

Predicting Human Preferences for LLM Response Enhancement

20203955 Park WonKyu, 20214677 Lee YuJung, 20201799 Jung SeungHwan

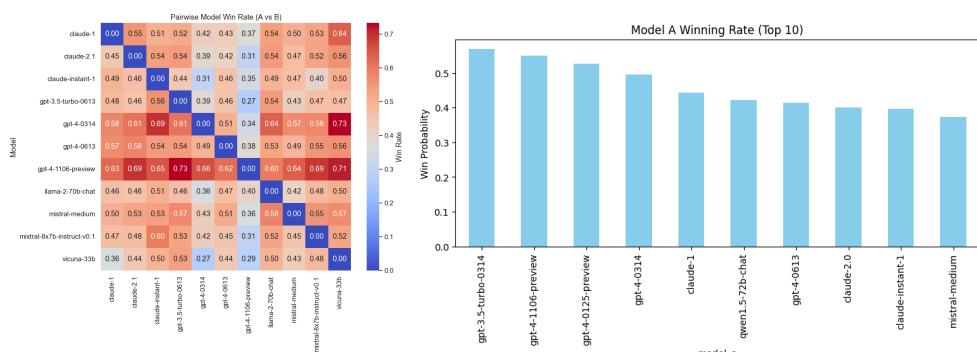
Abstract

In this project, we participated in the LLM Classification Finetuning Kaggle competition with the goal of accurately predicting human preferences between two competing model responses. Our initial approach established a hybrid lexical–semantic framework, leveraging the SentenceTransformer (*all-MiniLM-L6-v2*) to combine handcrafted lexical features with semantic embeddings within a stacking ensemble of heterogeneous classifiers. While this ensemble achieved high efficiency and an accuracy (≈ 0.48), it struggled to capture fine-grained contextual reasoning and differentiate nuanced ties.

To address this limitation, we introduced a DeBERTa-v3-small + LoRA fine-tuning model, enabling deeper semantic comparison through end-to-end representation learning. Although the hybrid ensemble achieved slightly higher overall accuracy, the LoRA-based model was strategically chosen for its ability to generate more balanced and semantically discriminative predictions, particularly in tie or contextually similar cases. Building on the DeBERTa model, we conducted an in-depth error analysis to uncover its structural and reasoning limitations and proposed future extensions to enhance contextual understanding and alignment with high-level human preference signals.

Data Overview

The dataset for the competition comprises 57,477 pairs of LLM responses to user prompts, each labeled by human preference (A-win, B-win, or Tie), and is globally balanced across labels. The responses were generated by 11 distinct models, among which the GPT-4 family dominated with a 0.6–0.7 win rate.



Each data instance includes a prompt and two responses (`response_a`, `response_b`), which were concatenated into a single input sequence to train a model capable of discerning subtle preference differences. The average sequence length was approximately 607 tokens, occasionally leading to long processing times or exceeding the model’s maximum token limit, which required additional consideration during model design.

Modeling

Step 1: Baseline Model

Using handcrafted lexical and structural features (length, punctuation, and sentence counts), a simple Logistic Regression baseline was trained with an 80–20 split.

The model achieved a validation accuracy of 0.44, log loss of 1.07, and macro F1-score of 0.35, showing reasonable separation between A-win and B-win but poor tie detection. This result

established the lexical-only baseline for subsequent hybrid and contextual extensions.

Step 2: Embedding-based Model

Next, a semantic-only embedding model was built to assess the standalone effectiveness of contextual representations without any lexical or stylistic features.

We adopted the *all-MiniLM-L6-v2* SentenceTransformer for its strong balance between semantic expressiveness and

computational efficiency, making it suitable for large-scale pairwise comparisons.

Each prompt-response pair was independently embedded, and the difference between response embeddings was used as the input feature.

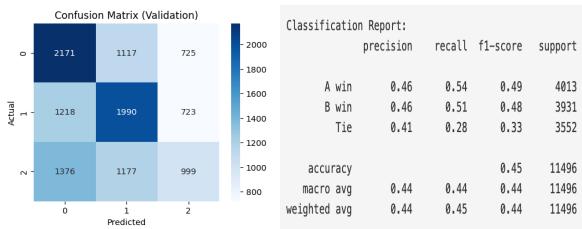
While this setup efficiently captured semantic similarity, it achieved only 0.41 accuracy and 0.36 macro F1.

Step 3-1: Feature-Enhanced Lexical Model

Following the *1. baseline* experiments, we further constructed a feature-enhanced ensemble trained on the engineered lexical features.

Hypothesis	Evidence	Conclusion
H1. Longer responses are preferred.	$t = 47.1, p < 0.001$	Significant verbosity bias — humans favor longer answers even when model identity is controlled.
H2. Punctuation-rich responses win more.	Kruskal $p < 0.001$	Annotators prefer responses with higher punctuation density, implying richer narrative structure.
H3. Linguistic complexity correlates with preference.	$\text{lexical_div_diff } p < 1e-200$	Greater word diversity and lower repetition correlate with higher win probability.
H4. Certain words influence preference.	company ($p = 0.001$), brace ($p = 0.018$), apologize/ sorry ($p < 0.001$, negative)	“Company”, “Brace”, “Knee”, “Progression” words predict higher win; apology-type terms correlate negatively.

Using the scaled lexical feature matrix, data were partitioned via stratified splitting to maintain class balance for validation. The model combined four heterogeneous classifiers—Logistic Regression, Random Forest, Gradient Boosting, and LightGBM—under a **soft-voting** ensemble. This ensemble leveraged complementary strengths: the interpretability of statistical models, the variance reduction of bagging, and the error-correcting capacity of boosting, thereby capturing both linear and nonlinear lexical relationships.



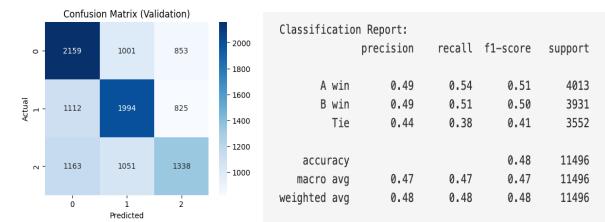
The trained model achieved a validation accuracy of 0.4489 and a Log Loss of 1.0508. Although

recall was balanced between A-win and B-win, the confusion matrix revealed notable overlap, suggesting that surface-level lexical features alone could not fully capture **nuanced semantic preferences**, despite offering some stylistic discrimination.

Step 3-2: Hybrid Model with Stacking Ensemble

To address the surface-level limitations of the lexical ensemble, we built a hybrid model that integrates lexical–stylistic features with semantic embeddings from pre-trained language models.

A Stacking Ensemble was implemented this time, utilizing Logistic Regression, Random Forest, and LightGBM as base learners, with a multinomial Logistic Regression meta-learner to maximize the capture of lexical–semantic interactions. Unlike simple Voting, Stacking models the inter-feature dependencies and balances the complementary strengths.



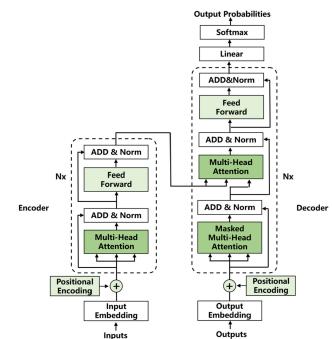
After hyperparameter tuning, the hybrid model achieved a validation accuracy of **0.48** and a macro F1-score of **0.47**, outperforming the lexical-only ensemble (0.45). The confusion matrix showed improved recall for Tie cases and reduced overlap between A-win and B-win, demonstrating stronger contextual and semantic discrimination.

Step 3-3: DeBERTa-v3-small + LoRA Fine-Tuning (Calibration)

While the hybrid model combined lexical and semantic features, it lacked token-level interaction between paired responses.

To address this, we adopted **DeBERTa-v3-small with LoRA fine-tuning**,

enabling end-to-end contextual comparison within a shared attention space. LoRA efficiently updates only low-rank attention weights, preserving the pretrained model’s contextual capacity while keeping the training lightweight.

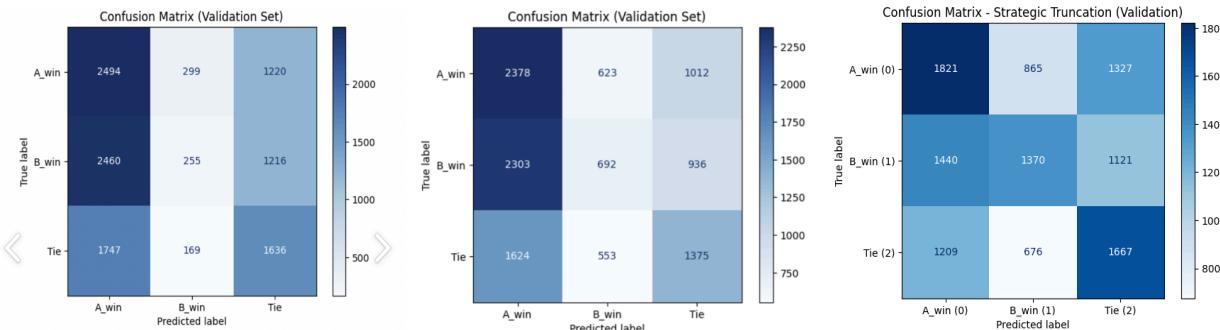


However, the initial model exhibited a strong A_win prediction bias and a severe drop in B_win recall [Figure 1], primarily due to the 512-token limit, which truncated essential parts of longer responses. Even with hyperparameter and LoRA adjustments, improvements remained marginal because of structural information loss [Figure 2].

To mitigate this, we introduced **Strategic Truncation**—retaining the prompt’s head (core question) and responses’ tails (final arguments)—and applied Symmetric Data Augmentation to balance A/B positional bias. Although these adjustments slightly improved

B_win recall, the token constraint fundamentally persisted.

This refined pipeline partially mitigated the token limitation and strengthened discriminative learning for fine-grained preference differences. Recognizing that the model’s raw output probabilities were skewed as a result, we subsequently applied **Isotonic Regression calibration**. This post-hoc calibration corrected probability distortion and improved overall prediction reliability, yielding roughly **10%** relative improvement in validation accuracy and a lower **log-loss** score.



Validation strategy and results

All experiments used an 80–20 stratified train–validation split to maintain label balance across A-win, B-win, and Tie cases. Random seeds were fixed for reproducibility, and models were evaluated using Accuracy, Macro F1, and Log Loss, consistent with the Kaggle metric. Confusion matrix analysis further helped interpret model behavior and identify class-specific biases.

Model	Feature Type	Accuracy	LogLoss	Macro F1	Kaggle Score	Notes
Baseline (LR)	Lexical only	0.44	1.07	0.35	1.08	Simple lexical + length features
Embedding	Semantic only	0.41	1.07	0.36	1.08	all-MiniLM-L6-v2
Hybrid Stacking	Lexical + Semantic	0.48	1.03	0.47	1.075	Semantic embeddings (MiniLM) + Stacking
DeBERTa + LoRA	Contextual Fine-tuning	0.42	1.07	0.42	1.09	Calibration with isotonic regression

	notebook4fe338d67d - Version 2 Succeeded · 6d ago	1.07637		simple LoRA - embedding Succeeded · 4h ago · Notebook simple LoRA Version 11	1.07671
	notebook4fe338d67d - Version 6 Succeeded · 7h ago · Notebook notebook4fe338d67d Version 6	1.09611		Fork of simple LoRA bbfe79 - Version 1 Succeeded · 1h ago · Notebook Fork of simple LoRA bbfe79 Version 1	1.07524

Although the Hybrid Stacking model achieved the highest validation accuracy and macro F1 overall, its improvement was largely driven by clearer separability in binary cases (A-win vs B-win), rather than true semantic understanding. In contrast, Tie prediction—where two responses are contextually comparable—represents a far more challenging and human-aligned capability, requiring the model to discern subtle equivalence in reasoning depth, coherence, and contextual appropriateness.

The DeBERTa + LoRA model after isotonic calibration exhibited more balanced probability distributions and superior recognition of Tie instances, indicating a deeper grasp of the probabilistic and context-dependent nature of human judgment. Therefore, despite slightly lower aggregate metrics, it was selected as our final model for its stronger alignment with human-like reasoning and greater potential for scalable extension.

Error & Bias Analysis with DeBERTa + LoRA(final model)

Despite efforts to alleviate the 512-token limit and the resulting information loss, deeper error analysis shows that the DeBERTa-small model fundamentally prioritizes surface-level fluency and structural clarity over contextual reasoning and practical utility—core dimensions of human preference.

ID	Prompt	Response A	Response B	True Label	Pred Label	Key Observation
164264 1931	<i>"How long should I wait for a bus before deciding it's not coming?"</i>	"You should wait for a bus for 5 minutes before deciding it's not coming."	"It depends on several factors such as time of day, location, bus frequency, and weather... generally wait 10–15 minutes, or check transit updates."	B_win	A_win	Model favored the literal answer that directly addressed the question, missing the value of practical and conditional reasoning.
232319 8756	<i>"There are 13 birds sitting in a cherry tree. A hunter shoots two dead. How many are left?"</i>	"If there were initially 13 birds and the hunter shoots and kills 2, then there are $13 - 2 = 11$ birds left."	"There are no birds left. The gunshot scared the rest away."	B_win	A_win / Tie	Model favored structured reasoning (<i>therefore</i> , numbered steps) and arithmetic clarity over real-world reasoning and commonsense validity.

In Case 1 (ID 1642641931, Bus Wait Time), the model misjudged the short, definitive answer as equally valid, overlooking the superior conditional reasoning of the true winning response that adapted to real-world variability such as time and frequency.

In Case 2 (ID 2323198756, Birds in a Tree), it preferred an arithmetically neat but contextually unwanted answer, failing to recognize the common-sense inference that all birds would fly away after the gunshot.

Across both examples, the model demonstrates reliance on syntactic completeness and explicit logical cues, rather than discerning the depth, adaptability, and real-world validity underlying human judgment.

Limitation & Future works

The DeBERTa-small + LoRA approach provided an efficient and lightweight fine-tuning framework, yet its performance was fundamentally limited by the 512-token truncation constraint and fixed prompt-response pairing, leading to persistent information loss in longer inputs despite strategic head–tail truncation. In addition, restricted computational resources constrained LoRA hyperparameter tuning to a narrow range, preventing full optimization.

Error analysis further revealed that the model struggled to capture high-level human preference features—such as deep contextual reasoning, practical utility, and creative nuance—often misjudging riddle validity or conditional reasoning quality.

Future work should address these structural limitations by adopting longer-context architectures (ex. Longformer) to eliminate truncation, refining LoRA hyperparameter optimization and output calibration, and integrating human feedback alignment techniques to better model the qualitative, utilitarian, and creative aspects underlying human preference.

Reproducibility Notes

- Environment: Python 3.10.14 (Kaggle GPU – NVIDIA Tesla P100 16 GB)
- Frameworks: transformers 4.57.1, datasets 4.0.0, peft 0.17.1, torch 2.6.0
- Random Seed: 42 (NumPy, PyTorch, Scikit-learn)
- Runtime: ≈ 2 min inference + few calibration
- Reproducibility: All experiments are fully reproducible in Kaggle’s no-internet GPU environment. The dataset and base model (DeBERTa-v3-small) are pre-mounted under /kaggle/input/, and our LoRA fine-tuning was directly executed in the same offline Kaggle environment to produce the custom checkpoint (checkpoint-22992).