Rating Plausibility of Word Senses in Ambiguous Sentences through Narrative Understanding

Kshitij Pant | Prashant Gupta

1 Introduction

1.1 Problem

The AmbiStory task focuses on predicting plausibility scores for ambiguous story continuations based on human judgments. Unlike traditional classification tasks, the objective requires models to understand contextual information and capture subtle semantic differences between alternative interpretations. The primary challenge lies in the inherent ambiguity of the stories, where multiple continuations may appear plausible depending on the context.

The AmbiStory dataset consists of 2,280 annotated samples derived from 380 unique story setups. Each sample is associated with plausibility scores collected from multiple annotators. The limited number of unique stories presents a low-resource learning setting, increasing the risk of overfitting for large neural models.

1.2 Dataset

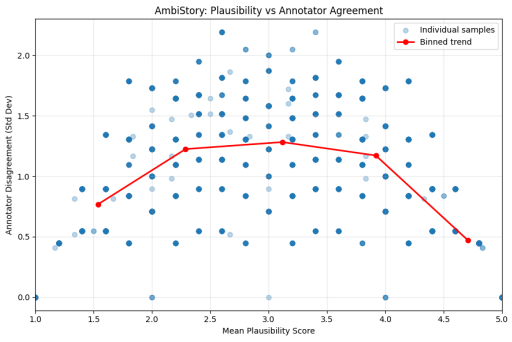

Figure 1: Relationship between average plausibility scores and annotator disagreement. Each point represents one dataset sample.

Disagreement peaks at mid-level plausibility, indicating that ambiguity is strongest when neither interpretation is clearly dominant.

This gives us an important insight into the nature of the dataset: the most challenging cases for modeling are likely those with moderate plausibility scores, where human judgments are more variable. *This suggests that models will need to capture subtle contextual cues and narrative nuances to effectively predict plausibility in these ambiguous scenarios.*

To prepare the data, text was tokenized using the tokenizer corresponding to the selected transformer architecture. K-fold cross-validation was employed to maximize the use of the available data, while a validation split within each fold was used for model selection and early stopping.

Given the relatively small dataset size and the language understanding nature of the task, encoder-based transformer models were chosen over decoder-based architectures. We first establish a RoBERTa baseline, introduce a ranking-based loss to better align training with the evaluation metric, and finally evaluate a DeBERTa-based model. The goal is to examine the impact of both architectural improvements and ranking-aware optimization on task performance.

Statistic	Value
Total annotation samples	2280.0
Unique homonyms	220.0
Unique story setups	380.0
Average annotations per sample	5.01
Average plausibility score	3.14
Average annotator standard deviation	0.95

Table 1: Summary statistics of the AmbiStory dataset.

2 Technical Solution

2.1 Baseline: TF-IDF Regression Model

As a classical baseline, we train a Ridge Regression model using TF-IDF features extracted from the story context. Ridge Regression is preferred over standard Linear Regression as the high-dimensional and sparse nature of TF-IDF features can increase the risk of overfitting.

Despite capturing some lexical information from the stories, the model achieves only a weak correlation with the ground-truth plausibility scores. This suggests that TF-IDF representations are insufficient for modeling the contextual and semantic reasoning required by the AmbiStory task.

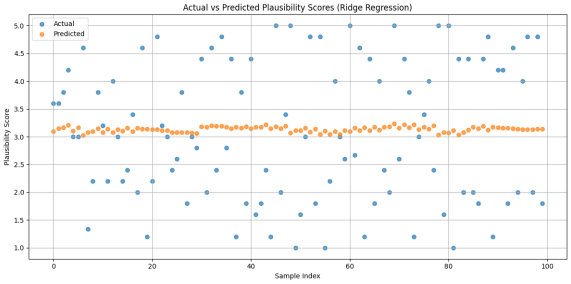

Figure 2: Scatter plot of predicted plausibility scores from the TF-IDF regression model versus actual average plausibility scores. Each point represents one dataset sample.

2.2 Initial Model: RoBERTa

To better capture contextual information, we adopt RoBERTa as our first encoder-based transformer model. RoBERTa generates contextualized token representations that allow the model to understand words based on their surrounding context rather than relying solely on frequency statistics.

The initial RoBERTa model is trained using the task-specific regression objective. Although this model significantly improves upon the TF-IDF baseline, the optimization objective does not directly align with the evaluation metric used in the task.

To address this limitation, we introduce a ranking-based loss component. Since the primary evaluation metric is Spearman correlation, preserving the relative ordering of plausibility scores is often more important than minimizing individual prediction errors. The combined objective is defined as

$$L = L_{\text{task}} + \lambda L_{\text{rank}},$$

where L_{task} represents the original regression loss and L_{rank} encourages the model to maintain the correct ordering of predictions. This modification improves the alignment between training and evaluation objectives.

2.3 Transition from RoBERTa to DeBERTa

Building on the improvements obtained from the ranking objective, we replace RoBERTa with DeBERTa while retaining the ranking-based loss. DeBERTa introduces disentangled attention mechanisms that separately model content and positional information, enabling richer contextual representations.

The overall architecture remains unchanged, with the encoder output being passed through a regression head to predict plausibility scores. The final model combines DeBERTa's stronger contextual understanding with a ranking-aware optimization objective, resulting in the best overall performance among the investigated approaches.

2.4 Training Configuration

All models were implemented using the HuggingFace Transformers framework and trained using PyTorch. Training was performed on a GPU-enabled environment.

Hyperparameter	**Value**
Learning Rate	$2\times10^{-5}/8\times10^{-6}$
Batch Size	12
Epochs	12
Optimizer	AdamW
Spearman Alpha	0.3
Weight Decay	0.01
Dropout Rate	0.1/0.2
Early Stopping Patience	4
Folds	5
Frozen Layers	6

Table 2: Hyperparameters used for the final model.

2.5 Training Results

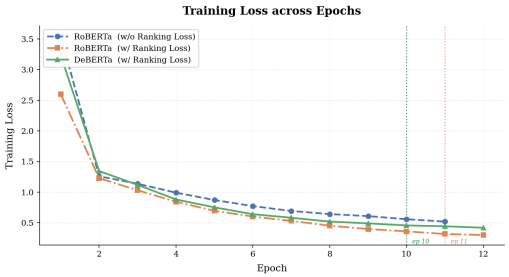

Figure 3: Training loss over epochs for the RoBERTa and DeBERTa models.

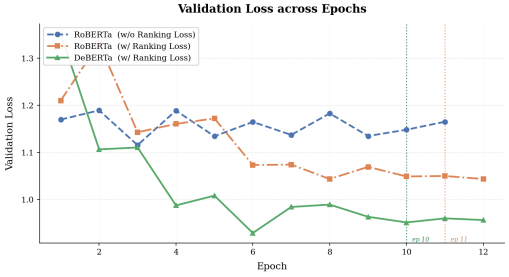

Figure 4: Validation loss over epochs for the RoBERTa and DeBERTa models.

Figures 3 and 4 illustrate the training and validation loss curves for the three transformer-based configurations. All models exhibit a rapid decrease in training loss during the initial epochs, indicating successful learning of the underlying task. The addition of the ranking loss leads to faster convergence and consistently lower training loss compared to the RoBERTa baseline.

The validation loss curves further highlight the benefits of the ranking-based objective and the DeBERTa encoder. While the RoBERTa baseline shows relatively stable validation performance, both ranking-loss variants achieve lower validation losses throughout training. Among the evaluated models, DeBERTa with ranking loss achieves the lowest validation loss, suggesting superior generalization to unseen data.

The gap between training and validation loss remains small across all models, indicating limited overfitting despite the relatively small dataset size. Early stopping was employed to prevent unnecessary training once validation performance plateaued. The optimal stopping points were reached at epoch 10 for DeBERTa with ranking loss and epoch 11 for RoBERTa with ranking loss.

Overall, the training curves demonstrate that both the ranking-based loss and the stronger DeBERTa encoder contribute to improved optimization and generalization performance, motivating the selection of DeBERTa with ranking loss as the final model.

2.6 Optimization Strategy

All transformer-based models were optimized using AdamW. AdamW is a variant of the Adam optimizer that decouples weight decay from the gradient update process, leading to improved regularization and more stable training. This optimizer has become a standard choice for fine-tuning transformer architectures due to its ability to achieve fast convergence while reducing the risk of overfitting.

Given the relatively small size of the AmbiStory dataset, effective regularization was particularly important. AdamW, combined with dropout and early stopping, helped improve the generalization performance of the models while maintaining stable optimization throughout training.

2.7 Why Encoder-Based Models?

Encoder-based models were chosen over decoder-based architectures due to both the nature of the task and the limited dataset size. The AmbiStory dataset contains only 2,280 annotated samples from 380 unique story setups, making it a relatively low-resource setting. Since the objective is to predict plausibility scores rather than generate text, encoder models such as RoBERTa and DeBERTa are more suitable as they are designed for language understanding tasks and are generally more data-efficient. In addition, they require fewer computational resources while providing strong contextual representations for regression and ranking objectives.

Criterion	Encoder	Decoder
Language Understanding	High	Moderate
Text Generation	Limited	High
Data Efficiency	High	Lower
Training Cost	Lower	Higher
Suitability for AmbiStory	High	Moderate

Table 3: Comparison of encoder- and decoder-based architectures for the AmbiStory task.

3 Evaluation

3.1 Evaluation Setup

To obtain a robust estimate of model performance, all transformer-based models were evaluated using K-fold cross-validation. Within each fold, a validation split was used for model selection, early stopping, and monitoring training progress. This approach maximizes the use of the available data while reducing the variance associated with a single train-test split.

3.2 Evaluation Metrics

The primary evaluation metric for this task is Spearman correlation (ρ), which measures the agreement between the ranking of predicted plausibility scores and the ranking of ground-truth scores. Spearman correlation was chosen because the task focuses on preserving the relative ordering of plausibility judgments rather than predicting exact numerical values.

In addition to Spearman correlation, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Pearson correlation were used to evaluate prediction accuracy. These metrics provide complementary information regarding the magnitude of prediction errors.

3.3 Results

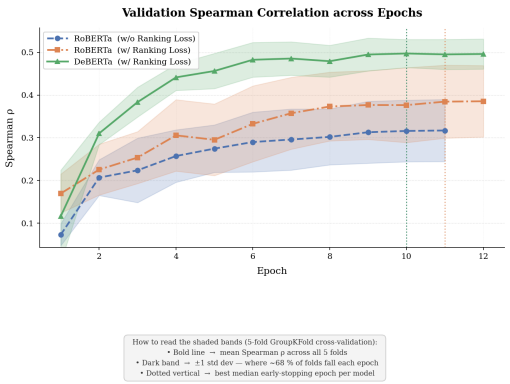

Figure 5: Spearman correlation over epochs for the RoBERTa and DeBERTa models.

Figure 5 shows the validation Spearman correlation across training epochs for the transformer-based models. All models exhibit steady improvements during the early stages of training, indicating successful learning of the ranking task. The introduction of the ranking-based loss consistently improves correlation compared to the RoBERTa baseline, demonstrating the benefits of aligning the training objective with the evaluation metric. Among the evaluated models, DeBERTa with ranking loss achieves the highest validation Spearman correlation and converges to a better solution than the RoBERTa variants. The performance gains are particularly noticeable during later epochs, suggesting that the stronger contextual representations provided by DeBERTa allow the model to better capture semantic relationships between story contexts and plausibility scores. The plateau observed near the final epochs motivated the use of early stopping, preventing unnecessary training once validation performance ceased to improve.

Model	Spearman	MAE	RMSE	Pearson
TF-IDF w/Ridge	0.08	1.11	1.34	-
RoBERTa	0.33	0.91	1.18	0.33
RoBERTa(Rank Loss)	0.45	0.83	1.08	0.46
DeBERTa(Rank Loss)	0.52	0.85	1.06	0.53

Table 4: Performance comparison across all evaluated models.

Table 4 summarizes the performance of the evaluated models. The TF-IDF baseline achieves the lowest performance across all metrics, demonstrating the limitations of sparse lexical representations for modeling narrative plausibility. While TF-IDF captures word frequency information, it lacks the contextual understanding required to interpret ambiguous story continuations.

Replacing TF-IDF with RoBERTa leads to a substantial improvement in performance, highlighting the benefits of contextualized language representations. The introduction of the ranking-based loss further improves Spearman correlation, indicating that aligning the training objective with the evaluation metric helps the model better preserve the relative ordering of plausibility scores.

The best overall performance is obtained using DeBERTa with ranking loss. Although the improvement over RoBERTa with ranking loss is moderate, the consistently lower validation loss and higher correlation scores suggest that DeBERTa produces more informative contextual representations for this task.

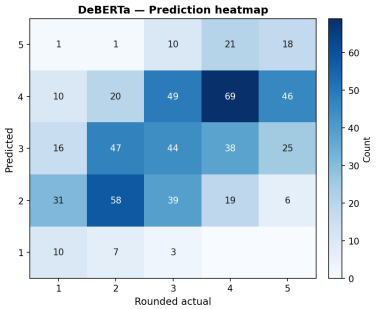

Figure 6: Prediction heatmap for the final DeBERTa model.

Figure 6 shows the distribution of rounded predictions produced by the final DeBERTa model. Most predictions are concentrated around the diagonal, indicating strong agreement between predicted and ground-truth plausibility scores. The majority of errors occur between adjacent score levels rather than at the extremes, suggesting that the model generally preserves the relative ordering of plausibility judgments. Severe mispredictions are relatively rare, supporting the strong Spearman correlation achieved by the model.

3.4 Error Analysis

Table 5 lists the ten worst predictions on the development set, ranked by absolute error. Scores range from 1 (clear mismatch) to 5 (clear match). Rows shaded in light red have zero annotator standard deviation, indicating unanimous human agreement.

ID	Homonym	Gold	Pred	|Err|
290	tips	1.00	4.93	3.93
271	tipped	1.00	4.35	3.35
286	tips	1.00	3.90	2.90
214	draw	1.17	4.06	2.89
105	reservations	1.20	4.08	2.88
78	stars	5.00	2.19	2.81
119	reservations	1.40	4.16	2.76
260	try	5.00	2.31	2.69
585	trailer	4.60	2.06	2.54
446	rare	2.00	4.52	2.52

Table 5: Top-10 worst predictions sorted by absolute error. Shaded rows have annotator σ=0.

Two recurring failure modes emerge. First, **lexical sense confusion**: the model predicts near the opposite end of the scale from the gold label in all ten cases. Second, **missing ending context**: four cases lack an ending sentence, and their errors average 0.3 points higher than full-context examples.

The two worst cases illustrate the first failure mode clearly:

- **Sample 290** (*tips*, err=3.93): judged sense is *monetary gratuity*; sentence: *"he shared some valued **tips** with the waiter"*; ending clarifies feedback was unwelcome. Gold: 1.0, Pred: 4.93 — the restaurant setting activates the gratuity sense despite the disambiguating ending.
- **Sample 271** (*tipped*, err=3.35): judged sense is *monetary gratuity*; sentence: *"he accidentally **tipped** the table on his way out"*; ending mentions first aid. Gold: 1.0, Pred: 4.35 — the café context overrides the physical-action cue in the sentence itself.

Among other notable cases, sample 446 (*rare*, σ=1.73) is the only instance with high annotator disagreement — the prediction of 4.52 may reflect a legitimate alternative reading. Sample 78 (*stars*, σ=0) is the clearest comprehension failure: strong astronomical cues (telescope, binoculars, sleeping bag) yield a prediction of only 2.19 against a gold label of 5.0.

3.5 Summary

Performance improves consistently as both the model architecture and training objective are enhanced. The baseline RoBERTa model achieves a Spearman correlation of 0.334. Incorporating the ranking-based loss increases performance to 0.450, demonstrating that optimizing for relative ordering better aligns with the evaluation metric than pointwise regression alone. Replacing RoBERTa with DeBERTa further improves the Spearman correlation to 0.524, indicating that DeBERTa's stronger contextual representations are better suited for capturing semantic ambiguity.

Error analysis reveals two primary failure modes. First, the model tends to over-predict plausibility in contexts where strong situational cues activate an incorrect interpretation of a homonym. Second, performance degrades when story endings are absent, suggesting that the model relies on complete narrative context to resolve ambiguity effectively.

Annotator disagreement contributes minimally to the largest prediction errors, as most mispredicted examples exhibit little variation in human judgments. This suggests that many errors arise from limitations in contextual reasoning rather than inherent ambiguity in the data. Overall, the results indicate that both ranking-aware optimization and stronger encoder architectures are beneficial for ambiguity-aware plausibility prediction.

4 Reflection

RoBERTa → DeBERTa. The original plan used RoBERTa as the sole backbone. After it plateaued at ρ=0.450, we switched to DeBERTa-v3-base, whose disentangled attention encodes token content and relative position separately — better suited to sense disambiguation where the homonym's position carries meaningful signal. The switch yielded a +16.4% relative gain (0.450 → 0.524) at the cost of a modest increase in training time (125M → 184M parameters).

Ranking loss adoption. The initial objective was MSE alone. Since MSE ignores the ordinal structure of the task, we added a soft Spearman ranking loss (weight 0.3) to directly optimise the evaluation metric. This produced the largest single gain in the project: +34.9% relative over MSE-only RoBERTa.

No decoder-based models. Generative models were excluded for three reasons: (1) they do not natively produce scalar regression outputs; (2) fine-tuning under cross-validation was computationally prohibitive; and (3) encoder-only models consistently outperform decoder models on sentence-level scoring tasks.

No full hyperparameter optimisation. A broad HPO sweep was replaced by a manual grid search over four configurations per model, covering learning rate, dropout, batch size, and early stopping patience. Full HPO would have required an order of magnitude more GPU hours than available, and DeBERTa is known to be robust to hyperparameter choice within reasonable ranges. We acknowledge HPO as the most promising avenue for further gains.

Dataset challenges. The dataset presented two key challenges. First, a subset of examples lacked an ending sentence, reducing the contextual information available for disambiguation and increasing prediction difficulty. Second, annotator agreement varied considerably across samples, introducing label noise into the plausibility scores. These challenges motivated the use of a ranking-based loss, which is less sensitive to noisy labels and better aligned with the relative ordering of human judgments.

Future work. Several directions could extend this work. First, **hyperparameter optimisation** via Optuna or Ray Tune remains the most immediate gain, given our grid search covered only four configurations per model. Second, **data augmentation** — for instance, back-translation of low-resource homonym contexts or paraphrasing the pre-context while preserving the target sentence could alleviate the class imbalance between high- and low-agreement examples. Third, **larger backbones** such as DeBERTa-v3-large (400M parameters) or domain-adapted variants may close the remaining gap, particularly on cases where fine-grained world knowledge is required (e.g. *rare* beef vs. rarity). Finally, an **ensemble** of RoBERTa and DeBERTa regressors could combine their complementary strengths, as the two models exhibit different error profiles across homonym categories.

Feature	Baseline	RoBERTa(MSE)	RoBERTa(Rank Loss)	DeBERTa(Rank Loss)
Loss	MSE	MSE	MSE + Soft-ρ	MSE + Soft-ρ
Params	—	125M	125M	184M
Speed	Instant	Fast	Moderate	Slowest
Spearman ρ	0.08	0.334	0.450	**0.524**

Table 6: Comparison of all systems across key design dimensions.