Response Helpfulness Evaluation with Deep Learning

Tejas Chandramouli, Sailakshmi Gangisetty, Samhitha Pudipeddi, Sarah Rayen, Arya Shankardas

Project Overview

01 Dataset

NVIDIA HelpSteer3 (contains detailed human annotations of response preferences)

02 Model

DistilBERT-based neural network designed for precise scoring

03 Application

Interactive and user-friendly scoring app developed using Gradio

Method and Model

Data Preparation & Modeling

Tokenized texts and created preference labels on -3 to +3 scale, normalized during training

Built neural architecture with DistilBERT encoder

Used mean pooling for better text representation

Implemented preference-based MSE loss function

Model Evaluation & Enhancements

Evaluated using accuracy metrics: Direction accuracy and Spearman correlation

Addressed score compression issue

Enhanced scoring clarity through linear normalization

Performance Results

Direction Accuracy: 63% (how frequently our model aligns with human preference)

Spearman Correlation: 0.35–0.45 (consistency between model predictions and human scores)

Interactive Demo Application

Built using Gradio for easy accessibility

Enables direct comparison of responses and visualizes scoring distributions

Future Directions

Evaluate larger models (e.g., RoBERTa, BERT-large) to improve performance

Implement data augmentation strategies

Integrate scoring system with reinforcement learning (RLHF)

Results & Demo

THANK YOU