ConText at SemEval-2026 Task 5: Rating Plausibility of Word Senses in Ambiguous Stories through Narrative Understanding

Fakeha Faisal[1*] **Syeda Zaidi**[1*] **Rubab Shah**[1*] **Azkaa Nasir**[1*]
Sandesh Kumar[1] **Abdul Samad**[1]

[1]Department of Computer Science, Habib University, Karachi, Pakistan
{ff08288, rs08104, sz08469, an08017}@st.habib.edu.pk
{sandesh.kumar1*, abdul.samad}@sse.habib.edu.pk

Abstract

Here, we report our system for SemEval-2026 Task 5 (Gehring et al., 2026), which predicts graded plausibility scores for target word senses in narrative context. We explore embedding-based similarity, transformer fine-tuning, and a three-stage curriculum combining WiC pretraining, Wasserstein distribution learning, and KL-based calibration. Our best model, **DeBERTa-xLarge** with curriculum training, achieves **78% accuracy** within one standard deviation and a **Spearman Correlation of 0.70**, with an overall test score of 0.74. Results show that distribution modeling better aligns with human plausibility judgments than single-score prediction.

1 Introduction

Word sense ambiguity arises when a word admits multiple plausible interpretations in context Cruse (2011). Traditional Word Sense Disambiguation (WSD) treats this as a categorical problem, selecting a single correct sense Navigli (2009). However, in natural narratives, interpretation is often subjective: different human annotators may assign varying degrees of plausibility to the same sense.

SemEval-2026 Task 5 reframes WSD as a graded prediction task. Instead of selecting one sense, systems must predict a plausibility score (1–5) that reflects the average judgment of independent human annotators, rather than a single gold label. This formulation models plausibility as a spectrum and requires capturing the average of human ratings.

We explore embedding-based similarity methods, transformer fine-tuning, and a curriculum-based distribution learning framework. Our results show that explicitly modeling rating distributions leads to stronger alignment with aggregated human judgments. Our code is publicly available on GitHub. † †

[*]Equal Contribution
[*]https://github.com/fakehafaisal/SemEval-Task-5-2026

2 Related Work

We drove our inspiration primarily from Yap et al. (2020), which introduced GlossBERT, a model that treats word sense disambiguation (WSD) as a sentence–gloss matching task using BERT's contextual embeddings. The model pairs each sentence containing an ambiguous word with its possible glosses and learns to select the most appropriate sense based on contextual similarity through a ranking mechanism, achieving an F1 score of 80.4 on benchmark datasets such as SemCor and Senseval-3. Although GlossBERT is designed to select a single correct sense, its underlying idea of context gloss similarity was adapted for SemEval 2026 Task 5 by transforming it into a regression model that predicts a plausibility score.

While Yap et al. (2020) approaches WSD as a classification task, ConSeC Barba et al. (2021) models it as a continuous sense comprehension problem, scoring the relatedness of a sentence with respect to each gloss using BERT-like contextual embeddings, enabling the model to represent finer-grained semantic differences in terms of how likely each meaning is rather than as a binary choice. We also see that some works use graph-based approaches to address WSD, such as Kwon et al. (2021), which enhances graph-based WSD systems by removing unrelated contextual words using knowledge bases like WordNet and BabelNet, showing that selective context filtering is effective across multiple benchmarks.

Furthermore, we also looked into WSD approaches using LLMs and found that while they are good for WSD because of rich context representation, they also face some limitations. Liu et al. (2023) found that models such as GPT-3, BERT-large and T5 performed well (83%) on single-sense WSD tasks but showed poor accuracy of only 41% on multi-sense WSD data which had a poor correlation (0.34) to

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1840–1845

human plausibility ratings. Additionally, findings by Yae et al. (2025) showed that providing full narrative context rather than single isolated sentences can improve the ability of LLMs to disambiguate word senses. While we didn't employ LLMs due to computational constraints, this finding inspired us to concatenate pre-context, sentence, and ending to form stories for better narrative context.

3 Dataset

3.1 Structure and Format

The training set contains **2,280** annotated entries, while the validation set includes **588** entries. Finally, the test set includes **930** entries.
Each entry contains the `homonym` (the ambiguous word whose one meaning is being evaluated) and `judged_meaning` (the candidate sense meaning whose plausibility is being rated), alongside a short story formed by three fields: `precontext` (preceding sentences forming the background), `sentence` (the main sentence containing the target homonym), and `ending` (an optional continuation or conclusion). It also contains human plausibility judgments, obtained from five distinct human annotators who each rate how plausible they consider the judged meaning to be for the given short story containing the target homonym. These scores are on a scale 1–5. The other fields include: `average` (the mean of the five human ratings, used as the gold-standard plausibility score), and `stdev` (the standard deviation across the ratings, reflecting annotators agreement). Finally, `example_sentence` provides a sentence showing a clear usage of the homonym where it is expressing its judged meaning.

Another dataset we employ is Word-in-Context (WiC), which is formulated as a binary classification task that determines whether a target word is used with the same meaning in two different contexts Pilehvar and Camacho-Collados (2019).

4 Methodology

We employed multiple approaches which evolved progressively. Initially, we worked with embedding-based similarity models, introduced agreement-aware data augmentation and architectural scaling, and then transitioned to a distributional ordinal modeling framework trained using a three-stage curriculum.

4.1 Input Representation

Our input consists of the story (pre-context, target sentence marked with `[TGT]` tokens, and ending) and the judged meaning (definition and example sentence). We adopt a cross-encoder formulation in which the full story and the candidate meaning are jointly encoded, allowing fine-grained interaction between contextual and definitional representations.

4.2 Phase 1: Embedding-Based Similarity Models

We initially framed the task as semantic similarity estimation.

(1) Story vs. Meaning Similarity. We computed cosine similarity between contextualized embeddings of the entire story and the judged meaning. The similarity score was mapped to the plausibility scale.

(2) Target Word vs. Meaning Similarity. We refined this by extracting the contextual embedding of the marked target word and comparing it directly with the meaning embedding.

(3) GWSD Augmentation with Agreement-Aware Grouping. To mitigate data sparsity, we identified Graded Word Sense Disambiguation (GWSD) dataset Cassotti and Tahmasebi (2025), whose structure closely aligned with that of SemEval, although its rating scale ranged from 0 to 4 rather than 1 to 5, and did not include an example sentence. The GWSD scale differs structurally from SemEval's 1–5 scale: a score of 0 in GWSD denotes annotator uncertainty or abstention rather than minimal plausibility, making it semantically distinct from scores 1–4. We therefore treat 0 as a special neutral category and map it to the SemEval midpoint (3), while mapping the remaining ordinal scores monotonically: $1 \rightarrow 1$, $2 \rightarrow 2$, $3 \rightarrow 4$, $4 \rightarrow 5$. This preserves ordinality within the plausible range while placing uncertain annotations at the scale center. We acknowledge this introduces uneven interval spacing; future work could explore filtering out GWSD-0 instances entirely. Example sentences were generated via Gemini Flash 2.5 and were not manually reviewed or validated against genuine linguistic usage, and may introduce noise into the training signal.

Rather than treating all additional data uniformly, we partitioned instances using empirical mean plausibility (μ) and standard deviation (σ) into three

groups: **G1** (low mean, low variance: $\sigma \leq 1.5$, $\mu \leq 3$), **G2** (high mean, low variance: $\sigma \leq 1.5$, $\mu > 3$), and **G3** (high variance: $\sigma > 1.5$). G1 and G2 represent high-agreement instances differing primarily in plausibility polarity, while G3 captures inherently ambiguous examples with strong annotator disagreement.

We progressively incorporated these subsets (G1), (G1+G2), and (G1+G2+G3) during experimentation. Performance improved consistently as supervision increased, while inclusion of G3 exposed the model to genuine semantic ambiguity.

(4) Ensemble Training. We applied bagging across multiple transformer encoders to stabilize embedding-based predictions.

(5) Additional Text Augmentation. For further text augmentation, We experimented with random insertion and deletion, back-translation (German, French),synonym replacement (WordNet),gloss back-translation, hypernym concatenation augmentation, and Synthetic gloss and example generation.

Augmentation was applied with awareness of agreement structure. High-agreement groups (G1, G2) were treated as reliable supervision, while high-variance examples (G3) were preserved without aggressive transformation to avoid amplifying annotation noise. Gloss back-translation and lexical enrichment proved more stable than fully synthetic generation.

Although these methods improved robustness, embedding similarity alone struggled to model ordinal structure and human disagreement.

4.3 Phase 2: Strengthening Representation and Supervision

Before moving to full distribution modeling, we explored intermediate improvements to representation quality and optimization.

(6) Example-Sentence Anchored Comparison Here, we reframed the task to now compare the embedding of the target word in its example sentence to its embedding in the story. While alignment improved, predictions remained scalar and failed to capture rating variance.

(7) WiC Transfer Learning For binary sense discrimination, we employed intermediate training on the Word-in-Context (WiC) Pilehvar and Camacho-Collados (2019) task. We mapped the WiC labels to endpoints of the AmbiStory graded scale. This helped the model to learn sense distinctions prior to fine-grained plausibility modeling leading to improved final performance.

(8) Model Scaling We compared BERT, RoBERTa, ELECTRA, MPNet, DeBERTa-Large, and DeBERTa-xLarge. Scaling to DeBERTa-xLarge yielded best performance, particularly when combined with expanded SemEval+GWSD dataset.

(9) Variance-Aware Optimization Experiments Here, we tested loss reweighting methods through two approaches, i.e. increasing the weight for low-variance data points and testing the Wasserstein loss. The experiments showed that explicit modeling of ordinal distance leads to better results that match human judgment standard. The observations led us to transition from using scalar regression to complete distribution modeling as our new approach.

4.4 Phase 3: Distribution-Matching Framework

Another approach we used was to model human plausibility judgments as distributions rather than single labels. Inspired by the SORD framework Díaz and Marathe (2019), which introduces soft labels for ordinal regression, we adopted a distributional prediction strategy.

Now, instead of soft labels using Gaussian smoothing around a single score, we directly normalized the empirical human rating distribution. This method maintains actual semantic uncertainty which exists via disagreements between the annotators while preventing the creation of artificial noise patterns.

4.5 Three-Stage Curriculum Training

After conducting the above experiments in isolation, we chose the best yielding approaches and adopted a curriculum learning strategy.

4.5.1 Stage 1: WiC Pretraining

Before plausibility prediction, we adapted the backbone model using the Word-in-Context (WiC) task. We applied target token marking, sentence-pair cross-encoding, partial layer freezing (updating only upper transformer layers), and cross-entropy optimization. This stage produced a sense-aware backbone prior to graded plausibility learning.

4.5.2 Stage 2: Wasserstein Distribution Learning

We trained the model to predict a probability distribution over five ordinal classes.

To explicitly model ordinal distance, we optimized Earth Mover's Distance (EMD) Hou et al. (2016), which penalizes predictions proportionally to their ordinal displacement, using the hybrid loss $\mathcal{L} = \mathbf{0.8}\,\mathcal{L}_{\mathbf{EMD}} + \mathbf{0.2}\,(\mathbb{E}_p - \mathbb{E}_{\hat{p}})^{\mathbf{2}}$, where the mixing coefficients were selected based on validation set performance and the regression component aligns expected values with correlation-based evaluation metrics.

4.5.3 Stage 3: High-Agreement KL Calibration

Finally, we fine-tuned the model on filtered high-consensus subsets (G1 and G2; low variance examples).

We optimized Kullback–Leibler (KL) divergence Joyce (2011) between predicted and empirical distributions using $\mathcal{L} = \mathbf{0.7}\,\mathcal{L}_{\mathbf{KL}} + \mathbf{0.3}\,(\mathbb{E}_p - \mathbb{E}_{\hat{p}})^{\mathbf{2}}$, which sharpens distribution alignment for confident examples while preserving ordinal structure. The mixing coefficients were similarly selected based on validation set performance.

4.6 Final Integrated Pipeline

Our final system integrates the strongest components identified across experimentation:

- WiC-based intermediate pretraining
- SemEval + GWSD (G1+G2+G3) supervision
- DeBERTa-xLarge encoder
- Gloss back-translation augmentation
- Distribution-aware hybrid loss (EMD + expectation regression)
- KL-based calibration on high-agreement subsets
- Full distribution prediction instead of single-score regression

This sequential refinement allowed us to transition from similarity-based modeling to robust distribution prediction, systematically combining data expansion, agreement-aware augmentation, transfer learning, model scaling, and ordinal-aware optimization into a unified framework.

4.7 Training Setup

All stages use batch size 16, a linear learning rate scheduler with warmup, and early stopping on development loss. Stage-specific learning rates are 1×10^{-5} for WiC pretraining and distribution learning, and 5×10^{-6} for KL calibration. Due to computational constraints, we were unable to run experiments across multiple random seeds. We therefore caution that small differences in (0.01–0.02) between adjacent configurations in Table 1 should be interpreted as indicative trends rather than statistically significant gains.

4.8 Output

The model outputs a probability distribution over the five plausibility classes. The final prediction is computed as the expected value of this distribution. The continuous expected score is used for correlation-based evaluation, while the rounded value is used for discrete evaluation.

5 Results and Discussion

Performance was evaluated using **Spearman Correlation** (ρ), measuring rank-order agreement with human plausibility ratings, and **Accuracy Within Standard Deviation** (Acc), measuring the proportion of predictions falling within one standard deviation of the mean human rating. Our best system (DeBERTa-xLarge, full 3-stage curriculum) achieves **78% accuracy** and ρ **= 0.70**, for an **average score of 0.74**, ranking **31st out of 79** participating teams.

Table 1 presents an implicit ablation isolating the contribution of each design decision. Phases 1–5 reveal that embedding similarity consistently plateaued at 58% accuracy despite expanded data and ensemble training. Comparing full story embeddings to meaning embeddings (Phase 1) produced only a weak signal (ρ = 0.33); localising to the target word (Phase 2) gave marginal improvement. This showed a fundamental limitation which is that cosine similarity produces scalar estimates with no ordinal structure and cannot represent the spread of human opinion. Only increasing data quantity or model variety (Phases 3–5) is not enough for graded plausibility.

Phase 6 disrupted the plateau by attaching comparisons to the usage of the target word in its example sentence rather than an abstract definition, combined with back-translation and a learning rate scheduler, reaching 69% accuracy and ρ = 0.46. The increase in Spearman correlation across 7a–7g reflects the change to complete cross-encoder fine-tuning, with WiC pretraining (7f) giving the strongest single boost (ρ = 0.68) by doing

Phase	Models	Data Augmentation / Key Approach	Acc (%)	Spearman ρ	Best Model
1	BERT-base, RoBERTa-base	Story vs. judged meaning embedding similarity	44	0.33	BERT-base
2	BERT-base, RoBERTa-base	Target word vs. judged meaning similarity	51	0.36	BERT-base
3	BERT-base, RoBERTa-base	Same as 2 + GWSD (G1 subset)	58	0.42	BERT-base
4	BERT-base, RoBERTa-base, DistilBERT, DeBERTa-v3-base, ELECTRA-small	Same as 3 + Ensemble Training (bagging)	58	0.43	BERT-base
5	BERT-base, RoBERTa-base, DeBERTa-large	Same as 2 + Random insertion/deletion	58	0.42	BERT-base
6	MPNet-base-v2, DeBERTa-large	Example sentence anchoring + Back Translation (German) + LR scheduler	69	0.46	MPNet-base-v2
		Expanded Supervision and Augmentation Experiments			
7a	DeBERTa-large	SemEval only	66	0.62	DeBERTa-large
7b	DeBERTa-large	SemEval + GWSD (G1+G2)	68	0.66	DeBERTa-large
7c	DeBERTa-large	SemEval + GWSD (G1+G2+G3)	69	0.67	DeBERTa-large
7d	DeBERTa-large	Gloss Back-Translation (G1,G2) + SemEval	65	0.59	DeBERTa-large
7e	DeBERTa-large	Hypernym concatenation + SemEval	75	0.60	DeBERTa-large
7f	DeBERTa-large	WiC $\rightarrow$ SemEval pretraining	69	0.68	DeBERTa-large
7g	DeBERTa-large	WiC $\rightarrow$ SemEval + GWSD (G1+G2+G3)	68	0.69	DeBERTa-large
		Distribution Modeling and Curriculum Training			
8a	DeBERTa-large	Soft-label distribution learning (KL + MSE) + GWSD	77	0.66	DeBERTa-large
8b	DeBERTa-xlarge	SemEval + GWSD (G1+G2+G3)	69	0.74	DeBERTa-xlarge
8c	DeBERTa-large	Wasserstein (EMD) loss	80	0.65	DeBERTa-large
8d	DeBERTa-large	Variance-weighted loss scaling ($\sigma < 1.5$)	75	0.65	DeBERTa-large
8e	DeBERTa-xlarge	Full 3-stage curriculum (WiC pretraining + EMD hybrid + KL calibration)	**78**	**0.70**	**DeBERTa-xlarge**

Table 1: Expanded summary of modeling approaches, augmentation strategies, and performance metrics across all experimental phases.

sense discrimination before graded prediction. Expanded GWSD supervision (7b$\rightarrow$7c) further improved both metrics.

Phase 8 introduced distribution modeling. Rather than predicting a single score, the model outputs a probability distribution over five ordinal classes. Optimising EMD loss (8c) and KL calibration fits better with evaluation metrics that compare against aggregated human judgments. The full 3-stage curriculum (8e) brings all of this together for the best overall results.

Error Analysis: Per-group analysis (Table 2) reveals additional patterns. G3 instances ($\sigma > 1.5$) actually hit the highest accuracy (93.0%), likely because high-variance items tend to have higher mean scores which are easier to predict; however, their Spearman correlation (0.46) is only moderate, meaning the model struggles to rank within that ambiguous middle ground. G1 (low mean, low variance) flips this: lowest accuracy (68.0%) but highest correlation (0.56), so the model orders low-plausibility examples well but gets the actual scores wrong. G2 has the weakest correlation (0.36), meaning high-plausibility cases are surprisingly the hardest to rank despite decent accuracy. Additionally, instances where story context diverges strongly from the provided example sentence reduce anchoring effectiveness, and some cases are just genuinely ambiguous, where no single distribution can satisfy all annotators.

Group	Acc (%)	ρ
G1 (low mean, low var.)	68.0	0.56
G2 (high mean, low var.)	82.0	0.36
G3 (high variance)	93.0	0.46

Table 2: Per-group performance of best model (DeBERTa-xLarge, SemEval+GWSD G1+G2+G3).

6 Conclusion

We present a progressive modeling framework for SemEval-2026 Task 5, evolving from embedding-based similarity methods to a curriculum-trained distributional prediction model. Our findings show that graded plausibility prediction is inherently ordinal and distributional: modeling the full empirical distribution of independent human ratings leads to stronger alignment with aggregated human judgments than scalar regression or similarity-based approaches.

7 Limitations

First, our approach relies on large transformer encoders and multi-stage training, resulting in high computational cost and limited scalability. The system is currently evaluated only on English narratives, restricting cross-lingual generalization.Loss mixing coefficients were tuned on the validation set

without systematic sensitivity analysis; their robustness across different data splits remains unverified.

Second, our distributional approach assumes annotator homogeneity and does not explicitly model individual annotator biases. As a result, systematic differences in annotator interpretation may be conflated with genuine semantic ambiguity, limiting the interpretability of predicted distributions.

Furthermore, our ablation results are reported for single runs without confidence intervals or significance testing across random seeds, owing to the high computational cost of multi-stage curriculum training on DeBERTa-xLarge. Marginal performance differences between configurations may therefore not reflect robust gains. Future work should validate key design decisions across multiple seeds.

Finally, performance gains from curriculum training are smaller for highly context-dependent homonyms, indicating sensitivity to data sparsity and imbalance.

8 Future Work

Future research could explore uncertainty-aware or annotator-aware modeling to better capture structured human disagreement, as well as parameter-efficient fine-tuning of instruction-tuned LLMs with distribution-aware objectives could combine narrative reasoning strengths with calibrated ordinal prediction. Extending the framework to multilingual settings and low-resource languages would test the robustness of curriculum-based distribution learning beyond English.

References

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021. Consec: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503.

Pierluigi Cassotti and Nina Tahmasebi. 2025. Sense-specific historical word usage generation. *Transactions of the Association for Computational Linguistics*, 13:690–708.

Alan Cruse. 2011. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press UK, Oxford.

Raúl Díaz and Amit Marathe. 2019. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4733–4742.

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.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. 2016. Squared earth mover's distance-based loss for training deep neural networks. volume abs/1611.05916.

James M. Joyce. 2011. *Kullback-Leibler Divergence*, pages 720–722. Springer Berlin Heidelberg, Berlin, Heidelberg.

Sunjae Kwon, Dongsuk Oh, and Youngjoong Ko. 2021. Word sense disambiguation based on context selection using knowledge-based word similarity. *Information Processing and Management*, 58(4):102551.

A. Liu and 1 others. 2023. We're afraid language models aren't modeling ambiguity.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10:1–10:69.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Word-in-context: Evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1267–1273. Association for Computational Linguistics.

J.H. Yae, N.C. Skelly, N.C. Ranly, and 1 others. 2025. Leveraging large language models for word sense disambiguation. *Neural Computing and Applications*, 37:4093–4110.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. 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.