UAlberta at SemEval-2026 Task 5: Disambiguating Stories via Task Decomposition

David Basil, Junhyeon Cho, Chirooth Girigowda, Guoqing Luo, Sahir Momin, Sevryn Robinson, Ning Shi, Grzegorz Kondrak
Alberta Machine Intelligence Institute
Department of Computing Science
University of Alberta, Edmonton, Canada
{dbasil1,gkondrak}@ualberta.ca

Abstract

We describe our system for predicting sense plausibility in short narratives. Our approach centers on task decomposition: instead of predicting a score directly, we break the problem into simpler subtasks and combine their outputs. We further improve performance by ensembling complementary signals, including word sense disambiguation and fine-tuned embedding models. We also find empirical support for the one-homonym-per-translation principle of Hauer and Kondrak (2020a). Our best ensemble system achieves competitive performance in the official evaluation. Our code and data are available on GitHub.

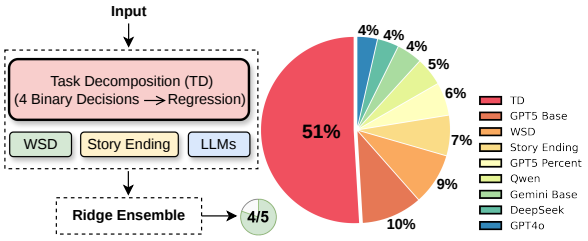

Figure 1: Overview of our system. The input story is processed separately by multiple components. TD reduces Task 5 to four binary decisions followed by regression. Its outputs, together with those from WSD, Story Ending, and direct LLM prompting, are combined using a ridge regression ensemble to produce the final plausibility score. The pie chart shows the normalized ridge coefficients as the relative contributions of each component, of which TD accounts for more than half. `GPT5 1Shot` is omitted for clarity (see text for details).

1 Introduction

This paper describes our submission to SemEval-2026 Task 5: Rating Plausibility of Word Senses in Ambiguous Sentences through Narrative Understanding (Gehring et al., 2026). The task is based on the AmbiStory dataset (Gehring and Roth, 2025). Each instance in this dataset contains a short narrative of four or five sentences. The fourth sentence contains a polysemous word ("homonym") whose meaning is ambiguous, with two candidate senses. The objective is to predict, on a 1–5 scale, how plausible it is that the homonym expresses each candidate sense in context. The task reflects the assumption that meaning in natural language may be graded rather than strictly categorical. The creators of the dataset report promising results using few-shot prompting, making approaches based on large language models (LLMs) the most widely explored avenue for this task.

This paper extends our prior work on the theoretical and empirical study of relationships among closely related semantic tasks. We have argued that certain semantic tasks can be reduced to one another under a shared framework where word meanings are represented as discrete concepts (Hauer and Kondrak, 2022). We applied such concept-based approaches in prior SemEval tasks (Hauer et al., 2020, 2022; Ogezi et al., 2023; Shi et al., 2024, 2025), most notably in Hauer et al. (2021), which leveraged translation as a signal for meaning in the word-in-context (WiC) task. We continue this investigation by examining whether word sense disambiguation (WSD) and multilingual methods can be applied to Task 5. Specifically, we posit that Task 5 is a non-binary variant of the target sense verification (TSV) task (Pilehvar and Camacho-Collados, 2019), which was shown to be theoretically equivalent to WSD by Hauer and Kondrak (2022), and thus can be reduced to WSD.

Task 5 is equivalent to TSV under an assumption known as the *sender axiom* (Hauer and Kondrak, 2020b, 2023). Specifically, we assume "that any given sentence is intended by the sender to have a single specific meaning, even if it may appear ambiguous to the receiver." Under this formulation, uncertainty is not inherent to the text, but is rather a reflection of the reader's imperfect estimation. Assume that we have a TSV oracle which has per-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2697–2708

fect access to the sender's true intent. The oracle returns TRUE if word w in context C is intended to express the sense s, and FALSE otherwise. We can then define a reduction from Task 5 to TSV as follows:

$$\text{Task5}(C, w, s) = \begin{cases} 5 & \text{if TSV}(C, w, s) \text{ is TRUE} \\ 1 & \text{otherwise} \end{cases}$$

The exclusion of intermediate scale values follows naturally from the presence of an oracle. Because the oracle determines the sender's intention with absolute certainty, the output maps strictly to one of the task's two extreme labels, i.e., *"the sense is not plausible at all given the context"* (score = 1) or *"the sense is the only plausible meaning given the context"* (score = 5). In practice, because the AmbiStory dataset features intentional ambiguity where the sender axiom may not strictly hold, our computational models instead estimate this intent, yielding continuous scores across the full 1-5 scale.

As established by the initial AmbiStory baselines, existing approaches to Task 5 and related tasks often rely on LLMs. While LLMs achieve WSD performance comparable to state-of-the-art specialized systems (Meconi et al., 2025), prior work also shows that they struggle with simple numerical reasoning tasks, including magnitude comparison (Li et al., 2025). This limitation may pose a significant challenge for Task 5, whose Spearman correlation metric emphasizes the relative magnitude of numerical outputs. Furthermore, recent work demonstrates that LLM performance on reasoning tasks is highly sensitive to question format (Song et al., 2025) and that decomposing complex tasks for LLM prompting improves generalization (Khot et al., 2023).

Motivated by these observations, we address Task 5 using an ensemble built around a task decomposition (TD) model. Rather than asking an LLM to directly predict a plausibility score, TD breaks the problem into several simpler subtasks, and combines their outputs via regression. In addition to TD, our ensemble incorporates three complementary approaches: continuous WSD, fine-tuned embeddings focused on the story ending, and targeted prompting across multiple LLMs. Together, these signals form a diverse ensemble that captures complementary indicators of the sense expressed in context.

We also investigate whether the principle of "one homonym per translation" (OHPT) can serve as a disambiguation signal in Task 5. The OHPT principle, which was proposed by Hauer and Kondrak (2020a), and justified theoretically in Habibi et al. (2021), states that a truly homonymous word token is necessarily disambiguated by its lexical translation. However, its role as a predictive signal for downstream tasks such as WSD remains underexplored.

Our results show that the ensemble achieves competitive performance, ranking second on the official leaderboard at the time of submission. Analysis shows that TD is the strongest individual component, and that combining diverse signals improves prediction quality. Additional experiments provide empirical evidence that translation can serve as a useful disambiguation cue in this setting.

2 Methods

In this section, we describe our methods for Task 5. We begin with the LLM-based approaches, first presenting our TD method, which breaks the task into simpler subtasks, followed by approaches that prompt the model to predict plausibility scores directly. We then turn to systems that do not use LLM prompting: an embedding approach, a continuous WSD method, and a translation-based method grounded in the OHPT principle.

2.1 Task Decomposition (TD)

TD attacks Task 5 by decomposing plausibility prediction into simpler binary decisions to be made by an LLM. Inspired by findings that LLMs perform better when complex reasoning is reduced to simpler subproblems (Khot et al., 2023), we employ four prompts, each eliciting a binary decision: (1) which of the two senses of the homonym is more plausible in context, (2) whether it is clear which sense is expressed, (3) whether at least one meaning fits the story, and (4) whether the given meaning is definitively correct. In addition, the zero-shot template provided by the organizers (Gehring and Roth, 2025) lets the model predict a plausibility score directly. The four binary outputs are encoded as binary features, and the plausibility response is included as a scalar feature. Together with a binary indicator specifying whether the instance contains an optional ending, these features are passed to a regression model trained on the training set, which produces a final plausibility score in the range of $[1, 5]$. The exact prompts are shown in Tables 6 and 7 in the Appendix.

2.2 Prompt Engineering

Another approach to Task 5 is to prompt LLMs directly (see Table 7). Since LLMs are sensitive to prompt design (Zhao et al., 2021), we vary the prompt format while keeping the stories and candidate senses fixed, in order to examine how different formulations affect model judgments. The official prompt provided by the organizers mirrors the human annotation procedure; the model rates the plausibility of each candidate sense on a five-point scale. We simplify the official prompt to create a baseline setting (`Base`) by removing unnecessary wording, which serves as a foundation for constructing additional variants. A more fine-grained scoring format (`Percent`) allows clearer distinctions between interpretations. Its outputs are linearly rescaled to match the five-point scale. Finally, to study in-context learning (`1Shot`), we prepend a fully labeled example to the test instance (Brown et al., 2020).

2.3 Story Ending

The Story Ending model predicts plausibility by jointly encoding the story context, candidate sense, and story ending. This is similar to the WSD encoding strategy of GlossBERT (Huang et al., 2019), where context-gloss pairs are jointly encoded to model compatibility between the context and a candidate sense. Inspired by the structure of the AmbiStory dataset, where alternative endings are intended to support different candidate senses, we seek to capture the interaction between the candidate sense and the story ending. Each input sequence concatenates the sense description, precontext, ambiguous sentence, homonym, and ending, with a special separator token before the ending, if present. The sequence is passed through an encoding model followed by a regression head to produce a scalar plausibility score. The model is fine-tuned on the training set.

2.4 Word Sense Disambiguation (WSD)

Motivated by the interpretation of Task 5 as a non-binary variant of TSV, we use continuous WSD scores as a proxy for sense plausibility. For each story, a WSD system assigns to each candidate sense a probability value, which we linearly rescale to the plausibility range $[1, 5]$ of Task 5.

2.5 One Homonym Per Translation

In order to apply the OHPT principle to Task 5, we translate the full instance into a target language, and identify the lexical translation of the ambiguous homonym using a word alignment system. If the translation is correct and its lemma appears among the possible translations in the lexical resources, OHPT produces a discrete sense decision, which can then be mapped onto the plausibility scale in various ways.

3 Experimental Setup

In this section, we present the implementation details of our methods for Task 5, and explain how they are combined into a unified system.

3.1 Ensemble

The final prediction is produced by an ensemble that combines the outputs of the systems described earlier (TD, Story Ending, WSD, and seven LLM prompting variants) to produce a single continuous plausibility score. Since some component models are trained on the official training set, to avoid data leakage, we construct and evaluate the ensemble using only the development set. We split the development set into training and validation subsets using a 70/30 ratio. To preserve the plausibility score distribution, we apply stratified sampling by dividing the gold scores into five bins. After selecting the best configuration on the validation subset, the ensemble is retrained on the full development set.

To determine the optimal ensembling strategy, we evaluated both ridge regression and XGBoost (Chen and Guestrin, 2016). We performed a grid search over the hyperparameter space of both models, measuring their performance on the validation subset. For XGBoost, we independently varied the combination of input systems, the learning rate, the number of estimators, and the group size across a set of candidate values. For ridge regression, we tuned both the combination of input systems and the regularization parameter, α.

We found that both models achieved comparable performance on the validation subset. Given these results, we selected ridge regression as our final ensemble method for its interpretability and speed. The grid search revealed that integrating all 10 systems listed in Table 1 yielded the best performance. Therefore, the optimal configuration uses ridge regression over these systems with a regularization parameter of $\alpha = 10$. Finally, the ensemble predictions are further adjusted using post-processing scaling, described and ablated in Section 4.4.

3.2 Tools and Resources

For TD, we query both GPT5 and GPT-4o from the GPT series (OpenAI et al., 2024) of LLMs. The resulting 10 outputs, together with the ending flag, are ensembled with an XGBoost regressor, using hyperparameters shown in Table 3.

For Story Ending, we fine-tune the DeBERTa encoder (He et al., 2021). Human ratings are linearly adjusted to the range $[0, 1]$ for training. Optimization combines a regression objective with a contrastive loss, where normalized scores are grouped into five bins to define positive and negative pairs. Contrastive logits are computed using cosine similarity between in-batch embeddings. Additional training details can be found in Appendix A.1.

For WSD, we use ConSeC (Barba et al., 2021) in interactive mode, which takes as input the homonym, the relevant text, and a set of lemma-definition pairs, and returns probabilities between 0 and 1 indicating the likelihood of each sense. We implement a Python interface to query the model programmatically.

For OHPT, we select Spanish as the target language, GPT-4o as the translator (the prompt is shown in Table 5), and SimAlign (Jalili Sabet et al., 2020) as the word aligner. To determine the set of possible lexical translations for each homonym, we match the dataset's sense description to its most similar gloss in BabelNet (Navigli and Ponzetto, 2012). In most cases, the string match is exact, as both AmbiStory and BabelNet use concept glosses from Princeton WordNet (Gehring and Roth, 2025). We then retrieve the translations from the corresponding target-language BabelNet synset, to which we add the lemma from the translated example sentence, if applicable.

For direct prompting baselines, we evaluate several LLMs from different model families, including GPT, DeepSeek (Guo et al., 2025), and Gemini (Team et al., 2025) through their APIs. In addition, we include the fully open-source model Qwen (Team, 2025) from HuggingFace (Wolf et al., 2020), which allows local execution and improves reproducibility. We use most systems with the organizer-defined prompt, resulting in methods we denote `GPT-4o`, `DeepSeek`, and `Qwen`. We query Gemini using the `Base` prompt, which we refer to as `Gemini Base`. Finally, we employ GPT5 with each of our variant prompts, resulting in `GPT5 Base`, `GPT5 Percent` and `GPT5 1Shot`. By varying both prompts and model families, this setup enables us to assess the robustness of our systems and to analyze how prompt variation and model diversity contribute to the ensemble. Specific versions of every LLM model mentioned in this paper can be found in Table 4.

4 Results

We evaluate our methods on the AmbiStory dataset using the official evaluation metrics. We report the performance of the ensemble system, its component models, and the baselines on the validation and test sets. We then analyze the ensemble structure, compare system variants, and present error analysis as well as OHPT-focused experiments.

4.1 Baselines

We employ three baselines. The majority and random baselines, both provided by the organizers, always predict 4 and a random integer between 1 and 5, respectively. The Spearman correlation of the majority baseline is undefined. We also define a more frequent sense (MFS) baseline: given the pair of senses associated with a story, it assigns 5 to the sense that is more frequent in Princeton WordNet (Miller et al., 1990) and 1 to the other.

4.2 Official Results

Table 1 shows the Spearman correlation results.[1] TD is clearly the strongest individual component. TD and Story Ending, both supervised systems trained on the training set, show the largest drop in performance from validation to test, suggesting possible overtuning to the development set. Overall, the ensemble improves over the best component system, achieving a Spearman correlation of 0.840 on the test set, comparable to the estimated human upper bound of 0.834 reported by Gehring and Roth (2025). These results demonstrate that combining diverse signals through ensembling is effective for Task 5.

Performance among LLM-based methods is consistently high but variable. All LLM-based approaches outperform the two non-LLM systems (WSD and Story Ending). Among the GPT-based methods, `GPT5 Base` performs best, while the `1Shot` and `Percent` prompts slightly degrade results. While Gehring and Roth (2025) found that 4-shot prompting improved performance on this task, in our experiments, 1-shot prompting slightly worsens results. Similarly, allowing scores in a wider

[1] All three evaluation metrics are reported in Table 2, with similar conclusions.

System	Val.	Test
Random	0.079	-0.014
MFS	0.142	0.098
Story Ending	0.601	0.538
WSD	0.605	0.639
DeepSeek	0.653	0.641
Qwen	0.631	0.653
Gemini Base	0.635	0.694
GPT-4o	0.743	0.724
GPT5 Percent	0.733	0.735
GPT5 1Shot	0.748	0.743
GPT5 Base	0.778	0.763
TD	0.831	0.803
Ensemble	**0.843**	**0.840**

Table 1: Spearman results on the validation and test sets.

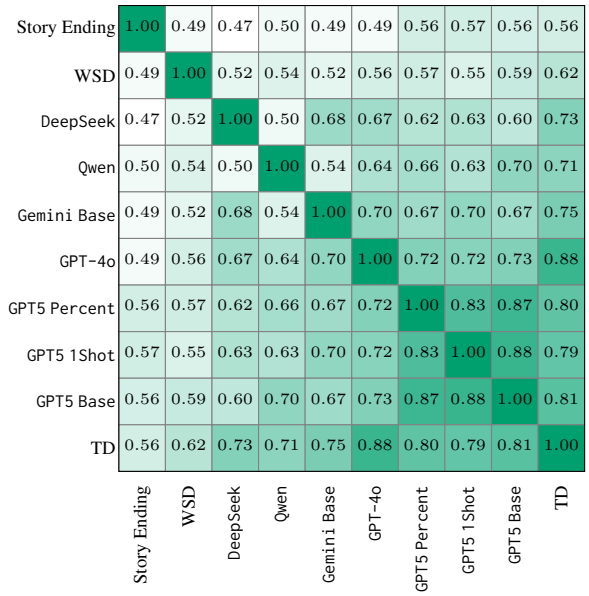

Figure 2: Correlations between outputs of component systems on the test set.

range (0–100) does not improve performance, possibly because the base prompt explicitly anchors the score interpretation ("3 = neither implausible nor plausible").

4.3 Ensemble Analysis

Figure 1 on the first page of the paper shows the ridge coefficients learned by the ensemble.[2] TD receives the largest weight, accounting for more than half of the total contribution. WSD and Story Ending are assigned higher weights than most direct LLM prompts, except for the strongest variant, GPT5 Base. This indicates that the ensemble does not simply favor the highest individual performer but also values signals that add complementary information.

Figure 2 presents the correlations between component outputs on the test set. The GPT5 variants (including TD) are highly correlated with one another, all above 0.75. Correlations between GPT and other LLM families such as Qwen and Gemini are lower, and correlations with our non-LLM systems (WSD and Story Ending) are lower still. This aligns with the ridge weights: systems that are less correlated with the dominant GPT variants tend to receive higher ensemble weights, suggesting that diversity of signals contributes to improved performance.

[2] GPT5 1Shot is excluded from the chart, as its contribution to the final ensemble is marginal at best, with a negative coefficient of -0.030.

4.4 Impact of Scaling

To better align predictions with the shared task metrics, we apply a monotonic post-processing step to the ensemble outputs. As shown in Figure 3, values below 2 are moved closer to 2 and values above 4 are moved closer to 4. The scaling strength is tuned on the validation set. Because task accuracy counts predictions within one point of the gold score, extreme values offer little benefit. Since the transformation is monotonic, it preserves the ranking of predictions and therefore does not change Spearman correlation.

We compare three variants: integer predictions obtained by clipping and rounding, raw float predictions, and scaled float predictions. The integer version achieves 0.856 accuracy and 0.813 Spearman. The float version improves to 0.919 accuracy and 0.840 Spearman. The scaled version, which is our final system, further increases accuracy to 0.925 while maintaining the same Spearman score of 0.840.

4.5 Error Analysis

Error analysis on the test set shows that many incorrect predictions occur when annotation scores for both candidate senses are relatively high. This pattern became apparent during manual inspection of 20 randomly sampled errors, defining errors as instances that fail the official Accuracy-within-Standard-Deviation criterion. In many of these cases, both candidate senses have relatively high

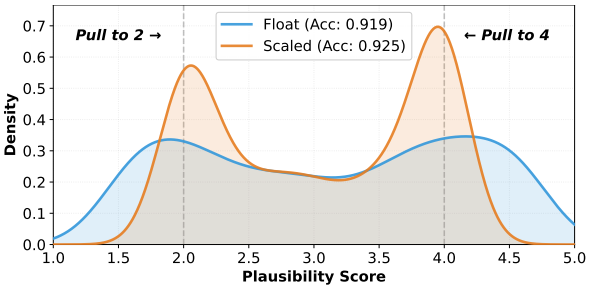

Figure 3: Distribution of float and scaled ensemble predictions.

plausibility scores, with one often appearing inflated. This effect is especially pronounced for instances where one sense is more widely used than the alternative. For example, in a narrative where riders examine a *crop*, the story concludes, "*It was an expensive brand that had clearly been bought to impress*." The mention of an expensive brand suggests the sense "*handle of a whip*"; however, the "*agricultural yield*" sense received a high plausibility score from annotators. We attribute this to the fact that when a sense is presented in isolation, as in the AmbiStory annotation process (Gehring and Roth, 2025), it may appear plausible simply because the rare alternative sense does not come to mind. When both senses are contrasted, as in our TD method, plausibility judgments may shift downward.

Motivated by this observation, we analyze all test-set errors programmatically. Error cases exhibit higher combined human ratings across both senses on average (3.35 vs. 3.13 for non-errors), consistent with our qualitative observations.

4.6 OHPT Analysis

Because OHPT produces discrete sense selections rather than continuous plausibility scores, it is not included in our main results table. Instead, we assess it using (1) ROC-AUC against human annotations, and (2) binary sense disambiguation accuracy.[3]

We compute ROC-AUC by treating human scores as a continuous ground truth, and OHPT sense selections as binary predictions. Although this reverses the typical ROC-AUC formulation, the metric remains appropriate as it evaluates ranking consistency between OHPT outputs and human annotations. We test three translation systems: GPT-4o, Qwen, and Google Translate (translation details are described in Table 5). Using these systems, OHPT achieves ROC-AUC scores of 0.679, 0.623, and 0.564 respectively, outperforming random (0.500) and MFS (0.556) baselines. GPT-4o performs best, followed by Qwen and Google Translate. This aligns with qualitative impressions of translation quality, and suggests that OHPT performance depends on translation accuracy.

We further test the hypothesis that when translation is correct and clearly disambiguates a sense, OHPT will select the contextually appropriate sense. We therefore restrict evaluation to instance pairs that can be treated as binary decisions. Specifically, we impose two conditions: (1) the in-context translation of the homonym appears in exactly one of the two expanded synsets, and (2) human annotators also prefer one example over the other, i.e., ties are excluded. With our strongest translation system (GPT-4o), 374 test instances satisfy these criteria. Treating the sense preferred by human annotators as gold, OHPT achieves 86% accuracy under this binary formulation. These results support the hypothesis that OHPT is effective for sense disambiguation.

5 Conclusion

We investigated a diverse set of approaches to Task 5. Our results show that aggregating complementary signals through a simple ridge-regression ensemble yields highly competitive performance. In particular, decomposing plausibility prediction into simpler subtasks was found to be very effective, and combining these signals with additional systems further improved results. At the time of submission, our system ranks second among the official submissions, achieving performance comparable to the reported human upper bound. Overall, our findings reinforce the value of task decomposition, signal diversity, and cross-lingual evidence for modeling graded lexical meaning.

Future work could expand the application of OHPT and the TD framework to a broader range of languages and semantic tasks. It remains to be seen how varying levels of resource availability and morphological richness affect the performance of these components. It would be interesting to evaluate them in standard WSD environments, where the contrastive structure and graded plausibility constraints of Task 5 are removed.

[3]This binary accuracy measure differs from the Accuracy-within-Standard-Deviation defined by the task organizers.

Limitations

The task structure supports contrast-based reasoning but may simplify the disambiguation problem compared to open-domain text, where the number of possible senses and contextual cues can be less clearly defined. Human plausibility ratings are inherently subjective, and variation among annotators limits the reliability of gold scores. In addition, the evaluation metrics likely reward predictions within a tolerance band, which may favor moderate scores over sharper distinctions. Finally, while our primary system was developed for English, the generalizability of these findings to more diverse languages or domains remains to be tested.

Several assumptions were required to evaluate the extension of the OHPT principle to this data. Due to the binary nature of OHPT, we utilized ROC-AUC and a restricted interpretation of accuracy rather than the official task metrics. Computing accuracy by restricting evaluation to cases with a binary preference may not be reflective of real-world contexts. Further, our investigation of OHPT was limited to Spanish as a translation language, and assumed the availability of high-quality translations and specific multilingual semantic resources.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

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*.

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.

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.

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*.

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)*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081).

Amir Ahmad Habibi, Bradley Hauer, and Grzegorz Kondrak. 2021. Homonymy and polysemy detection with multilingual information. In *Proceedings of the 11th Global Wordnet Conference*.

Bradley Hauer, Hongchang Bao, Arnob Mallik, and Grzegorz Kondrak. 2021. UAlberta at SemEval-2021 task 2: Determining sense synonymy via translations. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*.

Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, and Grzegorz Kondrak. 2020. UAlberta at SemEval-2020 task 2: Using translations to predict cross-lingual entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*.

Bradley Hauer, Seeratpal Jaura, Talgat Omarov, and Grzegorz Kondrak. 2022. UAlberta at SemEval 2022 task 2: Leveraging glosses and translations for multilingual idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.

Bradley Hauer and Grzegorz Kondrak. 2020a. One homonym per translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34.

Bradley Hauer and Grzegorz Kondrak. 2020b. Synonymy = translational equivalence. *Preprint*, arXiv:2004.13886.

Bradley Hauer and Grzegorz Kondrak. 2022. WiC = TSV = WSD: On the equivalence of three semantic tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Bradley Hauer and Grzegorz Kondrak. 2023. Taxonomy of problems in lexical semantics. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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)*.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. *Preprint*, arXiv:2210.02406.

Haoyang Li, Xuejia Chen, Zhanchao Xu, Darian Li, Nicole Hu, Fei Teng, Yiming Li, Luyu Qiu, Chen Jason Zhang, Li Qing, and Lei Chen. 2025. Exposing numeracy gaps: A benchmark to evaluate fundamental numerical abilities in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Domenico Meconi, Simone Stirpe, Federico Martelli, Leonardo Lavalle, and Roberto Navigli. 2025. Do large language models understand word senses? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: An on-line lexical database*. *International Journal of Lexicography*, 3(4).

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193.

Michael Ogezi, Bradley Hauer, Talgat Omarov, Ning Shi, and Grzegorz Kondrak. 2023. UAlberta at SemEval-2023 task 1: Context augmentation and translation for multilingual visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for 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, Volume 1 (Long and Short Papers)*.

Ning Shi, David Basil, Bradley Hauer, Noshin Nawal, Jai Riley, Daniela Teodorescu, John Zhang, and Grzegorz Kondrak. 2025. UAlberta at SemEval-2025 task 2: Prompting and ensembling for entity-aware translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Ning Shi, Senyu Li, Guoqing Luo, Amirreza Mirzaei, Ali Rafiei, Jai Riley, Hadi Sheikhi, Mahvash Siavashpour, Mohammad Tavakoli, Bradley Hauer, and Grzegorz Kondrak. 2024. UAlberta at SemEval-2024 task 1: A potpourri of methods for quantifying multilingual semantic textual relatedness and similarity. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*.

Seok Hwan Song, Mohna Chakraborty, Qi Li, and Wallapak Tavanapong. 2025. Is large language model performance on reasoning tasks impacted by different ways questions are asked? In *Findings of the Association for Computational Linguistics: ACL 2025*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*.

A Appendix

A.1 Story Ending Training

Each input to the DeBERTa embedding model that forms the backbone for the Story Ending model is formed by concatenating two segments. The first segment contains the sense description marked with "Sense:", the three-sentence precontext, the ambiguous sentence, and the homonym marked with "Homonym:". The second segment contains the candidate story ending, and may be empty. So, an input would take the format: "*Sense:* <sense> <precontext> <sentence> *Homonym:* <homonym> [SEP] <ending>". Inputs are tokenized using the associated tokenizer and passed through a `DeBERTa-v3-large` encoder. The representation corresponding to the [CLS] token is fed into a projection layer, followed by a feed-forward regression head with GELU activation and dropout, producing a scalar plausibility score. Human plausibility ratings on the 1–5 scale are linearly normalized to the interval $[0, 1]$ for training.

To reduce regression collapse toward the mean, we augment the regression objective with a supervised contrastive loss. Normalized ratings are discretized into five bins using the following thresholds:

$$\text{bin}(t) = \begin{cases} 1 & t \leq 0.29 \\ 2 & t \leq 0.49 \\ 3 & t \leq 0.69 \\ 4 & t \leq 0.89 \\ 5 & \text{otherwise} \end{cases}$$

Instances assigned to the same bin are treated as positive pairs, while instances from different bins are treated as negatives. Contrastive logits are computed using cosine similarity between in-batch embeddings. The total training objective is the (equally weighted) sum of the regression loss (formatted as binary cross-entropy) and the contrastive loss.

Model selection is performed based on development-set mean squared error. After training, Platt scaling is applied to calibrate predicted probabilities, and outputs are rescaled to the 1–5 range and rounded to obtain final discrete predictions. Training required approximately 90 seconds per epoch for a total runtime of about 45 minutes over 30 epochs on an NVIDIA Quadro RTX6000 GPU with 24 GB memory.

System	Val.			Test		
	Accuracy	Spearman	Average	Accuracy	Spearman	Average
Majority Baseline	0.576	-	-	0.558	-	-
Random Baseline	0.497	0.079	0.288	0.445	-0.014	0.215
MFS Baseline	0.350	0.142	0.246	0.360	0.098	0.229
Story Ending	0.706	0.601	0.654	0.639	0.538	0.588
WSD	0.684	0.605	0.644	0.686	0.639	0.663
DeepSeek	0.763	0.653	0.708	0.748	0.641	0.695
Qwen	0.644	0.631	0.637	0.625	0.653	0.639
Gemini Base	0.610	0.635	0.622	0.672	0.694	0.683
GPT-4o	0.734	0.743	0.739	0.712	0.724	0.718
GPT5 Percent	0.706	0.733	0.720	0.737	0.735	0.736
GPT5 1Shot	0.734	0.748	0.741	0.763	0.743	0.753
GPT5 Base	0.763	0.778	0.770	0.789	0.763	0.776
TD	0.898	0.831	0.865	0.892	0.803	0.848
Ensemble	0.932	0.843	0.888	0.925	0.840	0.882

Table 2: System performance comparison including all three metrics.

Parameter	Value
n_estimators	2000
learning_rate	0.01
max_depth	100
subsample	0.8
colsample_bytree	0.8
reg_lambda	10.0
reg_alpha	0.1
objective	reg:squaredlogerror
random_state	42
n_jobs	-1

Table 3: Hyperparameters for the XGBoost regressor used in TD.

Model	Source
DeBERTa	deberta-v3-large
DeepSeek	DeepSeek-R1
Gemini	Gemini-3-Pro
GPT-4o	gpt-4o-2024-11-20
GPT5	gpt-5-2025-08-07
Qwen	Qwen3-14B

Table 4: Pretrained models and their exact versions.

GPT-4o / Qwen Translation Prompt
You are an expert translator. Translate from {SOURCE_LANGUAGE} to {TARGET_LANGUAGE}. Provide only the translation without explanations.
Google Translate
Accessed via the googletrans Python package: https://pypi.org/project/googletrans/

Table 5: Translation methods for our experiments.

Subtask 1	**Subtask 2**
You MUST answer in exactly one line. Return ONLY: ANSWER: 1 or ANSWER: 2 Do NOT explain your reasoning. Do NOT output anything else. HOMONYM: homonym POSSIBLE MEANINGS: 1. meaning1 2. meaning2 CONTEXT: Precontext: precontext Sentence: sentence Ending: ending Which meaning (1 or 2) is more plausible?	You MUST answer in exactly one line. Return ONLY one of: ANSWER: CLEAR ANSWER: AMBIGUOUS CLEAR = One meaning is clearly more plausible. AMBIGUOUS = Multiple meanings plausible OR insufficient evidence. Do NOT choose meaning 1 or meaning 2. Do NOT explain. Do NOT output anything else. MEANING 1: meaning1 MEANING 2: meaning2 CONTEXT: Precontext: precontext Sentence: sentence Ending: ending Is this CLEAR or AMBIGUOUS?
Subtask 3	**Subtask 4**
You MUST answer in exactly one line. Return ONLY one of: ANSWER: NEITHER ANSWER: OK NEITHER = Neither meaning is plausible in the context. OK = At least one meaning is plausible (even if both could fit). Do NOT explain. Do NOT choose meaning 1 or meaning 2. Do NOT output anything else. MEANING 1: meaning1 MEANING 2: meaning2 CONTEXT: Precontext: precontext Sentence: sentence Ending: ending Does EITHER meaning fit the context?	You MUST answer in exactly one line. Return ONLY one of: ANSWER: CORRECT ANSWER: NOT Definitions: CORRECT = The judged meaning is unquestionably correct, guaranteed, fully supported by the context, and has extremely high plausibility (near 100% certainty). NOT = The judged meaning is only moderately plausible, weakly supported, context-dependent, ambiguous, or NOT guaranteed. Do NOT output anything except one of the answers above. Do NOT justify. CONTEXT: Precontext: precontext Sentence: sentence Ending: ending JUDGED MEANING: meaning Is the judged meaning GUARANTEED CORRECT?

Table 6: The prompts for each of four subtasks used in TD.

Official	You will see a short text in which one sentence is marked with "**". That sentence contains a HOMONYM, a word that can have multiple meanings. Your task: Rate how plausible the GIVEN MEANING of that homonym is IN CONTEXT. Use this scale: 1 = Not plausible at all 2 = Weak plausibility 3 = Ambiguous / both meanings similarly plausible 4 = Mostly plausible 5 = The only plausible meaning Return ONLY the number 1, 2, 3, 4, or 5. HOMONYM: {homonym} TEXT: Precontext: precontext **{sentence}** Ending: {ending} Meaning being judged: "{meaning}" Return ONLY a single digit (1–5). No words.
Base	You are an expert human annotator for word-sense plausibility. Your task is to judge how plausible a given word sense would feel to a human reader in the context of a short story. Base your judgment only on the provided story, the given sense definition, and the example. Return only the score from 1 to 5 without explanation, where: 1 = very implausible 2 = implausible 3 = neither implausible nor plausible 4 = plausible 5 = very plausible
Percent	You are an expert human annotator for word-sense plausibility. Your task is to judge how plausible a given word sense would feel to a human reader in the context of a short story. Base your judgment only on the provided story, the given sense definition, and the example. Return only the score from 0 to 100 without explanation, where: 0 = completely implausible 100 = completely plausible
1Shot	You are an expert human annotator for word-sense plausibility. Your task is to judge how plausible a given word sense would feel to a human reader in the context of a short story. Base your judgment only on the provided story, the given sense definition, and the example. Return only the score from 1 to 5 without explanation, where: 1 = very implausible 2 = implausible 3 = neither implausible nor plausible 4 = plausible 5 = very plausible Target word: "bugs" Story: "Anna was having a tough week. Her room was a mess, and her computer kept crashing. Frustrated by everything going wrong, she called Jen. She asked her friend to help her get rid of the bugs. They were crawling on the keyboard. Maybe that was the reason it didn't work." Sense Definition: "general term for any insect or similar creeping or crawling invertebrate" Sense Example: "The garden was full of bugs." Rating: 5 Sense definition: "a fault or defect in a computer program, system, or machine" Sense example: "There's a bug in the software." Rating: 1

Table 7: Prompt templates used for direct LLM plausibility rating. `Official` stands for the prompt provided by the task organizer, while `Base` provides a simplified baseline version. `Percent` employs a finer-grained 0–100 scale to examine the effect of scoring resolution. `1Shot` extends the `Base` prompt with a fully labeled example to evaluate in-context learning effects.