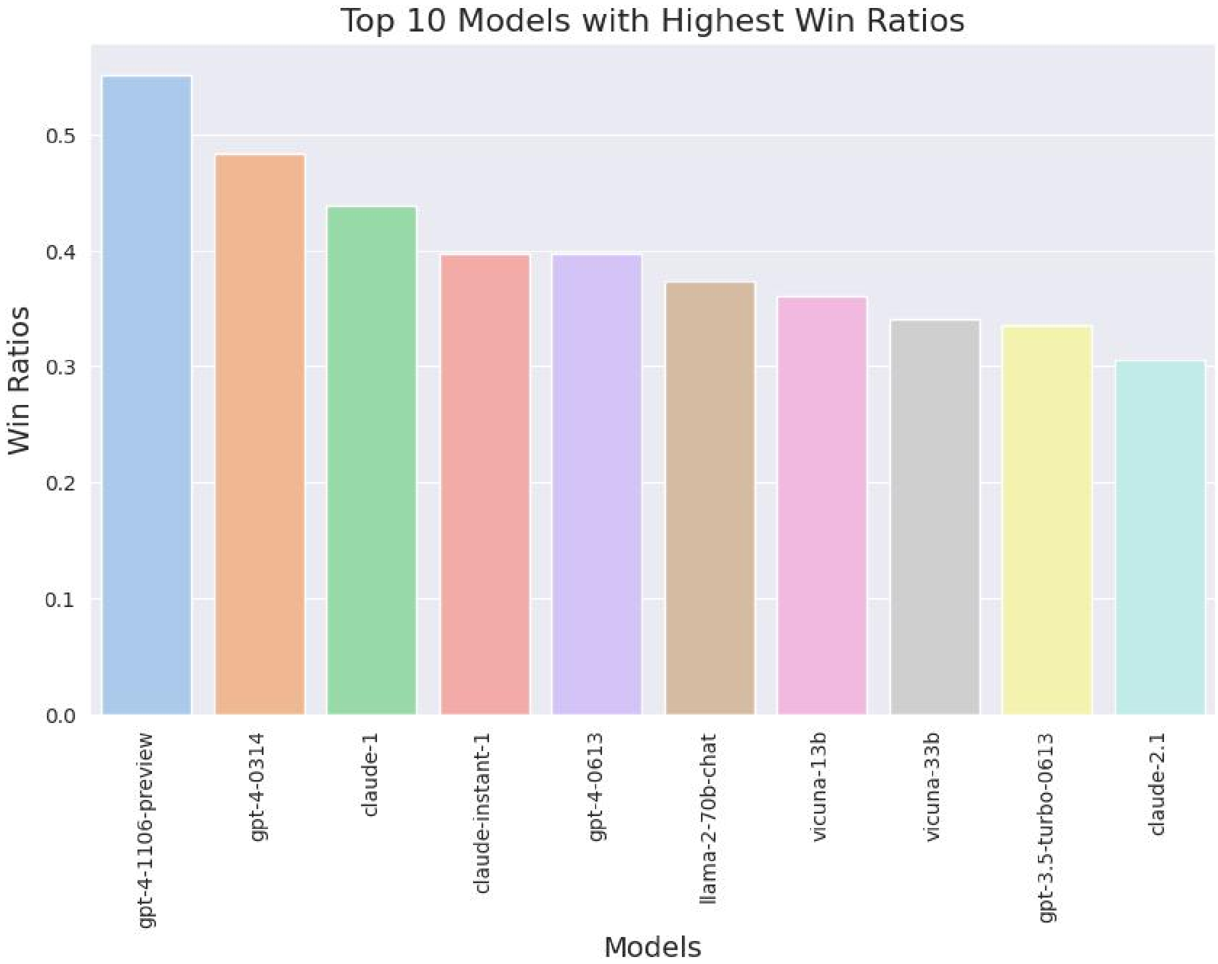
	id	model_a	model_b	prompt	response_a	response_b	winner_model_a	winner_model_b	winner_tie
0	30192	gpt-4-1106-preview	gpt-4-0613	["Is it morally right to try to have a certain...	["The question of whether it is morally right ...	["As an AI, I don't have personal beliefs or o...	1	0	0
1	53567	koala-13b	gpt-4-0613	["What is the difference between marriage lice...	["A marriage license is a legal document that ...	["A marriage license and a marriage certificat...	0	1	0
2	65089	gpt-3.5-turbo-0613	mistral-medium	["explain function calling. how would you call...	["Function calling is the process of invoking ...	["Function calling is the process of invoking ...	0	0	1
3	96401	llama-2-13b-chat	mistral-7b-instruct	["How can I create a test set for a very rare ...	["Creating a test set for a very rare category...	["When building a classifier for a very rare c...	1	0	0

Test Data

	id	prompt	response_a	response_b
0	136060	["I have three oranges today, I ate an orange ...	["You have two oranges today."]	["You still have three oranges. Eating an oran...
1	211333	["You are a mediator in a heated political deb...	["Thank you for sharing the details of the sit...	["Mr Reddy and Ms Blue both have valid points ...
2	1233961	["How to initialize the classification head wh...	["When you want to initialize the classificati...	["To initialize the classification head when p...

Exploratory Data Analysis

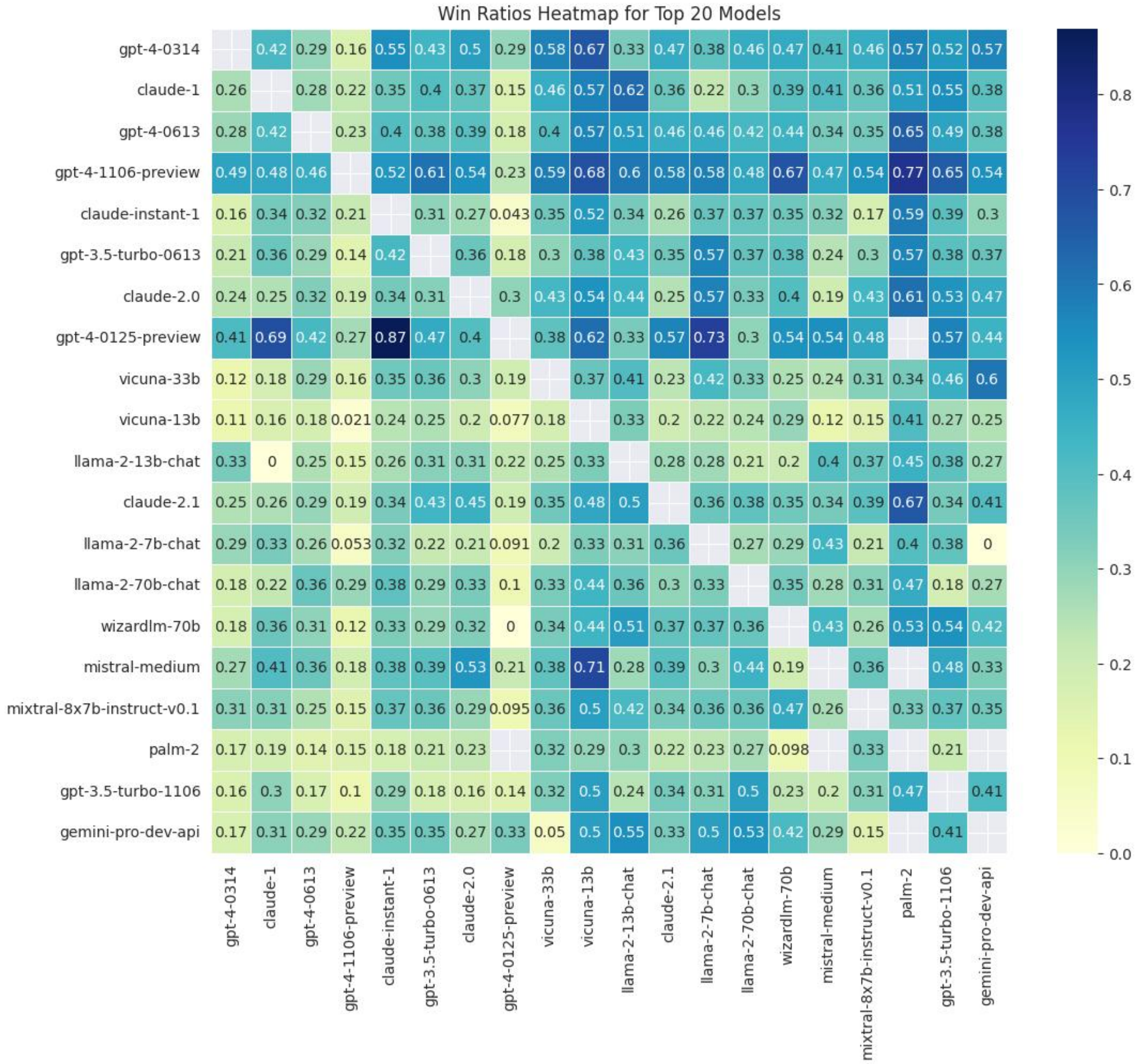




Exploratory Data Analysis

- **GPT-4 Models' Dominance**:
- The GPT-4 models (gpt-4-0314, gpt-4-0613, gpt-4-1106-preview, gpt-4-0125-preview) consistently show high win ratios against other models.

- **Consistent Performance:**
- claude-instant-1 and claude-2.0 also demonstrate strong performance with notable win ratios against several other models, except GPT models.



Model Configuration

Model Name: **microsoft/deberta-v3-xsmall**

Number of Labels: 3

Learning Rate: 2e-5

Weight Decay: 0.01

Warmup Steps: 10% of the total training steps

Training Epochs: 4

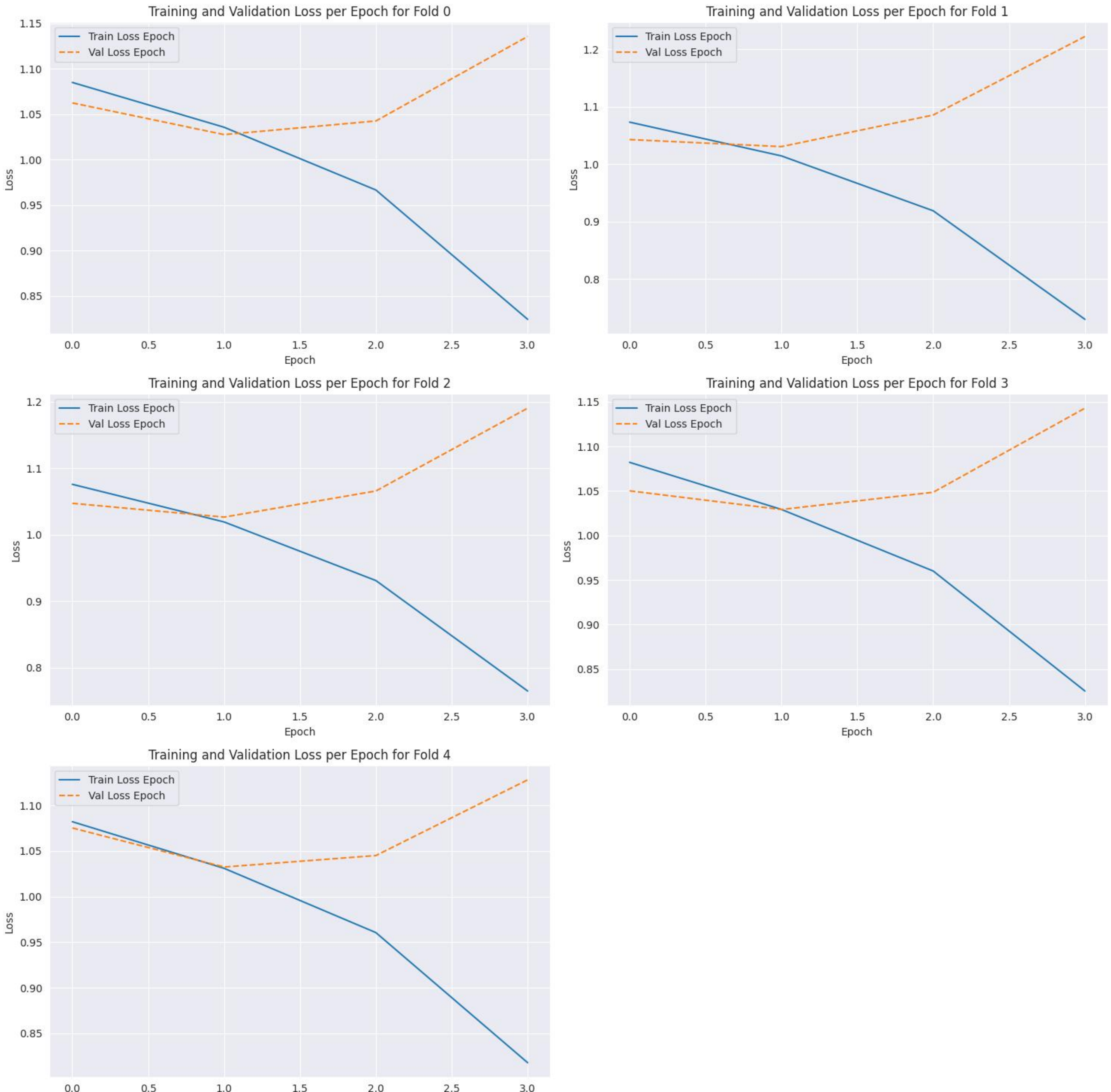
Batch Sizes:

- Train Batch Size: 16
- Evaluation Batch Size: 4

Model Evaluation

The model displayed effective learning during training.

The validation loss did not decrease as significantly, suggesting that the model could improve in generalization and better handle unseen data.



Conclusions and Discussion

While the model performed well, there are several areas for potential improvement:

- **Hyperparameter Tuning:** Further tuning of hyperparameters such as learning rate, batch size, and weight decay might yield better results. Techniques like grid search or Bayesian optimization could be employed.
- **Regularization Techniques:** Implementing additional regularization methods such as dropout or weight decay adjustments could help prevent overfitting.
- **Data Augmentation:** Augmenting the training data with techniques like back-translation or synonym replacement could enhance the model's ability to generalize.
- **Model Ensemble:** Combining the predictions of multiple models (ensemble learning) could improve the overall performance by reducing variance and bias.
- **Advanced Architectures:** Exploring more advanced or larger transformer models, or fine-tuning specifically on similar task datasets, could yield performance gains.
- **Extended Training:** Increasing the number of training epochs or using early stopping based on validation loss could help the model converge better and avoid overfitting.

References

1. Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (Year). Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. https://arxiv.org/pdf/2403.04132.

2. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., & Stoica, I. (Year). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv. https://doi.org/10.48550/arXiv.2306.05685

3. He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv. https://doi.org/10.48550/arXiv.2006.03654

4. He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv. https://doi.org/10.48550/arXiv.2111.09543

5. Chiang, W., Zheng, L., Dunlap, L., Gonzalez, J. E., Stoica, I., Mooney, P., Dane, S., Howard, A., & Keating, N. (2024). LMSYS - Chatbot Arena Human Preference Predictions. Kaggle. Retrieved from https://kaggle.com/competitions/lmsys-chatbot-arena

Thank You!