NCL-UoR at SemEval-2026 Task 5: Embedding-Based Methods, Fine-Tuning, and LLMs for Word Sense Plausibility Rating

Tong Wu[1], Thanet Markchom[2], and Huizhi Liang[3]
[1]Independent Researcher
[2]Department of Computer Science, University of Reading, Reading, UK
[3]School of Computing, Newcastle University, Newcastle upon Tyne, UK
tongwuwhitney@gmail.com, thanet.markchom@reading.ac.uk,
huizhi.liang@newcastle.ac.uk

Abstract

Word sense plausibility rating requires predicting the human-perceived plausibility of a given word sense on a 1–5 scale in the context of short narrative stories containing ambiguous homonyms. This paper systematically compares three approaches: (1) **embedding-based methods** pairing sentence embeddings with standard regressors, (2) **transformer fine-tuning** with parameter-efficient adaptation, and (3) **large language model (LLM) prompting** with structured reasoning and explicit decision rules. The best-performing system employs a structured prompting strategy that decomposes evaluation into narrative components (precontext, target sentence, ending) and applies explicit decision rules for rating calibration. The analysis reveals that structured prompting with decision rules outperforms both fine-tuned models and embedding-based approaches, and that prompt design matters more than model scale for this task.

1 Introduction

Word Sense Disambiguation (WSD) has traditionally been framed as selecting the single correct sense for a word in context (Navigli, 2009). Yet real-world language is often ambiguous, and multiple senses may be plausible with varying degrees of contextual support (Erk and McCarthy, 2009). SemEval-2026 Task 5 (Gehring et al., 2026) addresses this gap through the AmbiStory dataset (Gehring and Roth, 2025), which reframes WSD as a *graded plausibility rating* task over English narratives. Given a five-sentence story with an ambiguous homonym, systems must predict the human-perceived plausibility of a specific word sense on a 1–5 scale.

Three distinct modeling approaches are investigated: (1) *embedding-based methods* extracting similarity features from sentence embeddings for use with classical regressors; (2) *transformer fine-tuning* adapting pre-trained language models with LoRA for plausibility regression; and (3) *LLM prompting* using structured reasoning prompts with explicit evaluation criteria and calibration rules. The core strategy decomposes plausibility judgment into component-level evaluations of precontext, target sentence, and ending, then combines them into a final rating via decision rules.

On the test set, GPT-4o with structured prompting achieves $\rho = 0.731$ and Acc. = 0.794, outperforming fine-tuned and embedding-based methods. The code is available[1].

2 Background

Task Description. Given a five-sentence English narrative containing an ambiguous homonym, SemEval-2026 Task 5 (Gehring et al., 2026) asks systems to predict the human-perceived plausibility of a specific word sense on a 1–5 scale. The task uses the AmbiStory dataset (Gehring and Roth, 2025), a collection of short narrative stories designed to probe lexical ambiguity. Each story consists of a three-sentence precontext that establishes the narrative setting, a target sentence containing the homonym, and an ending sentence that may disambiguate it toward one of its senses. The input is the full narrative with a candidate word meaning; the output is the plausibility rating.

As a concrete example, consider the homonym *ring* with the candidate meaning "a characteristic sound." The story reads: *John looked at his savings and smiled. He had been careful with his money for months. Now, he finally felt ready to make a big decision for their anniversary. He told his girlfriend he would give her a ring. John was excited to finally buy the special piece of jewelry.* The ending confirms the *jewelry* sense, so "a characteristic sound" is implausible (gold = 1).

Each sample was rated by at least five annotators;

[1]https://github.com/tongwu17/SemEval-2026-Task5

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1930–1937

the gold label is the average rating. The dataset contains 2,280 training, 588 development, and 930 test samples (all in English).

Word Sense Disambiguation. The graded plausibility formulation departs from traditional WSD, which assumes a single correct sense per context (Navigli, 2009). Graded word sense assignment (Erk and McCarthy, 2009) acknowledges that senses exist on a plausibility continuum, motivating the regression formulation adopted in this work. This naturally motivates casting the task as graded plausibility regression.

Transformer Fine-Tuning. ELECTRA (Clark et al., 2020) introduces replaced token detection, a sample-efficient pre-training task that trains a discriminator over all input tokens rather than a masked subset. DeBERTa (He et al., 2021) improves attention with disentangled content and position representations. LoRA (Hu et al., 2022) enables parameter-efficient adaptation by injecting low-rank decomposition matrices into frozen pretrained weights.

Sentence Embeddings. For the embedding-based paradigm, Sentence-BERT (Reimers and Gurevych, 2019) derives semantically meaningful sentence embeddings via siamese networks, enabling efficient similarity computation between story contexts and word sense descriptions without task-specific fine-tuning. The resulting vectors serve as input features for downstream regressors.

LLM Prompting. Scaling language models enables few-shot task performance without fine-tuning (Brown et al., 2020). Prompt design and structured instructions play a central role in task performance (Liu et al., 2023). Recent work on LLM-as-a-judge (Zheng et al., 2023) shows that LLMs with structured evaluation criteria can approximate human judgments. This work uses several OpenAI GPT models (GPT-4o[2], GPT-4.1[3], GPT-5 mini[4], GPT-5.2[5]), Llama 3.2[6] (Grattafiori et al., 2024), and Ministral[7](Liu et al., 2026), exploring structured prompting strategies with explicit evaluation criteria and decision rules for graded plausibility prediction.

[2]gpt-4o-2024-08-06
[3]gpt-4.1-2025-04-14
[4]gpt-5-mini-2025-08-07
[5]gpt-5.2-2025-12-11
[6]llama-3.2-3B-Instruct
[7]ministral-3-8B-Instruct-2512

3 Methodology

Three approaches are investigated for predicting word sense plausibility ratings. Each takes as input a narrative story, a homonym, and a candidate word sense, and outputs a plausibility rating from 1 to 5. Figure 1 illustrates the overall system architecture.

3.1 Embedding-Based Methods

MPNet + Ridge Regression. The story and candidate meaning are encoded into sentence embeddings using `all-mpnet-base-v2` (Song et al., 2020; Reimers and Gurevych, 2019). From these, 8 features are extracted: cosine similarity, Euclidean distance, and dot product between embeddings, text length features, a binary ending indicator, and interaction terms. These are fed into a Ridge regressor (Hoerl and Kennard, 2000) with $\alpha = 1.0$.

RoBERTa + XGBoost. As an alternative configuration, 23 features are extracted using `all-roberta-large-v1`, a RoBERTa-based (Liu et al., 2019) sentence embedding model (Reimers and Gurevych, 2019): similarity features (cosine, Euclidean, Manhattan, dot product), lexical overlap (word overlap, Jaccard, character overlap), structural features (sentence/punctuation counts), and interaction terms. An XGBRegressor (Chen and Guestrin, 2016) is used with regularization and Spearman-based early stopping.

3.2 Transformer Fine-Tuning

Two families of models are fine-tuned with LoRA (Hu et al., 2022) for regression.

ELECTRA-base and ELECTRA-large + LoRA. Two ELECTRA variants (Clark et al., 2020) are fine-tuned: ELECTRA-base (110M parameters, full fine-tuning) and ELECTRA-large (335M parameters, LoRA with r=8, α=32). The input format is `[meaning] [SEP] [story]` with labels normalized from 1–5 to 0–1. For ELECTRA-large, mean pooling over all tokens (instead of only `[CLS]`) and Huber loss (δ=1.0) (Huber, 1964) are applied for robustness to annotator disagreement. Training uses batch size 32 with early stopping (patience 3) based on Spearman correlation.

DeBERTa-large + LoRA with Pairwise and Uncertainty Losses. DeBERTa-large (He et al., 2021) is fine-tuned with LoRA. The input concatenates the precontext, target sentence, and ending, separated by `[SEP]` from the word sense description, with three pooling methods considered:

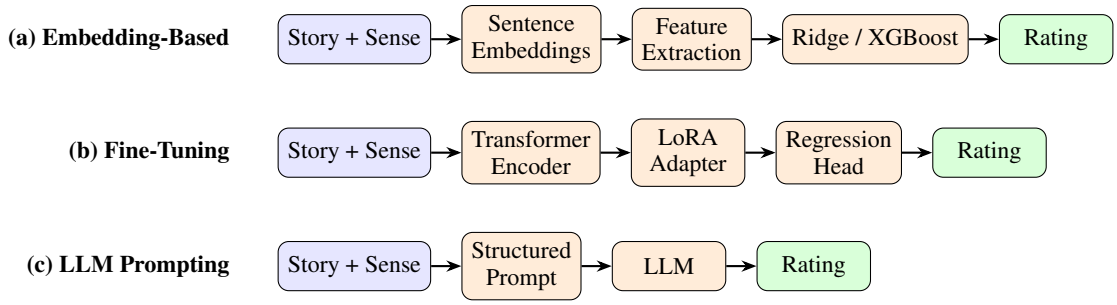


Figure 1: System overview of the three approaches. All approaches take the same input (narrative story, homonym, candidate word sense) and output a plausibility rating from 1 to 5.

[CLS] token pooling, mean pooling, and attention-based pooling. Beyond standard regression loss, two additional loss components are introduced to better optimize for the evaluation metrics:

- **RankNet pairwise loss** (Burges et al., 2005): Since Spearman correlation measures rank correlation, a pairwise ranking loss is added that encourages the model to correctly order sample pairs by plausibility. For a mini-batch, pairs (i, j) where $y_i > y_j$ are sampled, and the loss $\mathcal{L}_{\text{rank}} = -\log \sigma(\hat{y}_i - \hat{y}_j)$ is optimized.
- **Uncertainty-aware loss**: To incorporate human uncertainty in word sense plausibility ratings, the annotator standard deviation is used as a tolerance margin during training. Prediction errors within the human disagreement range incur no penalty, while errors exceeding this range are penalized linearly. Specifically, for each sample i, $\mathcal{L}_{\text{unc}} = \max(0,\ |\hat{y}_i - y_i| - \sigma_i)$, where σ_i denotes the standard deviation of annotators' scores for that instance.

The total loss is $\mathcal{L} = \mathcal{L}_{\text{reg}} + \lambda_r \mathcal{L}_{\text{rank}} + \lambda_u \mathcal{L}_{\text{unc}}$, where λ_r and λ_u are weighting hyperparameters.

3.3 LLM Prompting

Two prompting strategies are designed for LLMs.

Few-Shot Prompting (P1). The prompt consists of (1) a system message defining the task and 1–5 rating scale, emphasizing that "the ending is the most important factor for disambiguation"; (2) five few-shot examples selected from training data (one per rating level, choosing samples with zero annotator standard deviation); and (3) the user prompt with the target sample. Temperature 0 is used for deterministic output.

Structured Prompting with Decision Rules (P2). An improved prompt is designed that replaces few-shot examples with structured evaluation criteria and explicit decision rules:

1. *Component-wise evaluation*: The prompt instructs the model to separately evaluate three narrative components: precontext ("does the setup make this meaning likely?"), target sentence ("does the local usage support this meaning?"), and ending ("does the conclusion reinforce this meaning?"). The ending is identified as "the strongest source of evidence."
2. *Decision rules*: Explicit calibration rules constrain the rating: (a) "if the ending clearly contradicts the proposed meaning, the rating must be 1 or 2"; (b) "if evidence is mixed or unclear, choose the lower plausible rating"; (c) "a rating of 5 requires explicit confirmation in the ending and no contradictions elsewhere."
3. *Impartial framing*: The prompt positions the model as "an impartial evaluator" who should "base judgment only on the text given," reducing potential biases.

This strategy produces more calibrated predictions without requiring examples in the prompt.

4 Experimental Setup

Data Usage. Models are trained on the training set for development experiments. For test evaluation, fine-tuned and embedding-based models are retrained on combined train+dev data with an internal validation split (10–20%) for early stopping.

Evaluation Metrics. Two metrics are reported: (1) **Spearman correlation** (ρ) between predicted and gold ratings, measuring rank-order agreement; and (2) **accuracy**, defined as the proportion of predictions falling within one standard deviation of the mean annotator rating.

Implementation. Fine-tuning experiments used HuggingFace Transformers (Wolf et al., 2020) with LoRA via PEFT. LLM predictions were obtained via the OpenAI API for GPT models and locally via

Approach	System	ρ	Acc.
Embedding	MPNet + Ridge	0.174	0.560
	RoBERTa + XGBoost	0.114	0.537
Fine-tuning	ELECTRA-base	0.491	0.663
	ELECTRA-large + LoRA	0.644	0.709
	DeBERTa-large + LoRA	0.587	0.757
	+ uncertainty loss	0.606	0.767
Prompting	Llama-3.2-3B (P2)	0.108	0.526
	Ministral-3-8B (P2)	0.441	0.628
	GPT-5.2 (P1)	0.634	0.721
	GPT-5 mini (P2)	0.716	0.760
	GPT-5.2 (P2)	0.727	0.781
	GPT-4.1 (P2)	0.726	0.769
	GPT-4o (P2)	**0.749**	**0.818**

Table 1: Development set results. P1 = few-shot prompting; P2 = structured prompting with decision rules.

Approach	System	ρ	Acc.
Embedding	MPNet + Ridge	0.109	0.513
	RoBERTa + XGBoost	0.110	0.505
Fine-tuning	ELECTRA-base	0.482	0.625
	ELECTRA-large + LoRA	0.527	0.639
	DeBERTa-large + LoRA	0.492	0.676
	+ uncertainty loss	0.435	0.659
Prompting	Llama-3.2-3B (P2)	0.134	0.522
	Ministral-3-8B (P2)	0.472	0.594
	GPT-5.2 (P1)	0.635	0.713
	GPT-5 mini (P2)	0.696	0.743
	GPT-5.2 (P2)	0.717	0.760
	GPT-4.1 (P2)	0.722	0.767
	GPT-4o (P2)	**0.731**	**0.794**

Table 2: Test set results. P1 = few-shot prompting; P2 = structured prompting with decision rules.

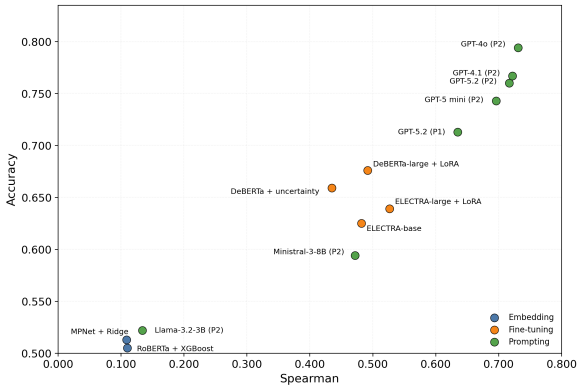


Figure 2: Test-set performance landscape (Spearman vs. Accuracy).

HuggingFace Transformers for open-source models (Llama, Ministral).

For the DeBERTa-large + LoRA system, the regression objective was evaluated using both Mean Squared Error (MSE) and Huber loss. The ranking loss weight λ_r was set to 0.25 or 0.5, while the uncertainty-aware loss weight λ_u was varied among 0.1, 0.3, and 0.5. For the LoRA component, the rank r was set to 4, 8, or 12, with $\alpha = 32$ and a dropout rate of 0.1. All other training parameters were held constant, including a batch size of 8, 10 training epochs, a learning rate of 1e-4, a warmup ratio of 0.1, and a weight decay of 0.01. Model selection was performed based on the highest average of Spearman correlation and accuracy on the development set, without cross-validation. The final model used mean pooling, MSE loss, $\lambda_r = 0.25$, $\lambda_u = 0.5$, and $r = 8$.

5 Results

5.1 Development Set Results

Table 1 presents results on the development set.

5.2 Test Set Results

Table 2 shows test set results. The best system (GPT-4o with P2) ranked 9th on the leaderboard.

Embedding features are insufficient for narrative reasoning. Both embedding-based methods achieve very low Spearman correlations ($\rho < 0.18$ on dev, < 0.14 on test), despite using rich feature sets. This suggests that handcrafted similarity features between story embeddings and meaning embeddings cannot capture *how* a narrative context supports or contradicts a specific word interpretation. The task requires compositional reasoning across multiple sentences, which static similarity metrics fail to model.

Fine-tuning captures contextual dependencies. Fine-tuned models substantially outperform embedding methods. ELECTRA-large + LoRA ($\rho = 0.644$ on dev) benefits from larger model capacity and mean pooling over all tokens. DeBERTa-large + LoRA with uncertainty loss achieves the best fine-tuning accuracy (0.767 on dev) by learning to down-weight high-disagreement samples. However, fine-tuning performance degrades on the test set ($\rho = 0.527$ for ELECTRA-large), suggesting difficulty in generalizing to unseen homonyms and story patterns. The substantial dev-to-test drop for ELECTRA-large + LoRA, from $\rho = 0.644$ to 0.527, likely reflects overfitting to the specific homonyms present in the development set: the model was retrained on combined train+dev data with only a small validation split (10–20%), limiting its ability to generalize plausibility representations to novel lexical ambiguities in the test set.

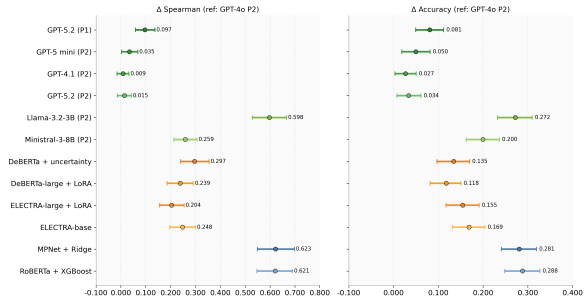


Figure 3: Paired bootstrap differences relative to GPT-4o (P2) with bootstrap confidence intervals.

Structured prompting outperforms few-shot prompting. For GPT-5.2, switching from few-shot prompting (P1: $\rho = 0.635$) to structured prompting with decision rules (P2: $\rho = 0.717$) yields a 0.082 improvement on the test set. The decision rules provide explicit calibration (e.g., "if the ending contradicts, rate 1–2"), which aligns model predictions with the annotation guidelines. This replaces the need for example memorization with principled reasoning.

Prompt design matters more than model scale. On the test set, GPT-4o with structured prompting ($\rho = 0.731$) outperforms GPT-5.2 with the same prompt ($\rho = 0.717$), suggesting that for this task, GPT-4o's reasoning capabilities are well-matched to the structured evaluation framework.

Statistical significance testing. Figure 3 shows paired bootstrap point estimates and 95% CIs relative to GPT-4o (P2). The tests (B=2000), under a paired resampling protocol with fixed random seed, show that GPT-4o is significantly better than GPT-5.2 (P1) on both Spearman ($\Delta\rho = 0.097$, 95% CI $[0.059, 0.137]$, $p < 0.001$) and accuracy (ΔAcc. $= 0.081$, $p < 0.001$). For GPT-4.1 (P2) and GPT-5.2 (P2), Spearman differences from GPT-4o are not significant ($p = 0.445$ and $p = 0.317$), though accuracy gains are small but significant ($p < 0.05$). GPT-5 mini (P2) is also significantly worse on both metrics ($p < 0.05$), as are both DeBERTa variants (p<0.001). Llama-3.2-3B (P2), Ministral-3-8B (P2), ELECTRA-large + LoRA, ELECTRA-base, MPNet + Ridge, and RoBERTa + XGBoost are worse on both metrics ($p < 0.001$). Overall, these results indicate that among GPT variants, prompt design rather than model scale is the main driver, and that the LLM prompting advantage over fine-tuning and embedding approaches is robust rather than due to chance.

Loss Configuration	ρ	Acc.
Huber + RankNet (λ_r=0.5)	0.587	0.757
MSE + RankNet + Uncertainty (λ_r=0.25, λ_u=0.5)	0.606	0.767

Table 3: Ablation of loss components for DeBERTa-large + LoRA on the development set.

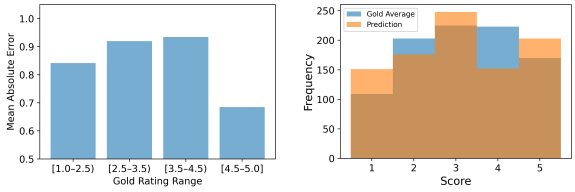
5.3 Ablation: Loss Components for DeBERTa

To better understand the contribution of each auxiliary objective to the fine-tuning behavior, Table 3 ablates the loss components for DeBERTa-large + LoRA on the development set.

Adding uncertainty loss improves both metrics on the development set, confirming that modeling annotator disagreement through learned uncertainty is beneficial. However, uncertainty loss degrades test set performance: DeBERTa-large + LoRA drops from $\rho = 0.492$ without uncertainty loss to $\rho = 0.435$ with it. The similar standard deviation distributions between development and test sets suggest this drop is unlikely to come from distribution shift and more likely reflects mild overfitting during development-set model selection, reducing generalization to unseen test examples.

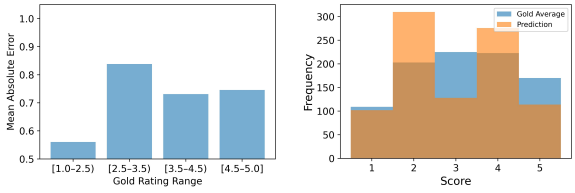
5.4 Error Analysis

High annotator disagreement increases prediction difficulty. On the test set, samples with high annotator standard deviation ($\sigma \geq 1.0$; $n = 420$) have a mean absolute error (MAE) of 0.962, compared to 0.765 for low-disagreement samples ($\sigma < 1.0$; $n = 510$).



(a) MAE by gold rating range (b) Gold vs. predicted score distributions

Figure 4: Error analysis of GPT-5.2 (P1) on the test set.



(a) MAE by gold rating range (b) Gold vs. predicted score distributions

Figure 5: Error analysis of GPT-4o (P2) on the test set.

Mid-range ratings are hardest to predict. In Figure 4, the left panel shows MAE by gold rating range. The highest MAE occurs for ratings in $[3.5, 4.5)$ (MAE = 0.934, $n = 223$), while extreme ratings near 1 or 5 are easiest (MAE ≈ 0.69). Extreme cases involve clear confirmation or contradiction by the ending, whereas mid-range ratings require nuanced judgment about partial evidence. The right panel of Figure 4 further reveals that the model's predictions cluster at integer values (1–5), while gold ratings are continuously distributed, indicating a discretization bias that limits fine-grained estimation. This stems from the prompt instruction to return a single integer. No experiments were conducted with non-integer ratings or averaging multiple stochastic API calls, as integer outputs were simpler to parse and validate under a deterministic setting. Exploring continuous-valued prompting strategies could reduce this bias.

Misleading precontexts cause catastrophic errors. Catastrophic errors typically involve homonyms where the narrative context unambiguously supports one sense, yet the model assigns high plausibility to another. For example, *shelved* in a library context led to prediction 5 for "hold back to a later time," while gold = 1.4. These failures highlight the tension between the ending-priority rule and human judgment: the model conflated meanings and rated "hold back to a later time" as 5, ignoring that both precontext and ending support only the "place on a shelf" sense. Future prompting strategies should explore more flexible weighting schemes that balance component-level evidence, accounting for sense-priming effects, rather than enforcing fixed hierarchies.

GPT-4o (P2) error analysis. Figure 5 shows the error distribution of GPT-4o (P2), which achieves an MAE of 0.702, lower than GPT-5.2 (P1) at 0.791. Gains from decision rules appear for low-plausibility ($[1.0, 2.5)$: MAE = 0.560) and mid-range ratings ($[3.5, 4.5)$: MAE = 0.731), suggesting that calibration rules help most when judgments are ambiguous. GPT-4o maintains better alignment with the continuous gold distribution, with less concentration at boundary values, indicating that structured prompting enables more calibrated estimates. This pattern suggests that explicit decomposition into precontext, target sentence, and ending mainly improves calibration rather than raw lexical matching, which is consistent with the overall advantage of prompting over embedding-based methods.

6 Conclusion

This paper presents a system for SemEval-2026 Task 5, comparing embedding-based, fine-tuning, and LLM prompting approaches for word sense plausibility rating. The key contribution is a structured prompting strategy with explicit decision rules, which achieves the best performance ($\rho = 0.731$, Acc. = 0.794 on test). The results demonstrate that: (1) embedding-based similarity features fail to capture narrative-level reasoning; (2) fine-tuned transformers with LoRA and auxiliary losses (pairwise ranking, uncertainty) improve over standard regression; and (3) structured prompting with component-wise evaluation and calibration rules outperforms both few-shot prompting and fine-tuning. Across all comparisons, the most consistent advantage comes from making the plausibility judgment more explicit: separating precontext, target sentence, and ending gives the model a simple task-specific scaffold that is more effective than relying on global similarity or larger model size alone. The error analysis also suggests that the remaining gap is less about lexical knowledge and more about calibration under mixed evidence, especially for mid-range labels and high-disagreement cases. Future work includes exploring ensemble methods that combine fine-tuned models with LLM predictors, and improving prompts to better handle conflicts.

Limitations

Structured prompting with decision rules improves calibration, but performance varies with prompt formulation and may require adaptation across models or settings. Evaluation is limited to English AmbiStory narratives and may not generalize to other domains, languages, or longer contexts. Paired bootstrap significance tests were conducted to assess statistical significance of test-set comparisons between systems. Predictions show concentration at integer points relative to the continuous gold ratings, which may limit output granularity and fine-grained calibration. The present study also evaluates only a small set of prompt variants, so some of the observed gains may depend on wording choices that were not explored systematically.

Acknowledgments

Thanks to the organizers for providing the dataset, evaluation infrastructure, and coordinating the task evaluation process.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 89–96, New York, NY, USA. Association for Computing Machinery.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of EMNLP*, 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.

Janosch Gehring and Michael Roth. 2025. AmbiStory: A challenging dataset of lexically ambiguous short stories. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 152–171, Suzhou, China. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Arthur E. Hoerl and Robert W. Kennard. 2000. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Peter J. Huber. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, and 1 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

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

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Prompt

P1: Few-Shot Prompting. The following shows the system message, an example, and the user prompt template. Four additional examples (ratings 2–5) follow the same format, selected from training samples with zero annotator standard deviation, ensuring that each in-context demonstration presents a consensus rating with no annotator disagreement.

System Prompt

You are evaluating whether a proposed meaning of a homonym is supported by its narrative context.
Input format:
- Homonym: The ambiguous word
- Meaning: The proposed interpretation
- Precontext: Background narrative
- Sentence: The sentence containing the homonym
- Ending: The conclusion (may be none)
Rating scale:
1 = Completely implausible. The meaning clearly conflicts with the narrative.
2 = Mostly implausible. Weak or contradictory support.
3 = Moderately plausible. Possible but ambiguous.
4 = Very plausible. Strong and consistent support.
5 = Highly plausible. Clearly intended and strongly confirmed.
The ending is the most important factor for disambiguation.
Return only a single integer (1–5). No explanation.

Few-Shot Example (1 of 5)

Homonym: drive | Meaning: hitting a golf ball off of a tee with a driver
Precontext: Lisa had always been competitive. Every weekend, she dedicated herself to her passion. She believed that her relentless practice would pay off someday.
Sentence: Her drive was what ultimately got her into the top university.
Ending: She made that long trip to show the course coordinators her dedication to going to that university, and they said that was one of the reasons why they accepted her.
Rating: 1

User Prompt

Homonym: {homonym}
Meaning: {judged_meaning}
Precontext: {precontext}
Sentence: {sentence}
Ending: {ending}
Rating:

P2: Structured Prompting with Decision Rules.

System Prompt (P2)

You are an impartial evaluator assessing whether a proposed meaning of a word is supported by the provided narrative context. Base your judgment only on the text given.
Word: {homonym}
Proposed meaning: {judged_meaning}
Narrative context
- Beginning (precontext): {precontext}
- Sentence containing the word: {sentence}
- Ending (conclusion): {ending}
Task
Rate the plausibility that the word {homonym} is used with the proposed meaning {judged_meaning} in this narrative.
Evaluation criteria
1. Precontext: Does the setup make this meaning likely or expected?
2. Target sentence: Does the local usage support this meaning?
3. Ending: Does the conclusion reinforce or confirm this meaning? This is the strongest source of evidence.
Decision rules
- If the ending clearly contradicts the proposed meaning, the rating must be 1 or 2.
- If evidence is mixed or unclear, choose the lower plausible rating.
- A rating of 5 requires explicit confirmation in the ending and no contradictions elsewhere.
Rating scale
1 Completely implausible: Clear contradiction.
2 Mostly implausible: Weak or conflicting evidence.
3 Moderately plausible: Possible but ambiguous.
4 Very plausible: Strong and consistent support.
5 Highly plausible: Clearly intended and explicitly confirmed.
Output format
Return only a single integer from 1 to 5. Do not include explanations, comments, or any extra text.