Ambirig at SemEval-2026 Task 5: Distributional Ordinal Modelling for Ambiguous Word Senses in Narrative Contexts

Soumyajit Roy
Setu
Bengaluru, India
roysoumyajit@icloud.com

Abstract

Word Sense Disambiguation (WSD) (Navigli, 2009) has traditionally been framed as selecting a single correct sense given context. However, natural language understanding by humans often involves ambiguity, underspecification, and graded plausibility judgments rather than categorical decisions. SemEval-2026 Task 5 explicitly addresses this gap by requiring systems to predict human-perceived plausibility scores for word senses in short narratives. In this paper, we present a systematic empirical study of modelling plausibility as an ordinal distribution prediction problem. We hypothesise that standard classification objectives fail to capture the ordinal nature of human uncertainty in this domain. While we experimented with complex auxiliary tasks, including Siamese networks, Task-Adaptive Pretraining (TAPT), and transfer learning from Natural Language Inference (NLI), our results show that these approaches fail in low-resource settings. Instead, we propose a streamlined architecture based on DeBERTa-v3-base utilising a GlossBERT-style cross-encoder optimised with Earth Mover's Distance (EMD) loss. By modelling the problem as ordinal regression over a probability distribution and enriching inputs with prototypical examples, our system achieves an accuracy of 73% and Spearman correlation of 0.593, establishing a robust baseline that outperforms complex parameter-heavy approaches.

1 Introduction

Word sense ambiguity is not merely a modelling inconvenience but a defining characteristic of human language understanding. In narrative contexts, multiple interpretations of a homonym may remain simultaneously plausible, even after observing surrounding context. SemEval-2026 Task 5 introduces a more nuanced challenge: rating the plausibility of a sense on a continuous 1-5 scale (Gehring et al., 2026). This shifts the paradigm from identifying a mode to modelling a distribution of human judgment.

This task raises a fundamental question: **should plausibility be predicted as a scalar value, or as a distribution reflecting human uncertainty?** Our work argues for the latter.

This task raises another fundamental modelling question: **How should models represent meaning when ambiguity is not noise, but signal?** Many standard WSD approaches, despite strong architectures, implicitly enforce sense resolution, exclusivity, or sharp ordering. In this work, we show that such inductive biases are misaligned with graded plausibility estimation.

1.1 Contributions

This paper makes the following contributions:

- **Distributional Ordinal Modelling with EMD:** We introduce an Earth Mover's Distance–based loss for modelling full plausibility distributions over Likert-scale ratings, explicitly respecting ordinal structure and human disagreement.
- **Architectural Control via GlossBERT-style Cross-Encoders:** By using a context–gloss cross-encoder architecture similar to GlossBERT, we isolate the effect of supervision and loss formulation, showing that changing the learning target matters more than changing the model.
- **Systematic Ablation of Common WSD Techniques:** Through extensive experiments, we demonstrate that widely used techniques, including scalar regression, contrastive learning, ranking losses, entropy regularisation, complementarity constraints, and task-adaptive pretraining, often degrade performance under genuine ambiguity.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 585–591

Together, these findings argue that preserving ambiguity - not resolving it - is essential for modelling graded semantic interpretation.

2 Task and Data

The task provides a short narrative context, a target homonym, and a specific sense definition. The goal is to predict the mean plausibility rating derived from multiple human annotators. The dataset is low-resource ($N = 2{,}280$ for training) and covers English. Each datapoint consists of a precontext (three sentences), an ambiguous sentence containing a homonym, optionally, an ending that may bias interpretation and a candidate sense definition.

The dataset exhibits substantial inter-annotator variance: a large proportion of instances have no clear consensus, with ratings spread across multiple Likert values. This makes the task fundamentally different from categorical WSD and motivates treating annotator disagreement as signal rather than noise.

Annotators rate the plausibility of each sense on a Likert scale from 1 (very implausible) to 5 (very plausible). Unlike traditional WSD datasets, annotations exhibit substantial variance, motivating distributional supervision.

3 Related Work

3.1 Gloss-based WSD

GlossBERT (Huang et al., 2019) and subsequent work (Bevilacqua et al., 2021) model WSD as sentence-pair classification between context and sense definitions. While effective for categorical disambiguation, these methods optimise pointwise objectives and implicitly assume a single correct sense.

3.2 Ordinal and Distributional Learning

Ordinal regression and distributional supervision have been explored in affective computing and semantic similarity tasks (Li and Lin, 2006; Baccianella et al., 2009). Earth Mover's Distance has been shown to effectively model ordered label distributions (Geng and Zhao, 2014), but remains underexplored in sense plausibility estimation.

3.3 Task-Adaptive Pretraining

Task-adaptive pretraining (TAPT) has demonstrated gains in semantic similarity and entailment tasks (Gururangan et al., 2020). Its interaction with ambiguity-heavy supervision, however, has not been systematically studied.

4 Modelling Plausibility as Ordinal Distributions

4.1 Input Enrichment

Standard WSD models often rely solely on the definition text. We propose an input enrichment strategy. We construct the input to maximise semantic interaction between the sense definition and the target context. Furthermore, we inject the example usage provided in the dataset to serve as a concrete "prototype" of the sense.

We construct the input sequence X by concatenating the enriched gloss and the marked context:

$$X = [\text{CLS}] \oplus T(d, e) \oplus [\text{SEP}] \oplus C' \oplus [\text{SEP}]$$

where $T(d, e)$ represents the template concatenation of the definition d, the separator literal `" | Example: "`, and the example usage e.

Target Marking: To act as an inductive bias for the self-attention mechanism, we explicitly mark the target word in the context C'. If the homonym w appears at index j, we inject boundary markers (quotes):

$$C' = \ldots w_{j-1}, \text{“}, w_j, \text{”}, w_{j+1} \ldots$$

Without target marking, the model's attention often drifted to non-ambiguous words in the sentence.

4.2 Model Architecture

We utilise DeBERTa-v3-base (184M parameters) (He et al., 2023). We found that larger models (DeBERTa-Large, 435M) exhibited overfitting under the limited training data, suggesting that scaling model size without additional supervision may not yield improvements in this setting. The `[CLS]` embedding is projected to a vector $z \in \mathbb{R}^5$ via a linear layer and normalised via Softmax to produce a predicted probability distribution $\hat{y}$.

We standardised on DeBERTa-v3 due to its superior performance on NLU benchmarks relative to RoBERTa, given our compute constraints.

4.3 Optimisation: Earth Mover's Distance

We treat the labels as a probability mass over ordinal ratings $k \in \{1, \ldots, 5\}$. We construct the target probability distribution $\mathbf{y}$ by calculating the normalised frequency of each Likert rating

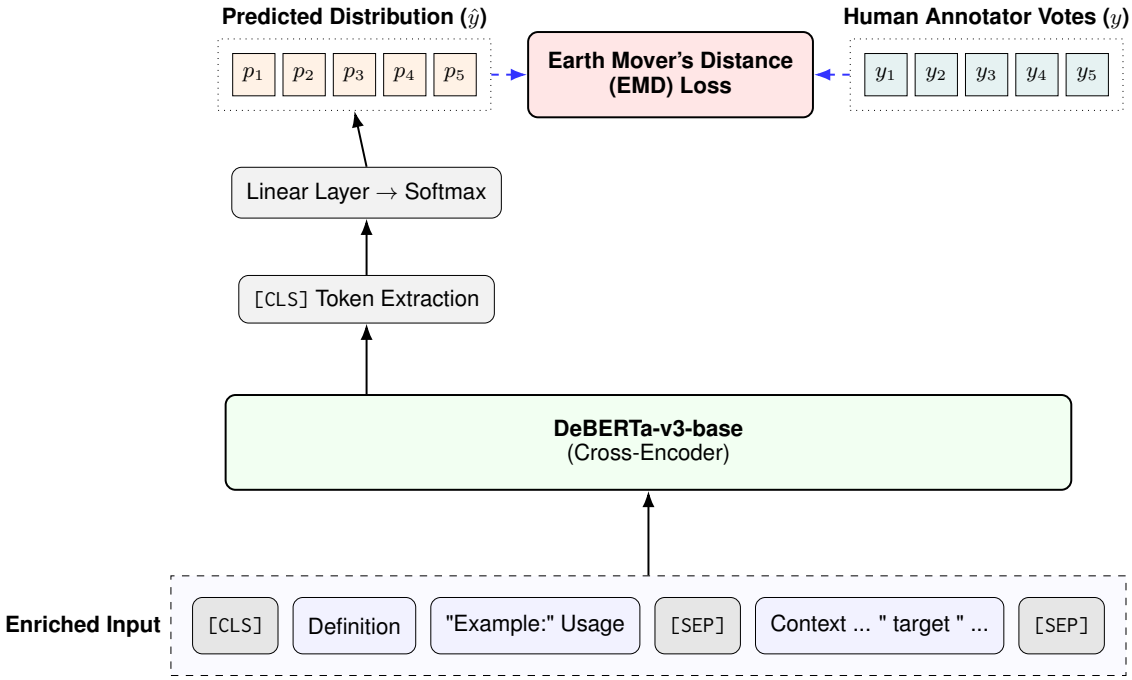

Figure 1: Overview of the proposed system. The enriched context and definition are processed via a Cross-Encoder. The [CLS] token is extracted and projected into a 5-class distribution, which is directly optimised against the raw human annotator distribution using Earth Mover's Distance.

(1 to 5) directly from the raw individual annotator votes. For example, if a sample receives five votes $[2, 1, 5, 3, 4]$, the target distribution is uniform:

$$\mathbf{y} = [0.2,\ 0.2,\ 0.2,\ 0.2,\ 0.2]$$

This explicitly preserves annotator disagreement rather than compressing it into an aggregated scalar mean.

We minimise the squared Earth Mover's Distance (EMD), which is computationally equivalent to the L_2 distance between the Cumulative Distribution Functions (CDFs).

Let $F_{\hat{y}}$ and F_y be the CDFs of the predicted and target distributions, where $F_y(k) = \sum_{j=1}^{k} y_j$. The loss is defined as:

$$\mathcal{L}_{\text{EMD}} = \frac{1}{5} \sum_{k=1}^{5} (F_{\hat{y}}(k) - F_y(k))^2$$

Why EMD? Unlike Cross-Entropy, which pushes the prediction only towards the mode, $\mathcal{L}_{\text{EMD}}$ provides gradients that push probability mass from distant classes towards the target (Hou et al., 2017). It naturally encodes the ordinal relationship.

5 Experimental Setup

We train on the official training set ($N = 2{,}280$) and use the development set ($N = 588$) for model selection and ablations. No external data was used to train our primary EMD submission. However, for specific ablation studies (e.g., the NLI initialisation detailed in Section 7.2), we utilised publicly available pre-trained checkpoints.

Model and Training. We fine-tune `microsoft/deberta-v3-base` with batch size 8 using AdamW (lr 1.5×10^{-5}, weight decay 0.01). Training runs for 5 epochs with cosine decay and 10% warmup, selecting the checkpoint with lowest validation loss.

All baseline models (e.g., MSE regression, contrastive learning) were subjected to identical training hyperparameters (batch size, learning rate, number of epochs), differing only in the loss function or auxiliary objective. This ensures fair comparison and isolates the effect of supervision. For all models, we swept learning rates $\in \{1 \times 10^{-5},\ 1.5 \times 10^{-5},\ 2{\times}10^{-5},\ 3{\times}10^{-5}\}$ and epochs $\in \{3, 5, 10\}$. The final EMD model utilises a learning rate of 1.5×10^{-5}, a batch size of 8, a cosine decay scheduler with a 10% warmup, and a sequence length of 300 tokens.

Notebooks containing preprocessing and model training scripts will be made publicly available.[1]

Evaluation. A prediction $\hat{y}$ is counted as correct if $|\hat{y} - \mu| \leq \sigma$, where μ and σ denote the mean

[1] https://github.com/roysammy/ambirig

Model Paradigm	Model	Loss / Strategy	Spearman ↑	Acc ($\pm\sigma$) ↑
Point Estimation	Cross-encoder regression	MSE	≈ 0.00	–
Distribution (unordered)	Soft-label regression	MSE	0.421	0.660
Metric Learning	Siamese cross-encoder	Contrastive	0.525	0.696
Consistency Regularisation	Cross-encoder + R-Dropout	EMD + KL (dropout)	0.513	0.687
Distribution (entropy)	Cross-encoder	Entropy loss	0.586	0.716
Distribution (ordinal)	Cross-encoder	Ordinal EMD	**0.593**	**0.730**
Distribution (ordinal)	Cross-encoder	CORAL	0.518	0.594
Hybrid (ordinal + ranking)	Cross-encoder	EMD + Pairwise	0.588	0.719
Structural Exclusivity	Cross-encoder	Complementarity	0.135	0.548
Domain Adaptation	TAPT + Cross-encoder	MLM + EMD	0.501	0.696

Table 1: Comparison of different models and loss strategies for word sense plausibility estimation.

Loss Variant	Formula	$\Delta\rho$	Analysis
Pure EMD	$\mathcal{L}_{\text{EMD}}$	0.000	Optimal Stability.
+ Ranking Loss	$+\lambda\mathcal{L}_{\text{rank}}$	-0.005	Redundant. EMD implies ranking.
+ Complementarity	$+\lambda\|\|p_A - (1-p_B)\|\|^2$	-0.458	Invalid. Ambiguity is non-exclusive.
+ Structural Entropy	$\mathbb{I}(\text{no\_ending}) \cdot H(p)$	-0.007	Not supported empirically.

Table 2: Constraints Ablation.

and standard deviation of human ratings. We report Spearman rank correlation (ρ) as the primary metric and Accuracy within Standard Deviation (Acc $\pm\sigma$) as the secondary metric.

6 Results and Ablation Analysis

6.1 Main Quantitative Findings

Following the shared task timeline, our primary ablation studies and architectural comparisons (Tables 1 and 2) report performance on the official development set. Our final submitted system achieved an accuracy of 66.6% and a Spearman correlation of 0.49 on the blind test set.

Table 1 presents the performance of our proposed system against various baselines and alternative architectures.

Point-estimate regression converged to predicting a narrow range of scores (≈ 3.4–3.5) across all inputs, effectively learning the global mean of the training distribution and failing to rank instances, while soft-label and contrastive models yield moderate gains ($\rho = 0.42$–0.53). Consistency regularisation with R-Drop improves stability but remains limited ($\rho = 0.51$). Entropy-based objectives better capture uncertainty ($\rho = 0.59$) but ignore ordinal structure. Our ordinal distributional model using Earth Mover's Distance achieves the best overall performance ($\rho = 0.593$, Acc $= 0.730$), confirming that jointly modelling order and annotator disagreement is essential. CORAL-style cumulative link ordinal regression improves over non-ordinal baselines but remains substantially below EMD, suggesting that monotonic threshold-based ordinal models are less suited for capturing graded annotator disagreement than distributional objectives. In contrast, exclusivity constraints and task-adaptive pretraining substantially degrade performance, underscoring the importance of ambiguity-preserving objectives.

6.2 Ablation Study: The Sufficiency of EMD

We tested whether adding explicit logical or structural constraints could improve the EMD baseline. As shown in Table 2, EMD captures the important inductive bias required for this task.

The similarity between EMD and pairwise ranking performance is expected: EMD implicitly encodes pairwise ordering constraints through cumulative distance minimisation. The lack of additive gains confirms that explicit ranking losses are redundant once ordinal structure is enforced.

6.3 Robustness Across Seeds and Data Splits

To assess robustness, we evaluate EMD across multiple random seeds and training splits. Across three seeds (13, 42, 87), EMD achieves a mean Spearman correlation of 0.584 with a standard deviation of 0.036, indicating stable performance and low sensitivity to initialisation.

We further perform five-fold cross-validation to estimate variance under reduced training data. While fold-level training achieves a mean correla-

tion of 0.587 ± 0.047, inference using fold-trained models yields lower performance due to reduced data per fold. We therefore follow standard practice and use the full training set for final inference, reporting cross-validation solely as a robustness diagnostic.

The improvement of Ordinal EMD over entropy-based objectives is consistent across seeds and exceeds one standard deviation of random initialisation, suggesting the gain is systematic rather than stochastic.

7 Analysis and Discussion

7.1 Is This Just GlossBERT with Ordinal Regression?

Architecturally, yes: both use context–gloss cross-encoders. Conceptually, no.

Aspect	GlossBERT	Our Model
Target	Single best sense	Distribution of plausibility
Supervision	Point label	Annotator distribution
Loss	CE / MSE	Ordinal EMD
Ambiguity	Collapsed	Preserved

Table 3: Comparison between GlossBERT and our model.

While our architecture resembles a GlossBERT-style context–gloss cross-encoder, the underlying modelling objective is fundamentally different. GlossBERT is designed to identify a single best sense through point-label supervision and cross-entropy or regression losses, thereby collapsing ambiguity by construction. In contrast, our model explicitly predicts the full distribution of human plausibility judgments and optimises an ordinal Earth Mover's Distance objective, preserving graded ambiguity rather than eliminating it. This distinction in supervision and loss formulation is central to the performance gains observed in our experiments.

7.2 Limitations of Complex Approaches

A significant contribution of this work is identifying techniques that degrade performance in low-resource ambiguity tasks.

Catastrophic Forgetting in TAPT: Task-Adaptive Pretraining (TAPT) (Gururangan et al., 2020) is standard for domain adaptation. However, our TAPT experiments degraded performance ($\rho = 0.501$). We attribute this to the small corpus size. Pre-training on ∼2,000 simple stories caused the model to overwrite the rich semantic knowledge of homonyms acquired from general pre-training (e.g., distinguishing "dough" as money vs. bread), effectively over-specialising the model to the limited domain.

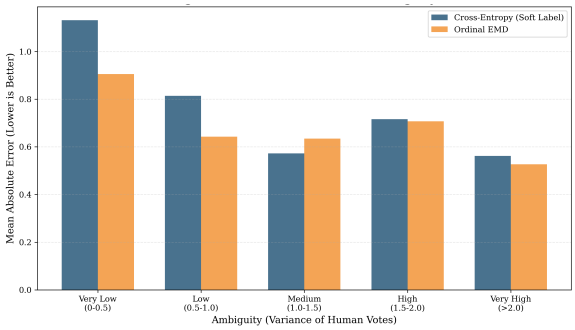

Figure 2: Mean Absolute Error (MAE) stratified by sample ambiguity (variance of human votes). The Ordinal EMD model (green) consistently achieves lower error rates than the Cross-Entropy baseline (red), demonstrating robustness across both high-consensus and high-ambiguity regimes.

Logic ≠ Plausibility (NLI Failure): Initialising weights from an NLI-tuned model (`cross-encoder/nli-deberta-v3-base`) resulted in poor correlation ($\rho = 0.447$). While NLI models excel at binary truth conditions (Entailment/Contradiction), they struggle with the graded, subjective nature of plausibility. This indicates that logical entailment and narrative plausibility require distinct representational subspaces.

7.3 Limitations of Complementarity Constraints

The dramatic failure of the Complementarity constraint ($\rho = 0.13$) is scientifically revealing. This constraint attempted to enforce that if Sense A is plausible, Sense B must be implausible ($P(A) \approx 1 - P(B)$). The failure suggests that ambiguity in this dataset is non-exclusive. Annotators frequently rate competing senses as equally moderate (e.g., rating both 3), or the context is vague enough to support neither. EMD accommodates these "flat" distributions naturally, whereas logical constraints force a polarity that contradicts the ground truth.

7.4 Robustness to Ambiguity

A core challenge of this task is that "ground truth" is often a high-variance distribution rather than a consensus. We analysed how our model performs across different levels of annotator disagreement (ambiguity), quantified by the variance of the hu-

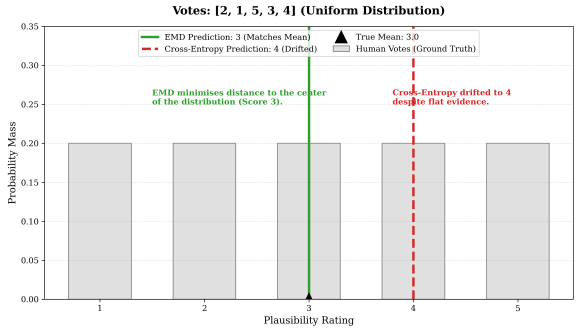

Figure 3: Qualitative analysis of Sample ID 117. The ground truth (grey bars) represents a highly ambiguous narrative where human annotators were uniformly divided. The EMD model (green) correctly identifies the distributional center (Score 3), whereas the Cross-Entropy model (red) drifts to Score 4, illustrating EMD's stability in high-entropy scenarios.

man votes.

Quantitative Analysis: As shown in Figure 2, we stratified the test set by vote variance. The Ordinal EMD model (green) consistently achieves lower Mean Absolute Error (MAE) compared to the Cross-Entropy baseline (red). Notably, in high-ambiguity scenarios (variance >2.0), EMD maintains robustness, whereas Cross-Entropy degrades. This confirms that minimising cumulative distance is a superior objective for modelling disagreement than minimising KL-divergence (i.e., Cross-Entropy / KL-divergence objectives), which tends to over-penalise flat distributions.

Qualitative Case Study: To visualise this behaviour, Figure 3 highlights Sample ID 117, a case of maximal ambiguity where human annotators were uniformly divided across ratings 1 to 5. The Cross-Entropy model, forced to pick a mode, drifted to a prediction of 4. In contrast, the EMD model correctly identified the distributional "center of mass" (Score 3). By respecting the ordinal topology of the label space, EMD effectively "hedges" its prediction in the center when faced with maximum uncertainty, minimising the expected error.

8 Conclusion

Our results show that resolving ambiguity in low-resource settings does not require large models or complex auxiliary tasks. Instead, performance improves through explicit input enrichment (definitions and examples) and loss-function alignment, specifically using Earth Mover's Distance (EMD) for ordinal regression. EMD provides a stable objective that captures graded plausibility without the instability of ranking losses or logical constraints.

Our findings indicate that scalar regression is insufficient for plausibility estimation, ordinal distributional losses better align predictions with human judgments, and additional pretraining or regularisation does not necessarily improve performance under ambiguity.

Future work may explore combining ordinal distributional objectives with large language models for zero-shot plausibility estimation.

Limitations

This work focuses on graded plausibility estimation for binary homonyms in short narrative contexts. While our findings generalise across several modelling paradigms, extending the approach to larger sense inventories or open-ended sense generation remains future work. Our models also rely on human-annotated plausibility distributions, enabling principled uncertainty modelling but limiting scalability across domains.

We fine-tune encoder-only architectures (DeBERTa) to maximise efficiency in low-resource settings and do not extensively compare against zero-shot or few-shot prompting of large generative models (e.g., GPT-4), leaving the trade-off between large-scale prompting and supervised precision unresolved. Although annotator disagreement is captured through ordinal rating distributions, exploring richer uncertainty representations and broader narrative contexts remains important future work.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Evaluation measures for ordinal regression. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 283–287.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

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.

Xin Geng and Quan Zhao. 2014. Label distribution learning. *CoRR*, abs/1408.6027.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. 2017. Squared earth mover's distance-based loss for training deep neural networks. *Preprint*, arXiv:1611.05916.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Ling Li and Hsuan-tien Lin. 2006. Ordinal regression by extended binary classification. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

A Appendix

A.1 Preprocessing

Text was tokenised using the standard DeBERTa-v3 tokenizer. We truncated sequences to a maximum length of 300 tokens, which was sufficient to capture the full narrative context for the vast majority of samples.

A.2 Infrastructure

All models were trained on a single NVIDIA A100 (40GB) GPU. Training a single seed took approximately 15 minutes. We used the HuggingFace `transformers` library and `accelerate` for mixed-precision training.