Fine-Tuning Gemma-2 with LoRA for Response Prediction Aligning with Human Preferences

Bocheng Dai, Tiantian Meng, Suye Shen

Mentor: Mingjian Jiang

PROBLEM & BACKGROUND

Motivation

Aligning chatbot responses with human preferences is a critical challenge in conversational AI. LLMs power chatbots but struggle with systematic biases, misalign model responses with human preferences.

Key biases include:

Verbosity Bias

Prompt:
"What are the main causes of climate change?"

Response A:
"Climate change is primarily caused by the emission of greenhouse gases, notably carbon dioxide, methane, and nitrous oxide, which result from human activities like burning fossil fuels, deforestation, and intensive agriculture."

Response B:
"The main cause of climate change is human activity."

Bias Explanation:
Despite Response B being concise and accurate, a model suffering from verbosity bias mistakenly **favor Response A** simply because it's longer, regardless of whether users prefer succinctness.

Position Bias

Prompt:
"What is the capital city of Australia?"

Ordering 1:
Response A: "Sydney is a major city but not the capital of Australia; Canberra is the capital."
Response B: "Canberra is the capital of Australia."

Ordering 2:
Response A: "Canberra is the capital of Australia."
Response B: "Sydney is a major city but not the capital of Australia; Canberra is the capital."

Bias Explanation:
Position bias occurs if the model consistently favors the earlier response regardless of content. Ideally, the prediction of human preference should remain stable when responses switch order. A biased model might **disproportionately prefer the first response** in each scenario.

Self-Enhancement Bias

Prompt:
"Can antibiotics cure viral infections?"

Response A (Overly confident but incorrect):
"Yes, antibiotics are highly effective at curing all kinds of infections, including viruses."

Response B (Accurate but cautious):
"No, antibiotics are effective against bacterial infections but do not cure viral infections."

Bias Explanation:
A model influenced by self-enhancement bias might mistakenly **favor the overly confident Response A**, despite its factual inaccuracy, due to its assertive tone, rather than the cautious but correct Response B.

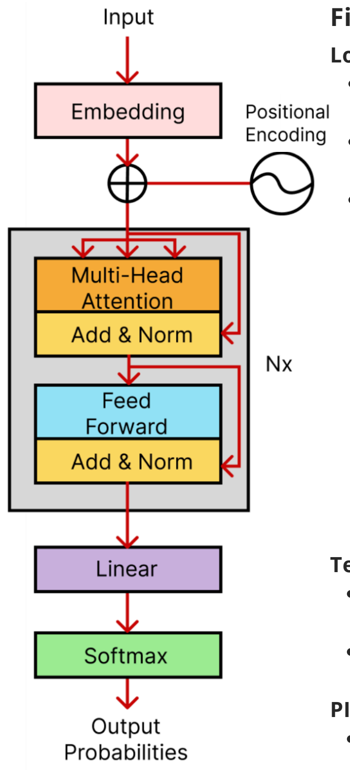
Existing Approaches & Limitations

- Traditional **feature-based ranking** models like XGBoost provide interpretability but **lack deep contextual reasoning**.
- Pretrained LLMs (e.g., Gemma-2) capture contextual nuances but remain susceptible to verbosity, position, and self-enhancement biases.

Our Goal

We **fine-tune Gemma-2-2B-it** with **LoRA** (Low-Rank Adaptation), **PISSA** Initialization, and **Test-Time Augmentation** (TTA) to improve chatbot preference prediction accuracy and mitigate systematic biases.

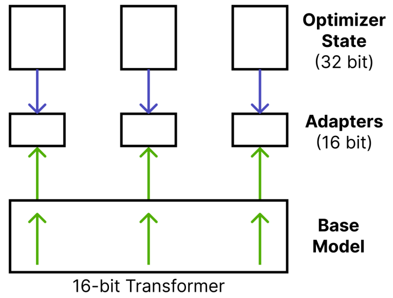
METHOD



Fine-tuning Gemma-2-2B-it

LoRA Integration

- Applied LoRA adapters to key transformer components.
- LoRA parameters: Rank = 64, Scaling Factor = 64, Dropout = 0.05.
- Only trainable layers are updated, keeping the base model mostly frozen.



Test-Time Augmentation (TTA) for Robustness

- Swaps response orders during inference to reduce position bias.
- Averages predictions across augmented samples to enhance preference prediction stability.

PISSA Initialization

- Optimized SVD-based initialization for LoRA to enhance convergence speed.

EXPERIMENT

Task

Given a **user prompt** and **two chatbot-generated responses**, predict which response a human prefers:

Data Input

Prompt:
"What is the scientific probability of earth being the only habitable planet in the universe?"

Claude-1's Response:
"The scientific probability of Earth being the only habitable planet in the universe is very low, for several reasons...."

GPT-3.5-turbo-1106's Response :
"The scientific probability of Earth being the only habitable planet in the universe is currently unknown. Given the...."

Human-Labeled Preference

Human Preference:
- winner_model_a: 0
- winner_model_b: 1
- winner_tie: 0

Model Output

Our model is expected to output probabilistic predictions indicating the likelihood of each response being preferred:

Model Prediction:
- winner_model_a: 0.40
- winner_model_b: 0.55
- winner_tie: 0.05

Evaluation Metrics

- **Accuracy:** Measures the proportion of correct predictions, assessing how well the model aligns with human preferences.
- **Log Loss:** Captures prediction confidence—lower values indicate better calibration and reliability of probability estimates.

Data

Primary Dataset

- LMSYS Chatbot Arena Human Preference Dataset
- 57,485 examples (80:20 train-test split)

Bias-Specific Validation Dataset

- Created using RLHF techniques + manual review for quality assurance.
- 6,000 examples (2,000 per bias type)

Baseline

Pretrained Gemma-2-2B-it

- No fine-tuning, serves as a pure LLM baseline.

LMSYS XGBoost Baseline

- Feature-based ranking model (structured learning).
- Helps analyze whether feature-based ranking models still perform well compared to fine-tuned LLMs.

RESULTS & ANALYSIS

Setting	Accuracy	Log Loss
Fine-Tuning with TTA, PISSA Initialization, Three-Scoring Layers	0.4726	1.0522
Fine-Tuning with No TTA, PISSA Initialization, Three-Scoring Layers	0.4341	1.0715
Fine-Tuning with TTA, No PISSA Initialization, Three-Scoring Layers	**0.4922**	**1.0373**
Fine-Tuning with TTA, PISSA Initialization, Single-Scoring Layer	0.3129	1.2574
XGBoost Benchmark (Feature Engineering)	0.4629	1.0582
No Fine-Tuning (Pretrained Gemma-2-2B-it)	0.2978	3.0237

Why did PISSA underperform?

- Likely over-constrained transformer layers, reducing adaptability.
- Shows that weight initialization optimizations don't always generalize well, emphasizing empirical validation over theoretical gains.

Why was XGBoost so competitive?

- Neural models were expected to outperform, but the margin was smaller than anticipated.
- Smaller LLMs (Gemma-2-2B-it) may not fully leverage deep contextual representations, making feature engineering a viable alternative.

Why do deeper scoring layers matter?

- Single-layer classifiers performed significantly worse, proving multi-layer architectures are essential for capturing human preference signals.

What does this mean for chatbot preference prediction?

- Strategic fine-tuning (LoRA + TTA) is crucial for aligning LLMs with human preferences.
- XGBoost remains strong, but larger LLMs + RLHF could push performance further.
- TTA effectively mitigates bias at no extra training cost, improving robustness.

CONCLUSION

Key Findings

- **LoRA + TTA (No PISSA)** achieved the best performance, confirming that robustness-enhancing augmentation (TTA) improves preference prediction while PISSA constraints reduce adaptability.
- **Multi-layer scoring architectures** significantly outperformed single-layer models, proving crucial for effective preference modeling.
- **Feature-based models like XGBoost remain strong competitors**, suggesting that LLM-based approaches still have room for improvement under computational constraints.

Limitations & Future Work

- **Scaling to larger models:** Extending this analysis with Gemma-2-9B-it to assess the impact of model size on preference prediction accuracy.
- **Optimizing LoRA fine-tuning:** Exploring alternative LoRA initialization techniques to improve stability and convergence.
- **Enhancing Bias Mitigation:** Developing data augmentation strategies beyond TTA to further reduce systematic biases in preference modeling.

REFERENCES

- Google AI. Gemma: Open weight efficient models, 2024.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zhi Hern Loh, Vikram N. Rangamani, Raghavendran G. Swaminathan, Christopher Ré, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2022.
- Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models, 2024.