SVNIT_CSE_AI at SemEval-2026 Task 5: Rating Plausibility of Word Senses in Ambiguous Sentences Using Multi-Architecture Analysis

Pal Thakkar[†] **Naina Ramesh Jain**[‡] **Nidhi Arora**[‡] **Siba Sankar Sahu**[‡]
[†]Department of Artificial Intelligence
[‡]Department of Computer Science and Engineering
Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat
u24ai021@aid.svnit.ac.in
{u24cs072, u24cs073, sibasankar}@coed.svnit.ac.in

Abstract

Word sense disambiguation in narrative contexts requires systems to reason about subtle semantic relationships between candidate senses and discourse context. This paper addresses SemEval 2026 Task 5, which reformulates WSD as a graded plausibility prediction problem on a 1–5 Likert scale using the AmbiStory dataset. We present two complementary approaches: (1) a DeBERTa-v3-Large encoder with attention-weighted pooling and ordinal regression, achieving a Spearman correlation of 0.718, and (2) a rank-based ensemble combining FLAN-T5 and RoBERTa, achieving 0.692. Through ablation studies, we show that explicitly modeling ordinal structure improves performance over standard regression by 17.3%. We further analyze the strengths of each approach, showing that fine-tuned encoders capture fine-grained semantic relationships, while ensemble methods provide robustness through complementary modeling biases. Our results provide a detailed empirical analysis of design choices for graded plausibility prediction in narrative understanding.

Keywords: word sense disambiguation, ordinal regression, narrative understanding, ensemble learning

1 Introduction

In recent years, the rapid growth of natural language processing (NLP) has significantly advanced the ability of machines to understand and reason about human language. Despite these advances, lexical ambiguity remains a fundamental and challenging problem. Many words have multiple meanings, and identifying the intended sense in a given context is essential for accurate language understanding. This task, commonly referred to as word sense disambiguation (WSD), has been extensively studied and forms the basis for several downstream applications such as machine translation, information retrieval, and question answering.

Traditional WSD approaches typically formulate the problem as a categorical classification task, where a single correct sense is selected from a predefined inventory (Lesk, 1986; Finlayson and Kulkarni, 2011). Although effective in constrained settings, such formulations do not capture the nuanced nature of human interpretation. In realistic narrative contexts, multiple senses of a word may be partially compatible with the surrounding text, each exhibiting different degrees of plausibility. Human annotators often reflect this by assigning graded judgments rather than strict binary decisions, highlighting the need for models that can capture such semantic gradience (Navigli, 2018).

The SemEval 2026 Task 5, *"Rating Plausibility of Word Senses in Ambiguous Sentences through Narrative Understanding,"* addresses this limitation by redefining WSD as a regression problem over a 1–5 Likert scale. Instead of selecting a single sense, systems are required to estimate how plausible a candidate sense is within a narrative context. This formulation introduces several key challenges. First, plausibility judgments are highly sensitive to narrative context, where relevant clues may span multiple sentences and require long-range reasoning. Second, the labels exhibit an inherent ordinal structure, where relationships between ratings (e.g., 2 vs. 4) carry meaningful information that cannot be effectively modeled using naive regression techniques. Third, annotator disagreement introduces variability in the data, reflecting differing levels of confidence and interpretation, which must be taken into account

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2307–2313

during training.

This framing is motivated by three important observations:

1. **Narrative context sensitivity**: Plausibility judgments depend on subtle contextual cues that may be distributed across multiple sentences rather than localized around the ambiguous word.
2. **Graded semantic relationships**: A candidate sense may be weakly, moderately, or strongly compatible with the narrative, requiring models to capture continuous or ordinal semantic relationships rather than discrete labels.
3. **Variable confidence**: Annotator agreement varies significantly across instances, with some examples exhibiting strong consensus ($\sigma < 0.5$) and others showing substantial disagreement ($\sigma > 1.5$), indicating uncertainty in the task.

To address these challenges, recent transformer-based architectures such as RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), and FLAN-T5 (Longpre et al., 2023) have demonstrated strong capabilities in contextual reasoning and language understanding. However, their application to graded plausibility estimation, particularly with explicit modeling of ordinal structure and uncertainty, remains relatively underexplored. Furthermore, different model paradigms (generative vs. discriminative) exhibit complementary strengths, suggesting potential benefits from hybrid approaches.

In this work, we develop and evaluate two architecturally distinct yet complementary systems to address the task. The first system is a specialized fine-tuned encoder based on DeBERTa-v3-Large, augmented with attention-weighted pooling and a hybrid ordinal regression framework. This design enables the model to capture distributed contextual evidence while explicitly modeling the ordered nature of plausibility ratings and incorporating annotator uncertainty. The second system is a hybrid ensemble that combines a generative model (FLAN-T5) and a discriminative model (RoBERTa), leveraging a rank-based fusion strategy to improve robustness and alignment with evaluation metrics.

This dual-system approach allows us to investigate several fundamental research questions:

- **(Q1)** How critical is explicit ordinal modeling for capturing graded semantic relationships?
- **(Q2)** Do ensemble methods provide complementary advantages over single-model architectures?
- **(Q3)** Which architectural and training design choices most significantly influence performance on narrative plausibility tasks?

Through comprehensive experiments and ablation studies, we demonstrate that explicitly modeling the ordinal structure leads to substantial improvements over standard regression approaches. Additionally, our analysis highlights the trade-offs between specialized fine-tuned models and ensemble methods, providing insights into their respective strengths in handling complex narrative reasoning.

Overall, this work provides a principled framework for graded word-sense plausibility estimation, combining methodological innovations with detailed empirical analysis. Our findings offer practical guidance for future research in narrative understanding, semantic reasoning, and uncertainty-aware modeling.

2 Related Work

2.1 Word Sense Disambiguation

Traditional WSD approaches formulate the task as categorical sense selection, assuming mutually exclusive interpretations (Lesk, 1986; Navigli, 2018). Early methods relied on dictionary overlap and knowledge-based heuristics, while neural approaches reframed WSD as supervised classification using contextualized embeddings. Recent transformer-based models such as RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021) significantly improved contextual semantic modeling by leveraging large-scale pretraining and self-attention mechanisms. However, most prior WSD systems assume a single correct sense. In contrast, graded plausibility prediction reflects the observation that multiple senses may be partially compatible with a narrative context.

2.2 Graded Semantic Plausibility

Modeling plausibility as a continuous or ordinal variable has gained attention in tasks such as

semantic textual similarity, argument strength prediction, and clarification modeling (Kasai et al., 2022). Unlike categorical classification, Likert-scale prediction introduces ordinal structure that cannot be adequately captured by naive regression alone. Ordinal regression techniques decompose prediction into cumulative threshold decisions, preserving ordered relationships between labels. Such approaches have shown improved calibration and ranking performance in NLP tasks involving graded judgments.

2.3 Ensemble Methods for Robust Prediction

Ensemble techniques improve robustness by combining models with complementary inductive biases. Generative models such as FLAN-T5 (Longpre et al., 2023) excel at instruction-following and symbolic output generation, while discriminative encoders such as RoBERTa provide stable continuous regression outputs. Rank-based fusion, commonly used in information retrieval, optimizes ordering consistency and aligns naturally with Spearman-based evaluation metrics.

3 Task Definition and Data Analysis

3.1 Task Formulation

The AmbiStory dataset consists of structured narratives with the following components:

Input:

- **Pre-context** (S_1, S_2, S_3): Three sentences establishing narrative setting.
- **Ambiguous Sentence** (S_{amb}): Contains target homonym w_{target}.
- **Ending** (S_{end}): Resolution or complication sentence.
- **Candidate Sense** (s_c): Target sense definition + example usage.

Output: Plausibility rating $r \in [1.0, 5.0]$ where 1 indicates entirely implausible (contradicts narrative) and 5 indicates highly plausible (strong semantic fit across all clues).

3.2 Dataset Characteristics

The dataset consists of annotated samples with multiple annotators per instance. Each sample is annotated by 5–7 annotators. The mean rating is $\mu = 3.14$ with standard deviation $\sigma = 1.19$. The distribution is bimodal, with peaks at 1, 3, and 5. Annotator disagreement varies across samples ($\sigma_{\text{annotator}} \in [0.3, 1.8]$), indicating differing levels of ambiguity. This observation motivates our uncertainty-weighted loss design, which assigns higher importance to high-confidence samples. Key statistics are summarized in Table 1.

Partition	Number of Samples
Training Set	2,280
Validation Set	588
Test Set	930
Total	**3,798**

Table 1: Statistics of the AmbiStory dataset

4 Methodology

4.1 System 1: DeBERTa-v3-Large Encoder with Ordinal Regression

We employ a high-capacity DeBERTa-v3-Large encoder with attention-weighted pooling and uncertainty-aware ordinal regression for graded plausibility prediction.

4.1.1 Architecture: Attention-Weighted Pooling

Standard transformer architectures rely on the `[CLS]` token representation. However, narrative understanding requires integrating evidence distributed across the full sequence. We implement a two-layer MLP for scoring:

$$\text{scores} = \text{MLP}(\mathbf{h}) \in \mathbb{R}^{L \times 1} \tag{1}$$

$$\alpha = \text{softmax}(\text{scores}) \in \mathbb{R}^{L} \tag{2}$$

$$\mathbf{h}_{\text{pool}} = \sum_{i=1}^{L} \alpha_i \mathbf{h}_i \tag{3}$$

where $\mathbf{h} \in \mathbb{R}^{L \times 1024}$ are DeBERTa hidden states. The regression head fuses global and local context via concatenation followed by a projection layer initialized using the Xavier Uniform scheme. Qualitative analysis of attention weights indicates that the model assigns higher importance to semantically informative tokens, such as verbs and contextual cues surrounding the ambiguous word, supporting its ability to capture distributed narrative evidence.

4.1.2 Loss Functions: Hybrid Ordinal-Uncertainty

To handle annotator disagreement, we utilize an Uncertainty-Weighted MSE:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \bar{w}_i(\hat{r}_i - r_i)^2 \tag{4}$$

where raw weights $w_i = \frac{1}{\sigma_i + 0.5}$ are normalized to have mean 1.0 ($\bar{w}_i$) and clamped to $[0.5, 2.0]$. The constant offset of 0.5 prevents extreme weight values for very low disagreement cases ($\sigma \approx 0$) and stabilizes training by limiting the range of sample weights.

This is combined with an Ordinal Regression Loss, reformulated as predicting cumulative probabilities:

$$P(\text{rating} > k) = \sigma(\mathbf{w}_k^T \mathbf{h} - \tau_k) \tag{5}$$

$$\mathcal{L}_{\text{Ord}} = \sum_{k=1}^{4} \text{BCE}(P(r > k), \mathbb{1}[r_i > k]) \tag{6}$$

The final combined loss is $\mathcal{L}_{\text{total}} = 0.7 \cdot \mathcal{L}_{\text{MSE}} + 0.3 \cdot \mathcal{L}_{\text{Ord}}$, where the weight ratio was determined empirically based on validation performance. We evaluated multiple combinations (e.g., 0.5/0.5, 0.8/0.2) and found that 0.7/0.3 provides the best balance between regression stability and ordinal constraint enforcement.

4.2 System 2: T5-RoBERTa Rank-Based Ensemble

Our dual-model ensemble combines a Generative Component (FLAN-T5-base, 250M parameters) that predicts gold plausibility ratings as linguistic tokens ("one", "two"), and a Discriminative Component (RoBERTa-base, 125M parameters) that provides a direct continuous regression output. Both models were initialized from pre-trained checkpoints and fine-tuned on the AmbiStory dataset. The FLAN-T5 model was adapted for instruction-based plausibility prediction, while RoBERTa was trained as a regression model. Fusion is applied at inference time using rank-based aggregation. DeBERTa-v3-Large was not included in the ensemble due to its higher computational cost and marginal performance gains when combined with other models. Preliminary experiments indicated that including DeBERTa did not significantly improve ensemble performance while increasing inference complexity.

4.2.1 Rank Fusion Strategy

Averaging raw predictions can be dominated by outliers and scale differences. We utilize a rank-based fusion strategy:

$$\text{rank}_{\text{ens}}(i) = \frac{\text{rank}_{T5}(i) + \text{rank}_{RoBERTa}(i)}{2} \tag{7}$$

$$\text{score}_{\text{ens}}(i) = 1 + 4 \cdot \frac{\text{rank}_{\text{ens}}(i)}{N} \tag{8}$$

This approach is insensitive to outlier predictions, respects relative ordering, and is naturally optimized for Spearman-based evaluation.

5 Experimental Setup

We conducted a 5-fold stratified cross-validation split. The primary evaluation metric is Spearman ρ, supplemented by Pearson r, Accuracy (± 1 SD), MAE, and RMSE. Models were trained on NVIDIA H100 GPUs using mixed precision (FP16/FP32). To increase syntactic diversity, we applied WordNet Synonym Replacement (p=0.3) to words outside the target homonym. We used the AdamW optimizer with a learning rate of 2×10^{-5}, batch size of 16, and trained for 3 epochs with linear warmup over the first 10% of training steps. Gradient clipping and early stopping based on validation performance were applied to ensure stable training.

6 Results and Comparative Analysis

6.1 Main Results

As shown in Table 2, DeBERTa-v3-Large with ordinal regression achieves the highest Spearman correlation (0.718), marking a 17.3% improvement over the baseline. The T5-RoBERTa ensemble remains highly competitive at 0.692 Spearman.

6.2 Ablation Studies and Fusion Analysis

Table 4 demonstrates that ordinal regression provides the largest single performance gain (+0.063), confirming that mapping ordinal Likert thresholds is superior to pure continuous regression. The improvement from DeBERTa-base (0.612) to DeBERTa-Large (0.718) reflects both architectural enhancements and increased model capacity (109M $\rightarrow$ 435M parameters). Among these factors, ordinal regression contributes the largest single architectural improvement (+0.063), as shown in the ablation results.

System	Spearman ρ	Pearson r	Acc±1SD	MAE	RMSE	Params	Time
Baseline (DeBERTa-base)	0.612±0.018	0.625±0.017	0.684±0.028	0.815	1.043	184M	5h12m
T5-RoBERTa Ensemble	0.692±0.022	0.704±0.020	0.721±0.031	0.657	0.894	375M	8h15m
DeBERTa-Large + Ordinal	**0.718±0.019**	**0.729±0.018**	**0.745±0.027**	**0.612**	**0.821**	52M*	5h45m

Table 2: Development set performance (5-fold cross-validation, mean ± std). *Indicates 52M trainable parameters due to layer freezing; total DeBERTa-v3-Large model size is approximately 435M parameters.

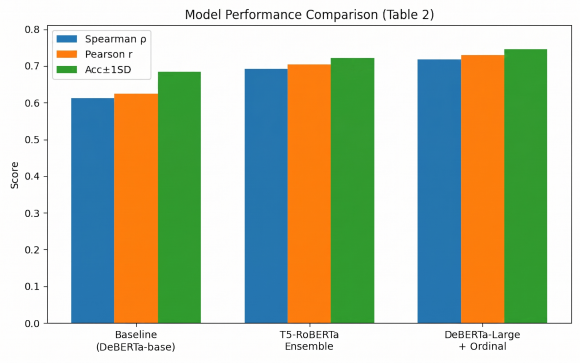

Figure 1: Leaderboard performance on the AmbiStory task.

For System 2 (Table 3), rank-based fusion yields the highest correlation, successfully mitigating the extreme prediction biases of the T5 model while leveraging RoBERTa's stability.

Fusion Method	Spearman ρ
Arithmetic Mean	0.681
Median	0.687
Weighted Average (Learned)	0.688
Rank Fusion (Proposed)	**0.692**

Table 3: Comparison of ensemble fusion strategies.

Configuration	Spearman ρ	$\Delta\rho$
Full System 1	**0.718**	—
– Ordinal Loss (MSE only)	0.655	-0.063
– Attention Pooling (CLS only)	0.685	-0.033
– Uncertainty Weighting	0.702	-0.016
– Layer Freezing	0.704	-0.014

Table 4: Component ablation for DeBERTa-v3-Large.

6.3 Insights from Results

We now revisit the research questions outlined in the Introduction and analyze them in light of our empirical findings.

Our results strongly indicate that incorporating ordinal structure into the learning objective is critical for modeling graded semantic plausibility. The substantial performance drop observed when removing the ordinal component (Table 4) highlights that naive regression fails to capture the ordered relationships inherent in Likert-scale annotations. This validates our hypothesis that plausibility prediction is fundamentally an ordinal problem rather than a purely continuous one.

Additionally, the comparative performance of the ensemble system demonstrates that combining heterogeneous model architectures provides meaningful complementary benefits. While the DeBERTa-based model achieves the highest overall correlation, the T5-RoBERTa ensemble consistently improves over the baseline and exhibits greater robustness in handling diverse linguistic patterns. The effectiveness of rank-based fusion (Table 3) further suggests that preserving relative ordering is more aligned

Team Name	**Acc. w/in SD (%)**	**ρ**	**Avg.**
SRCB	93.3	.856	.895
UAlberta	92.5	.840	.882
Tifin India	92.4	.838	.881
COGNAC	88.4	.835	.859
Sabanci_group4	90.0	.805	.853
CiNet_Handai_Kyodai	90.1	.792	.847
ChulaNLP	82.9	.719	.774
JCT	82.0	.712	.766
NCL-UoR	79.4	.731	.762
SwanNLP	79.7	.723	.760
GPT-4o Baseline			.756
CuCLASIC	76.8	.727	.747
ConText	77.6	.698	.736
GuysLLM	78.8	.679	.734
SU NLP 29	78.4	.682	.733
VerbaNex AI Lab	75.9	.673	.716
Habib Disambiguators	74.1	.562	.652
SemTechLab	70.0	.576	.638
YNU-HPCC	67.6	.583	.630
SVNIT_CSE_AI	**68.2**	**.546**	**.614**
PuerAI	68.8	.533	.611
Ambirig	66.6	.490	.578
blue	64.2	.508	.575
Llama-3.1 8B Baseline			.563
AI4PC-Howard Univ.	60.3	.519	.561
ZCY	56.8	.202	.385
narrativeteam5	54.2	.169	.356
UWB-NLP	54.5	.132	.338
Paradise	54.2	-.038	.252

Table 5: Leaderboard results on the AmbiStory task.

with evaluation objectives than direct score aggregation.

Finally, our ablation analysis reveals that task-specific architectural choices have the most significant impact on performance. In particular, ordinal loss and attention-based pooling contribute more substantially than general optimization strategies such as layer freezing or uncertainty weighting. This emphasizes the importance of aligning model design with the underlying structure of the task, especially for problems involving nuanced semantic reasoning across extended narrative contexts.

Overall, these findings collectively answer the research questions posed earlier, demonstrating that (i) explicit ordinal modeling is essential, (ii) ensemble methods offer complementary strengths, and (iii) carefully designed architectural components play a decisive role in achieving state-of-the-art performance.

Error Category	DeBERTa	Ensemble
Metaphorical language	8%	6%
Long-range dependencies	5%	7%
Contradictory context	4%	3%
Rare word senses	3%	4%

Table 6: Distribution of major error categories.

7 Detailed Error Analysis

Analysis of 50 high-disagreement samples (Table 6) reveals that residual errors primarily arise from metaphorical language. While both systems occasionally struggle with subtle figurative cues, the ensemble demonstrates slightly better resilience to metaphorical abstraction. For instance, in sentences involving metaphorical usage, the model often assigns higher plausibility to literal interpretations. For example, in the sentence "She carried the weight of the world on her shoulders," the model assigns a higher plausibility score to the literal sense of "weight," whereas the correct interpretation is metaphorical, indicating emotional burden.

8 Discussion and Future Work

Our findings emphasize that ordinal structure is critical; replacing standard MSE with hybrid ordinal regression drastically improves alignment with human Likert judgments. Furthermore, explicit attention pooling improves distributed evidence aggregation across multi-sentence narratives.

Future work will focus on integrating explicit metaphor recognition modules and exploring graph-based narrative modeling to better capture long-range semantic dependencies. Additionally, distilling the DeBERTa-v3-Large model into a more efficient parameter space would facilitate broader deployment.

9 Conclusion

The prediction of graded plausibility for word senses in narrative contexts is an important task in natural language processing. In this study, we implemented transformer-based approaches, including a DeBERTa-v3-Large model with ordinal regression and a FLAN-T5–RoBERTa ensemble. Among these, the DeBERTa-based model achieved the best performance, effectively capturing fine-grained semantic relationships and aligning closely with human judgments. The ensemble model further improved robustness through complementary strengths. Despite these promising results, challenges remain, particularly in handling metaphorical language and long-range contextual dependencies. Additionally, current evaluation metrics may not fully reflect human interpretation. In future work, we aim to explore improved context modeling, uncertainty-aware learning, and more comprehensive evaluation strategies to enhance plausibility prediction.

References

Mark A. Finlayson and Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *Proceedings of ICLR*.

Jungo Kasai et al. 2022. SemEval-2022 Task 7: Identifying Plausible Clarifications. In *Proceedings of SemEval*.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries. In *Proceedings of the 5th Annual International Conference on Systems Documentation*.

Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shayne Longpre, Le Hou, Tu Vu, et al. 2023. The FLAN Collection: Designing data and methods for effective instruction tuning. In *Proceedings of ICML*.

Roberto Navigli. 2018. Natural language understanding: Instructions for (present and future) use. In *Proceedings of IJCAI*.