blue at SemEval-2026 Task 5: NarrBERT : Narrative-Aware BERT for Word Sense Disambiguation

Rhea Singhal
Thapar Institute of Eng. & Tech.
Patiala, Punjab, India
rsinghal_be23@thapar.edu

Krish Sharma
Thapar Institute of Eng. & Tech.
Patiala, Punjab, India
ksharma8_be23@thapar.edu

Lakksh Sharma
Thapar Institute of Eng. & Tech.
Patiala, Punjab, India
lsharma_be23@thapar.edu

Jatin Bedi
Thapar Institute of Eng. & Tech.
Patiala, Punjab, India
jatin.bedi@thapar.edu

Abstract

This paper outlines the method submitted by team blue for the SemEval-2026 Task 5: Rating Plausibility of Word Senses in Ambiguous Sentences through Narrative (AmbiStory). The task requires predicting reasonable scores that match human thoughts and judgments instead of just picking a single correct sense as the output. This means that contextual reasoning with fine-grain contextual modeling is vital. In order to tackle this problem, we suggest a BERT-based cross-encoder regression model. This model encodes the entire narrative context, which includes the precontext, the ambiguous sentence, and the ending, along with candidate sense definitions and example usages. Unlike bi-encoder sentence-level methods, our model allows for token-level interaction between story cues and sense meanings. This interaction helps capture subtle narrative disambiguation signals. We conduct a systematic exploration of model architectures and training strategies, progressing from a sentence-transformer baseline to an optimised BERT cross-encoder. On the development set, our best configuration achieves a Spearman rank correlation of 0.66. On the official test set, the system achieves a Spearman correlation of **0.4866** and an Accuracy-within-Standard-Deviation of **0.6613**, substantially outperforming sentence-transformer bi-encoder baselines.

1 Introduction

Word Sense Disambiguation (WSD) in narrative contexts goes beyond traditional classification methods. Human interpretation of ambiguity is often more complicated than simply fitting it into categories.SemEval-2026 Task 5: Rating Plausibility of Word Senses in Ambiguous Stories through Narrative Understanding (AmbiStory) (Gehring et al., 2026) captures this idea. It requires systems to predict continuous plausibility scores that reflect human judgments. The cues for disambiguation in this task are often spread throughout the entire narrative, including the story's ending, making it inadequate to focus only on sentence-level modeling.

Figure 1 shows the overall task layout. Given a five-sentence narrative and a candidate sense definition, the system must provide a plausibility score ranging from 1 to 5 to match human annotations.

This task presents several challenges. Plausibility judgments can vary among annotators. The evaluation metrics focus on matching human disagreement. Furthermore, the narrative context needs to be understood as a whole. Initial experiments with sentence-transformer bi-encoders showed limited ability to model the detailed interactions between stories and sense descriptions, especially when disambiguation relies on subtle developments in the narrative. [1]

To solve these challenges, we suggest a narrative-aware BERT cross-encoder trained with a regression goal. Our model encodes the entire story context along with the candidate sense definition, allowing detailed token-level interaction between narrative cues and semantic gloss information. Unlike bi-encoder methods that process each input separately, our design directly models the relationship between context and meaning. This approach helps

[1] https://github.com/RheaSinghal/SemEval-Task5

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2427–2431

capture subtle resolution signals introduced at the end of the story.

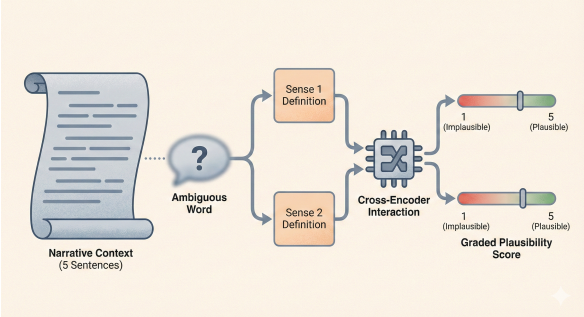

Figure 1: Overview of the AmbiStory plausibility task. The model uses the full 5-sentence narrative to project a continuous score onto a candidate sense definition.

2 Background and Related Work

Traditional Word Sense Disambiguation (WSD) treats sense prediction as a discrete classification problem, assuming mutual exclusivity among senses (Navigli, 2009; Raganato et al., 2017). . However, since sense applicability is inherently graded rather than strictly categorical (Erk and McCarthy, 2009), AmbiStory (Gehring et al., 2026) frames WSD as a continuous plausibility estimation task over narratives, allowing multiple senses to be partially compatible.

Neural approaches to WSD have evolved from LSTM-based architectures (Yuan et al., 2016) to Transformer-based models (Vaswani et al., 2017). Pre-trained models like BERT (Devlin et al., 2019) significantly improved contextual learning, enabling stronger semantic performance. A relevant development is Gloss-BERT (Huang et al., 2019), which reformulates WSD as a sentence-pair task by jointly encoding a target sentence and candidate sense gloss. This allows contextual tokens to directly attend to definition tokens, improving disambiguation.

Recent studies further distinguish between *Bi-Encoder* and *Cross-Encoder* architectures (Reimers and Gurevych, 2019; Hadiwinoto et al., 2019). Bi-Encoders encode context and gloss independently and compare them in embedding space, offering computational efficiency but limiting fine-grained interaction. In contrast, Cross-Encoders jointly process both inputs, enabling token-level attention between contextual cues and sense definitions. Cross-Encoders have been shown to outperform Bi-Encoders in tasks requiring precise semantic alignment.

Our work builds upon the Gloss-BERT framework by extending the input from a single sentence to a full five-sentence narrative, incorporating pre-context and story resolution. This modification enables the model to evaluate sense plausibility based on the complete narrative trajectory rather than localized lexical cues.

3 System Overview

The full layout of our approach is shown in Figure 2. Our system treats the graded word sense plausibility task as a semantic regression task. It models the direct interaction between narrative trajectory and sense semantics.

To accomplish this, we created a unified BERT-based Cross-Encoder architecture.

3.1 Component 1: Cross-Encoder Regression Architecture

The core engine of our system is `bert-base-uncased`. We chose it because it captures deep bidirectional context. Unlike Bi-Encoder architectures that compress the story and meaning into separate vectors and lose detailed interaction, our Cross-Encoder approach makes the model focus on every token interaction between the narrative cues and the meaning definition.

Given a narrative context C (which includes the pre-context, ambiguous sentence, and ending) and a candidate meaning S (which includes the definition and example usage), the input sequence is built as:

$$X_{input} = \texttt{[CLS]} \oplus C \oplus \texttt{[SEP]} \oplus S \oplus \texttt{[SEP]} \quad (1)$$

The model processes this sequence to produce a contextualized embedding for the `[CLS]` token, $\mathbf{h}_{\texttt{[CLS]}}$. This representation is projected via a linear regression head to a scalar score $\hat{y}_{raw}$:

$$\hat{y}_{raw} = \mathbf{W} \cdot \mathbf{h}_{\texttt{[CLS]}} + b \quad (2)$$

The model is trained using Mean Squared Error (MSE) loss compared to the human average ratings, enabling it to learn the *degree* of plausibility instead of just a correct or incorrect label.

3.2 Component 2: Distributional Calibration

While the Cross-Encoder effectively ranks senses, the raw regression outputs from BERT often suffer from "scale compression," clustering around the mean of 2.5 to 3.5 instead of using the full 1 to 5

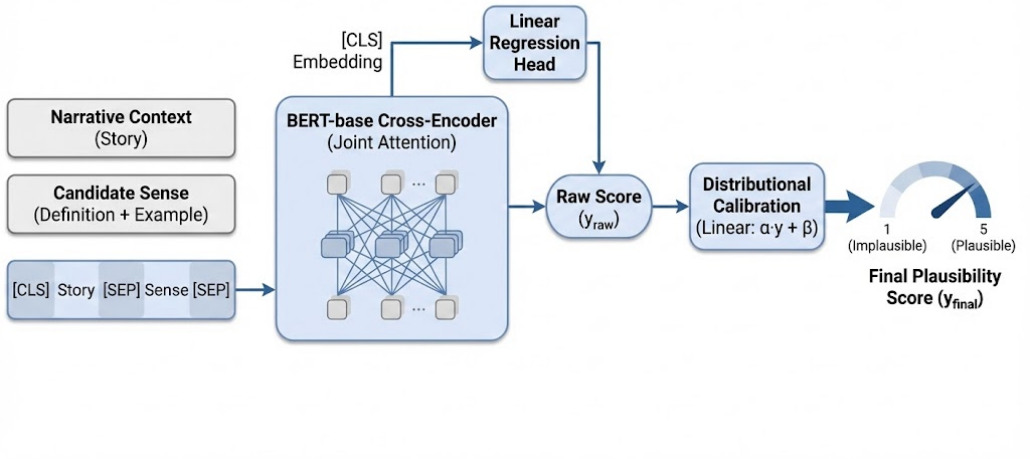

Figure 2: **System Architecture:** The BERT Cross-Encoder (left) jointly processes the narrative context and candidate sense. The raw regression output is then passed through a linear calibration layer (right) to align the scores with the human 1–5 plausibility scale.

range. To solve this, we use a linear calibration step trained on the development set.

For a raw prediction $\hat{y}_{raw}$, the final calibrated score $\hat{y}_{final}$ is computed as:

$$\hat{y}_{final} = \alpha \cdot \hat{y}_{raw} + \beta \tag{3}$$

where α (scaling factor) and β (shift) are coefficients fitted via least-squares regression. This component serves as a distribution aligner. It makes sure that highly plausible senses get scores close to 5.0, while implausible ones get scores near 1.0.

3.3 Illustrative Example

To illustrate the system's reasoning, consider a story containing the ambiguous word **"bank"**:

> *Context:* "John walked down the path. He saw the water flowing quickly. He sat down on the **bank**."

- **Candidate A (Financial Institution):** The definition involves "money" and "deposits." Although the word "bank" matches, the Cross-Encoder attention mechanism detects a semantic conflict between "water/flowing" in the context and "financial" in the definition.

- **Candidate B (River Side):** The definition involves "sloped land" and "water edge." The model attends to the explicit overlap of "water" in the story and "water" in the definition.

- **Result:** The raw BERT score might be 2.1 for A and 3.8 for B. After calibration, the system outputs $\hat{y}_{final}(A) \approx 1.2$ (Implausible) and $\hat{y}_{final}(B) \approx 4.9$ (Highly Plausible), matching human intuition.

4 Experimental Setup

4.1 Data and Preprocessing

We used the official AmbiStory dataset for SemEval-2026 Task 5. To create the input for our Cross-Encoder, we flattened the narrative structure. We combined the *pre-context*, *ambiguous sentence*, and *ending* to form the first segment, and we paired the *sense definition* with the *example usage* to create the second segment.

We applied the standard WordPiece tokenizer tied to `bert-base-uncased`. To ensure we didn't lose the important disambiguation cues at the end of the story, we set the maximum sequence length to 512 tokens. The average length of the stories plus definitions rarely exceeds this limit, so there was minimum truncation , mainly affecting the example usage field rather than the narrative itself.

4.2 Configuration and Hyperparameters

All experiments were conducted using the PyTorch framework and the Hugging Face `transformers` library.

- **Model Architecture:** We used `bert-base-uncased` with a regression

head (`num_labels=1`). The model trained to minimize the Mean Squared Error (MSE) loss between the predicted scalar and the human average rating.

- **Training Dynamics:** We applied the AdamW optimizer with a learning rate of 1×10^{-5} and a linear decay schedule. Because of GPU memory limits, we set a per-device batch size of 8 with gradient accumulation steps at 2, giving us an effective batch size of 16. The model trained for 10 epochs with mixed-precision (FP16) enabled to speed up convergence.

- **Checkpoint Selection:** To avoid overfitting, we assessed the model on the development set at the end of each epoch. We saved the checkpoint that got the highest Spearman rank correlation, focusing on ranking quality instead of just MSE loss.

- **Calibration:** The linear calibration parameters (α, β) were fitted using Ordinary Least Squares (OLS) regression on the development set predictions after the main training phase.

4.3 Evaluation Metrics

Following the official shared task guidelines, we report performance on two key metrics:

1. **Spearman Rank Correlation (ρ):**This is our main measure of ranking quality, checks how well the model arranges the likelihood of different meanings compared to human intuition.

2. **Accuracy within Standard Deviation:** This metric takes human subjectivity into account. A prediction is seen as correct if it falls within the range $[y_{human} - \sigma, y_{human} + \sigma]$, where σ is the standard deviation of the human ratings for that specific sample.

5 Results and Analysis

5.1 Official Evaluation Results

Our system was evaluated using the official SemEval-2026 Task 5 scoring script. On the hidden test set, our submission achieved a **Spearman rank correlation of 0.4866** ($p < 1.9 \times 10^{-56}$) and an **Accuracy-within-Standard-Deviation of 0.6613** (615/930 instances).

Spearman correlation measures the quality of plausibility ranking relative to human judgments, while Acc-within-SD evaluates whether predictions fall within the annotator disagreement range for each instance. The Acc-within-SD of 0.6613 indicates that the system's predictions align with the human consensus distribution for approximately 66% of test samples, which is notably higher than the naïve majority-class rate, suggesting the model captures meaningful plausibility signals even where exact ranking is difficult.

Model Configuration	Architecture	Spearman (ρ)
Sentence-Transformer	Bi-Encoder	0.52
Cross-Encoder (Initial)	BERT-Base	0.58
Cross-Encoder (Tuned)	**BERT-Base**	**0.66**
Cross-Encoder	BERT-Large	0.64

Table 1: Development set performance comparison across model variants.

5.2 Development Set Performance and Ablation Study

During development, we conducted systematic experiments to evaluate architectural choices and training strategies. Table 1 summarizes the performance progression.

The results show that switching from a bi-encoder to a cross-encoder architecture significantly improves the performance. The Sentence-Transformer baseline had a Spearman correlation of 0.52, which means it was not effective at modelling the fine-grained interaction between the narrative context and the candidate senses. Replacing this with a BERT-based cross-encoder improved performance to 0.58, which shows how important it is to encode both context and sense together.

Further hyperparameter tuning, including gradient accumulation, mixed-precision training, and Spearman-based checkpointing, raised the development set performance to 0.66. We also tested BERT-Large, scoring 0.64. However, the tuned BERT-Base configuration provided a better balance between performance and efficiency.

5.3 Test-Development Performance Gap

The tuned cross-encoder achieved a Spearman correlation of 0.66 on the development set, while the official test set score of 0.4866 indicates a notable drop. This gap reflects moderate overfitting to the development distribution. To mitigate this, we employed Spearman-based checkpoint selection rather than minimising training loss directly, and experimented with increased dropout (0.2) and weight decay ($\lambda = 0.01$). This did not fully close the gap, suggesting that the drop may partly reflect distri-

butional differences between the development and test narratives rather than overfitting alone.

It is worth noting that despite the drop in Spearman correlation, the system achieved a comparatively high Acc-within-SD of 0.6613. This suggests that while the exact ordering of senses may shift across annotation instances, the predicted plausibility scores remain broadly consistent with the human consensus distribution.distribution.

5.4 Quantitative Insights

The big difference in performance between the bi-encoder baseline (0.52) and the tuned cross-encoder (0.66) shows importance of model token-level interactions. Putting all five sentences into one embedding, like in sentence-transformer architectures, creates information bottlenecks that make it hard to see subtle cues that help you figure out what something means. The cross-encoder, on the contrary, lets the sense definition directly respond to certain lexical triggers and narrative resolutions.

Also, linear post-hoc calibration was helpful for matching raw regression outputs with the 1 to 5 rating scale used by people. Predictions tended to cluster in a narrow mid-range (2.0 to 4.0) without calibration. Calibration, on the other hand, spread out the distribution to better match how people score things, which improved the Accuracy-within-SD metric.

6 Conclusion

We presented a BERT-based cross-encoder regression system for SemEval-2026 Task 5, addressing the limitations of discrete WSD in modelling human plausibility judgments. On the development set, the tuned BERT-Base cross-encoder with linear calibration achieved a Spearman rank correlation of 0.66, substantially outperforming the sentence-transformer bi-encoder baseline ($\rho = 0.52$). On the official test set, the system achieved a Spearman correlation of 0.4866 and an Accuracy-within-SD of 0.6613. Our experiments demonstrate that for plausibility estimation, "reading the ending" necessitates joint attention between the narrative and the definition, rather than mere semantic similarity.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449, Singapore. Association for Computational Linguistics.

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.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 5300–5309, Hong Kong, China. Association for Computational Linguistics.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):1–69.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1374–1385, Osaka, Japan. The COLING 2016 Organizing Committee.