Rating Plausibility of Word Senses in Ambiguous Narratives

SemEval 2026 – Task 5

1. Introduction

Natural language is inherently ambiguous, particularly in the case of homonyms, where a single lexical form can express multiple distinct meanings. Traditional Word Sense Disambiguation (WSD) tasks typically assume that one word sense is objectively correct in a given context. However, this assumption does not fully reflect human language understanding, as ambiguity, underspecification and subjective interpretation often allow multiple meanings to remain plausible.

SemEval 2026 Task 5 addresses this limitation by shifting the focus from categorical sense selection to graded plausibility estimation. The task requires systems to predict how plausible a given word sense appears to human readers within a short narrative context. In this work, we present our team's approach to this task and analyze the experimental results obtained on the provided dataset.

2. Problem Description

SemEval 2026 Task 5, Rating Plausibility of Word Senses in Ambiguous Sentences through Narrative Understanding, focuses on modeling human judgments of word sense plausibility in short stories.

Each instance consists of a five-sentence narrative:

1. The first three sentences establish the context;
2. The fourth sentence contains an ambiguous homonym with two possible senses;
3. The fifth sentence optionally provides an ending that may bias interpretation toward a specific sense.

Given the narrative and one candidate sense of the ambiguous word, the system must predict a plausibility score on a scale from 1 to 5, reflecting the degree to which human annotators consider the sense reasonable in context.

Student Tasks

Dumitru Daniel-Antoniu	Documentation, Architecture, Evaluation,Testing & Validation
Lazăr Simina	Documentation, Architecture, Implementation, Evaluation, Testing & Validation
Danilă Nicoleta (Amargheoalei)	Documentation, Architecture, Testing & Validation

3. Dataset Description

The AmbiStory dataset was designed to elicit graded plausibility judgments in narrative contexts. Each story is annotated with plausibility scores for two candidate senses, across six variants (with/without endings). Human ratings were collected via Prolific using a 5-point Likert scale.

Each instance includes:

- Full story text;
- Target word (homonym);
- Candidate sense gloss;
- Mean plausibility score from multiple annotators;
- Standard deviation of ratings (used for tolerance-based accuracy).

4. Exploratory Data Analysis

Our analysis of the training set reveals:

- Score distribution: Plausibility scores span the full 1–5 range;
- Central tendency: Most frequent score is 3, indicating balanced ambiguity;
- Annotator agreement: Standard deviation ranges from 0.6 to 1.8, reflecting varying degrees of subjectivity;
- Bias effects: Endings tend to reduce ambiguity and increase annotator consensus.

These findings confirm the need for models that can handle both clear and ambiguous cases, and motivate the use of ordinal and rank-based objectives.

5. State of the Art

Traditional approaches to word sense disambiguation frame the task as a classification problem with a single correct sense. Earlier methods relied on feature-based models such as Support Vector Machines, while more recent work has adopted transformer-based language models, including BERT and RoBERTa.

Although these models achieve strong performance on standard WSD benchmarks, they generally do not account for graded plausibility or subjective interpretation. Task 5 differs from prior work by explicitly modeling uncertainty and narrative-driven ambiguity.

Related Work

Recent literature supports a shift toward graded semantic evaluation:

- AmbiStory (Gehring & Roth, 2025): Introduces narrative-based ambiguity and motivates plausibility modeling [1];
- Wang et al. (2021): Demonstrate that combining local and global context improves disambiguation [2];

- CoMeDi (Schlechtweg et al., 2025): Validate ordinal modeling and rank- based metrics for word-in-context tasks [3];
- WiC (Pilehvar & Camacho-Collados, 2019): Highlight the importance of contextualized embeddings for sense comparison [4];
- Word Sense Extension (Yu & Xu, 2023): Argue for flexible, context-adapted sense representations [5];

These works inform our architectural choices: dual context encoding, contextualized sense embeddings, fusion layers and ordinal regression.

6. Proposed Approach

We model the task as an ordinal regression problem, predicting a scalar plausibility score on a 1–5 scale, while explicitly optimizing for relative ordering between competing word senses.

Our system is based on a pretrained transformer-based language model, which encodes the full narrative context together with the candidate word sense. The model is fine-tuned on the AmbiStory training data to learn the relationship between narrative cues and human plausibility judgments.

To better align with human judgments, we combine:

- Regression loss (MSE): Approximates the mean human score;
- Pairwise ranking loss: Enforces correct ordering between competing senses of the same story.

Input formatting:

- Target word marked with `<TGT>` tokens;
- Sense gloss appended after the story using `[SEP]`;
- Maximum sequence length: 512 tokens.

7. System Architecture

The architecture of our system consists of three main components:

1. A tokenizer that converts text into subword units;
2. A pretrained transformer encoder that produces contextualized representations;
3. A regression head that outputs a single plausibility score.

We evaluate four model variants:

- BERT-base;
- BERT-STDM: BERT-base + 1 dense layer (256 units, ReLU);
- BERT-STDM-2L: BERT-base + 2 stacked dense layers (256 → 128 units, ReLU);
- RoBERTa-base;

- MiniLM-L6-v2 (Sentence Transformers).

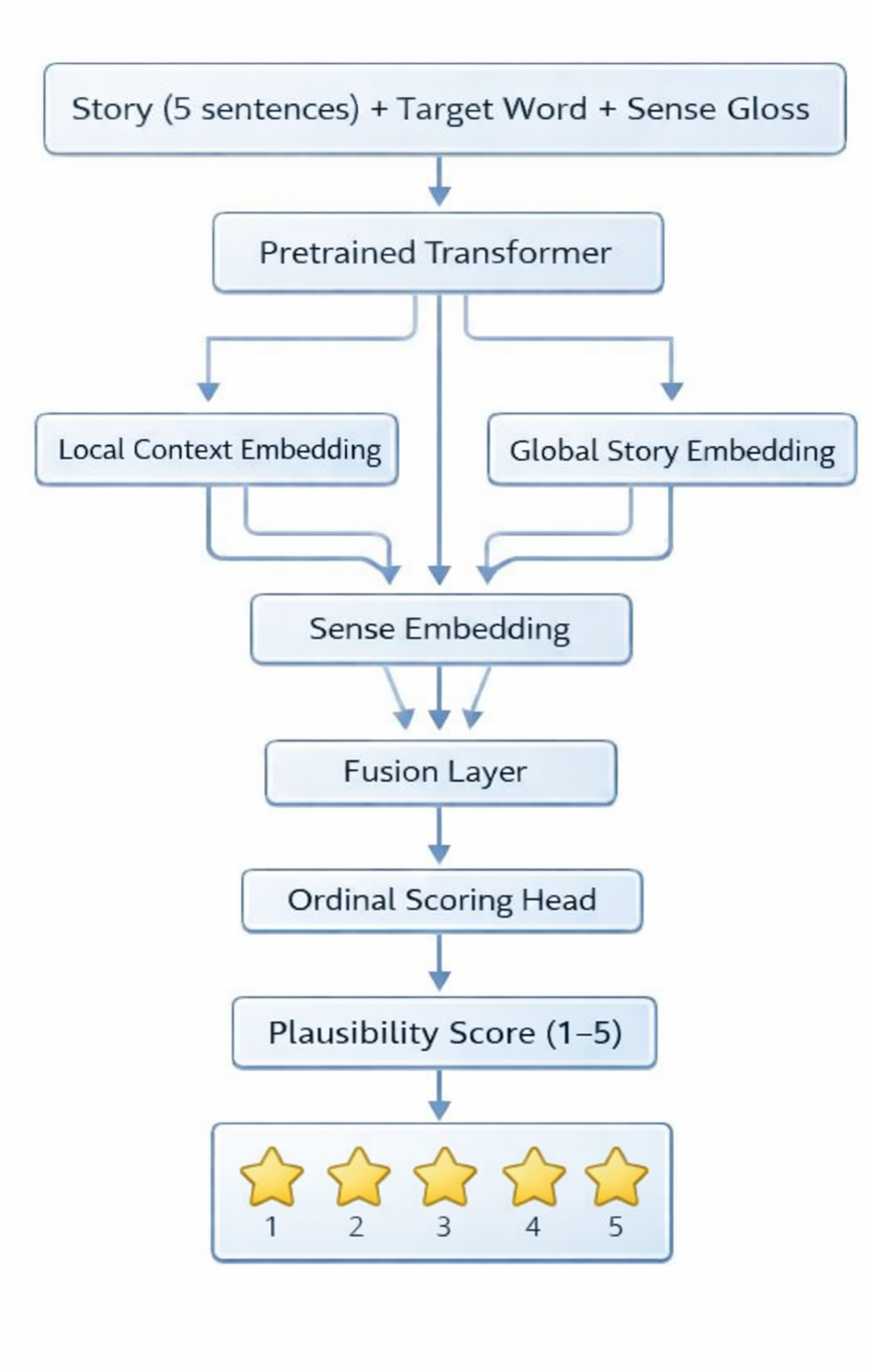

Figure 1: System architecture

Experimental Setup

- Epochs: 10 (5 for sentenceBERT and 20 for sentenceBERT with STDM);
- Batch size: 8 (16 for sentenceBERT with STDM, 32 for sentenceBERT);
- Learning rate: 2e-5 (3e-5 for RoBERTa, 1e-5 for sentenceBERT with STDM);
- Optimizer: AdamW;

- Scheduler: Linear warmup (10%);
- Loss: MSE + margin ranking (margin = 0.5);
- Hardware: NVIDIA T4 GPU (Google Colab).

8. Evaluation Metrics

System performance is evaluated using the official metrics defined by the task organizers:

- Spearman Correlation: measures rank-based agreement with human judgments;
- Accuracy Within Standard Deviation: measures whether predictions fall within the variability of human ratings.

9. Experimental Results

Model	Spearman ρ	Accuracy (±1 STD)
BERT-base	0.3352	0.6344
BERT-STDM	0.3668	0.6616
BERT-STDM-2L	0.3355	0.6497
RoBERTa-base	0.1193	0.5782
MiniLM-L6-v2	0.4478	0.5918

Table 1: Validation results

The validation set given by the competition was used for testing and the training set given by the competition was used for training.

Model	Spearman ρ	Accuracy (±1 STD)
MiniLM-L6-v2-STDM	0.1424	0.5118

Table 2: Test results

The test set given by the competition was used for testing, the training set given by the competition was used for training and the validation set given by the competition was used for validation.

Best model: BERT-STDM

It achieves the highest correlation and best tolerance-based accuracy, confirming the benefit of shallow task-specific adaptation.

10. Discussion

Our results indicate that adding a shallow task-specific dense module (STDM) improves performance over plain BERT-base. This suggests that task-specific adaptation helps the model better align narrative cues with sense plausibility.

Surprisingly, stacking two layers (STDM-2L) does not yield further gains, possibly due to overfitting or limited training data. RoBERTa-base underperforms, highlighting that pretrained representations alone are insufficient without targeted adaptation.

Error analysis shows that:

- Balanced ambiguity leads to near-random ordering between senses;
- High annotator disagreement correlates with lower model accuracy;
- Models struggle when narrative cues are subtle or pragmatic rather than lexical.

Interestingly, the all-MiniLM-L6-v2 model achieves the highest Spearman correlation (0.4478), outperforming all BERT-based variants in terms of rank agreement with human judgments. This suggests that MiniLM's sentence-level semantic representations capture relative plausibility differences effectively, despite its smaller size and lack of task-specific adaptation. However, its Accuracy@STD remains lower than BERT-STDM, indicating that MiniLM is less reliable in predicting absolute plausibility values within the annotator variability range.

11. Conclusion and Future Work

We presented a transformer-based approach to SemEval 2026 Task 5, modeling graded plausibility of word senses in narrative contexts. Our best-performing model, BERT-STDM, demonstrates that shallow task-specific adaptation improves alignment with human judgments.

The strong performance of MiniLM-L6-v2 highlights the potential of lightweight sentence-transformer architectures for graded semantic tasks, even without explicit task-specific layers.

Future directions include:

- Modeling annotator disagreement explicitly;
- Incorporating narrative structure (e.g., discourse relations, causal links);
- Exploring contrastive learning between competing senses;
- Multi-task training with related datasets (WiC, CoMeDi).

References

[1] J. Gehring and M. Roth. Ambistory: A challenging dataset of lexically ambiguous short stories. In *Proceedings of SEM 2025*. Association for Computational Linguistics, 2025.

[2] Wang M. et. al. Enhancing the context representation in similarity-based word sense disambiguation. In *Proceedings of EMNLP 2021*. Association for Computational Linguistics, 2021.

[3] Schlechtweg D. et. al. Comedi shared task: Median judgment classification and mean disagreement ranking with ordinal word-in-context judgments. In *Proceedings of CoMeDi 2025*. Association for Computational Linguistics, 2025.

[4] Pilehvar M. T. and Camacho-Collados J. Wic: The word-in-context dataset. In *Proceedings of NAACL-HLT 2019*. Association for Computational Linguistics, 2019.

[5] L. Yu and Xu Y. Word sense extension. In *Proceedings of ACL 2023*. Association for Computational Linguistics, 2023.