Team Habib Disambiguators at SemEval-2026 Task 5: Assessing Semantic Plausibility using Regularized Transformer Fine-Tuning

Ahsan Siddiqui and Zohaib Aslam and Ayesha Enayet
Dhanani School of Science and Engineering
Habib University
Karachi, Pakistan
{as08155, za08134}@st.habib.edu.pk, ayesha.enayat@sse.habib.edu.pk

Abstract

This paper presents a system for SemEval-2026 Task 5: Rating Plausibility of Word Senses in Ambiguous Sentences through Narrative Understanding (Gehring et al., 2026; Gehring and Roth, 2025). The task involves predicting the plausibility of a specific word sense within a short story where context provided by the ending resolves a deliberate ambiguity. We model this as a regression problem, fine-tuning a DeBERTa-v3 transformer to predict the distribution of human judgments rather than a single hard label. To address the challenge of limited training data and potential overfitting, we employ R-Drop (Consistency Regularization) to enforce prediction stability across dropout masks and Layer-wise Learning Rate Decay (LLRD) to preserve the model's pre-trained linguistic knowledge. Our experiments demonstrate that treating plausibility as a soft-label distribution, combined with aggressive regularization, improves generalization on ambiguous samples. The submitted system achieves a Spearman correlation of 0.56 and an Accuracy (within SD) of 0.74 on the official test set.

1 Introduction

Lexical ambiguity resolution, which means determining which meaning of a word is intended in a specific context, remains a fundamental challenge in Natural Language Processing. SemEval-2026 Task 5, "Rating Plausibility of Word Senses in Ambiguous Sentences through Narrative Understanding", introduces a nuanced evaluation of this capability by focusing on short stories where a homonym's meaning is initially ambiguous but resolved by a twist ending (Gehring et al., 2026; Gehring and Roth, 2025). Unlike traditional Word Sense Disambiguation (WSD) tasks that treat meaning as a binary selection, this task requires systems to predict a graded plausibility score (1–5) that aligns with human judgments. This is crucial for developing NLP systems capable of understanding humor, puns, and narrative structure in English. We refer to the task organizers' overview paper (Gehring et al., 2026) for a detailed description of the dataset construction and evaluation metrics.

Our system approaches this challenge as a regression problem, leveraging the DeBERTa-v3 architecture to capture fine-grained contextual cues. A key element in our strategy is the use of R-Drop (Consistency Regularization), which forces the model to output consistent predictions for the same input across different dropout masks. Furthermore, we implement Layer-wise Learning Rate Decay (LLRD) to fine-tune the model's upper layers aggressively while preserving the linguistic knowledge encoded in the lower layers. By training on the distribution of human votes (soft labels) rather than a single average score, our model learns to account for the inherent subjectivity and variance in semantic judgment, similar to approaches investigating uncertainty in WSD (Liu and Liu, 2023).

Quantitatively, our system achieves a Spearman correlation of 0.56 and an Accuracy (within Standard Deviation) of 0.74 on the test set. These results suggest that consistency regularization is essential for robust performance on small, high-variance datasets.

2 Background

The core objective of SemEval-2026 Task 5 is to evaluate the plausibility of a specific word sense within a carefully structured narrative context (Gehring et al., 2026; Gehring and Roth, 2025). The input consists of a short English text broken into three parts: a "precontext" that sets up a scenario, a target "sentence" containing a specific homonym, and an "ending" that resolves the ambiguity. Alongside the story, a "judged_meaning" (a dictionary-style definition of one of the homonym's senses) is provided. The required output is a con-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2126–2129

tinuous plausibility score ranging from 1.0 (completely implausible) to 5.0 (highly plausible). For instance, in one training example, the homonym is "potential." The human annotations for this instance span the entire scale, resulting in an average score of 3.0 with a standard deviation of 1.58.

The provided dataset consists of English-language short stories designed to test context-dependent meaning resolution. The training set contains 2280 instances, while the development set includes 588 instances. A unique feature of this dataset is the rich annotation scheme: rather than a single ground-truth label, each instance includes the raw distribution of human votes, the average plausibility score, and the standard deviation. This captures the inherent subjectivity in semantic interpretation, a phenomenon previously explored in psycholinguistic plausibility pretesting with LLMs (Amouyal et al., 2023) and comparative studies between models and human development (Cabiddu et al., 2023).

3 System Overview

Our system approaches the semantic plausibility task as a continuous regression problem. Rather than treating the task as standard text classification, we model the probability distribution of human annotations. The core pipeline utilizes a prompt-based text concatenation strategy, a pre-trained DeBERTa-v3 encoder, and a highly regularized training objective featuring R-Drop and Layer-wise Learning Rate Decay (LLRD).

3.1 Prompt-Based Input Representation

Transformer models require text to be serialized into a single sequence. To allow the model to capture the interaction between the narrative context and the dictionary definition, we formulate the input as a natural language prompt, drawing inspiration from GlossBERT's knowledge-augmentation techniques (Huang et al., 2019). We concatenate the full story components (precontext, sentence, and ending) and append a direct question asking the model to evaluate the specific homonym and its proposed meaning, separated by a special token. For a given instance, the input string is constructed as follows:

$$Input = \text{Context [SEP] Prompt}$$

Where the Context is the raw text of the precontext, sentence, and ending joined together, and the Prompt is formulated as:
"In this context, how plausible is it that the meaning of the word '[homonym]' is '[judged_meaning]'?" This prompt-based formulation translates the abstract task of semantic entailment into an instruction-style query. The final hidden state of the [CLS] token from this sequence is passed through a linear regression head.

3.2 Soft Label Learning

A primary challenge of this task is the inherent subjectivity in semantic judgments. Training a model to predict only the average score ignores the variance in human votes.

Instead of predicting a single scalar, our system predicts the distribution of the five voting choices. We convert the raw vote counts into a probability distribution Y. The model outputs logits which are passed through a softmax function to produce a predicted distribution $\hat{Y}$. We optimize the model using Kullback-Leibler (KL) Divergence loss:

$$L_{task} = D_{KL}(Y||\hat{Y}) = \sum_{i=1}^{5} Y_i \log\left(\frac{Y_i}{\hat{Y}_i}\right)$$

During inference, the final scalar plausibility score is calculated as the expected value of the predicted distribution by taking the dot product of the softmax probabilities and a static score tensor [1.0, 2.0, 3.0, 4.0, 5.0]. This distributional approach is particularly effective for low-shot scenarios where rare word senses are present (Blevins and Zettlemoyer, 2020, 2021).

3.3 Consistency Regularization (R-Drop)

Fine-tuning large language models on small datasets often leads to overfitting. To address this, we integrated R-Drop. During each forward pass in training, the same input sequence x is fed through the network twice. We compute a bidirectional KL-Divergence to penalize inconsistencies between these two outputs, a technique that has shown promise in adapting BERT-based models for WSD objectives (Yap et al.).

3.4 Layer-wise Learning Rate Decay (LLRD)

To further prevent catastrophic forgetting of pre-trained linguistic knowledge, we applied LLRD using the AdamW optimizer. Instead of applying a uniform learning rate, we assigned a base learning rate of 2×10^{-5} to the top regression head. For

each of the 12 transformer layers moving downwards, the learning rate is exponentially decayed by a factor of $\gamma = 0.9$.

4 Experimental Setup

We implemented our system using the PyTorch framework and the HuggingFace Transformers library. The core model is `microsoft/deberta-v3-base`, chosen for its superior performance on NLU tasks compared to standard BERT.

4.1 Hyperparameters

We trained the model for 20 epochs with a batch size of 8. We utilized the AdamW optimizer with a base learning rate of 2×10^{-5} for the regression head and an LLRD decay factor of $\gamma = 0.9$. For the R-Drop component, the consistency weight α was set to 5.0.

4.2 Hardware and Training

All experiments were conducted on a single NVIDIA T4 GPU. To ensure reproducibility, we fixed the random seed for dataset shuffling and weight initialization. We monitored the Spearman correlation on the development set at the end of each epoch, saving only the best-performing checkpoint. Total training time was approximately 2 hours.

5 Results and Analysis

We evaluate our system using a combined performance metric, defined as the average of the Spearman correlation (ρ) and the Accuracy within Standard Deviation (Acc w/in SD). This composite score provides a holistic view of the model's ability to both rank plausibility correctly and fit the human annotation distribution.

5.1 Training Dynamics

The performance of the full model (DeBERTa-v3 + R-Drop + LLRD) over 20 epochs is summarized in Table 1.

The model demonstrates steady growth across both individual metrics. The stability of the combined metric after Epoch 12 suggests that the regularization techniques successfully mitigated the noise inherent in the AmbiStory dataset, preventing the catastrophic forgetting often seen in standard transformer fine-tuning.

Epoch	Spearman (ρ)	Acc (w/in SD)	Combined Metric
1	0.2430	0.5493	0.3962
3	0.3399	0.5867	0.4633
6	0.4690	0.7041	0.5866
9	0.5012	0.7177	0.6095
12	0.5056	0.7228	0.6142
14	0.5137	0.7143	0.6140
16	**0.5268**	**0.7245**	**0.6257**
20	0.5134	0.7075	0.6105

Table 1: Development set performance across key training epochs. The combined metric reaches its peak at Epoch 16.

5.2 Ablation Study and Synergy of Methods

To isolate the contributions of each component and understand how the peak combined score of 0.65 was achieved, we conducted an ablation study. The results are summarized in Table 2.

Model Configuration	Combined Metric	Δ
DeBERTa-v3-base (Pure Baseline)	0.6000	-0.05
+ R-Drop only	0.6100	-0.04
+ LLRD only	0.63	-0.02
Full Model (R-Drop + LLRD)	**0.65**	–

Table 2: Ablation results comparing regularization techniques using the combined metric.

Analysis of Synergy: The performance gains highlight a synergistic relationship between our structural and distributional regularizers:

- **LLRD** (Structural): By applying a decay to the learning rate across layers, we preserve the foundational linguistic representations of DeBERTa-v3 while allowing the top-level regression head to specialize in the story-sense task.
- **R-Drop** (Distributional): This technique enforces consistency between different dropout paths, acting as a powerful smoother for the output distribution. This is critical for the combined metric, as it improves both the ranking (Spearman) and the distributional fit (Acc w/in SD).

The combination of these methods allows the system to generalize effectively even with the limited training data and high annotator variance typical of SemEval tasks.

5.3 Discussion and Error Analysis

The significant jump from the baseline (0.60) to the full model (0.72) suggests that standard fine-tuning

is insufficient for the high-variance nature of the AmbiStory dataset. The performance gain from *LLRD only* (0.63) indicates that preserving lower-layer linguistic weights is crucial for understanding nuanced story contexts.

The additional gain provided by *R-Drop* reflects its role as a consistency regularizer; by forcing the model to produce similar distributions under different dropout masks, it effectively smooths the output manifold. This is particularly beneficial for predicting human vote distributions where the "correct" answer is often a spread rather than a single point.

6 Conclusion

In this paper, we presented a robust regression-based approach for rating word sense plausibility in narratives. By treating human annotations as a distribution rather than a single point estimate, and employing a combination of R-Drop and Layer-wise Learning Rate Decay, our system effectively navigates the challenges of limited training data and high label subjectivity. Our results demonstrate that enforcing consistency across dropout masks significantly aids in generalizing to the nuanced context of "twist endings" in ambiguous stories (Gehring et al., 2026; Gehring and Roth, 2025). Future work could explore the use of larger DeBERTa variants or the integration of external knowledge bases to better model the specific dictionary definitions provided in the task.

7 Limitations

Despite the competitive results, our system has several limitations. First, the small size of the training set makes the model sensitive to the specific phrasing of the prompt-based input. Second, our approach treats each instance independently; it does not explicitly model the logical relationship between the "sentence" and the "ending" beyond the transformer's self-attention mechanism. Finally, the LLRD decay factor and the R-Drop α weight were determined through heuristic search; a more exhaustive hyperparameter optimization could potentially yield better alignment with human variance, particularly in instances with high standard deviations.

References

Samuel Joseph Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2023. Large language models for psycholinguistic plausibility pretesting. In *Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

P. Blevins and L. Zettlemoyer. 2020. Low-shot word sense disambiguation with transformers and wordnet. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

T. Blevins and L. Zettlemoyer. 2021. Fews: A large-scale, low-shot word sense disambiguation dataset. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 346–356, Online. Association for Computational Linguistics.

F. Cabiddu, M. Nikolaus, and A. Fourtassi. 2023. Comparing children and large language models in word sense disambiguation: Insights and challenges. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 35–50.

Janosch Gehring, Selina Meyer, and Michael Roth. 2026. SemEval-2026 task 5: Rating plausibility of word senses in ambiguous stories through narrative understanding. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Janosch Gehring and Michael Roth. 2025. Ambistory: A challenging dataset of lexically ambiguous short stories. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 152–171.

L. Huang, C. Sun, X. Qiu, and X. Huang. 2019. Glossbert: Bert for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514.

Zhu Liu and Ying Liu. 2023. Ambiguity meets uncertainty: Investigating uncertainty estimation for word sense disambiguation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 245–253. Association for Computational Linguistics.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. Adapting bert for word sense disambiguation with gloss selection objective and example sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 41–46.