ChulaNLP at SemEval-2026 Task 5: Regression-Calibrated LLM for Word-Sense Scoring

Wayu Limsuwan
Department of Statistics
Faculty of Commerce and Accountancy
Chulalongkorn University
wayulimsuwan@gmail.com

Attapol T. Rutherford*
Department of Linguistics
Faculty of Arts
Chulalongkorn University
attapol.t@chula.ac.th

Abstract

Word Sense Disambiguation (WSD) is typically framed as a classification task that selects one correct sense for a word. However, real language is often less clear-cut, as a homonym may support several plausible interpretations. SemEval 2026 Task 5 addresses this limitation by introducing plausibility rating, where models estimate how likely each sense is in a narrative context, aligning predictions with graded human judgments. We use GlossBERT and BEM as encoder-based baselines and show that large language models (LLMs) produce more accurate plausibility estimates. Building on this observation, we propose a regression-calibrated LLM model that applies linear regression to adjust raw LLM outputs to better match human annotation patterns. Our calibrated model achieves the highest within-standard-deviation accuracy among our evaluated systems, demonstrating that lightweight post-hoc calibration can substantially improve LLM performance on graded semantic judgment tasks.

1 Introduction

Word Sense Disambiguation (WSD) is traditionally framed as a classification task, where a model selects a single correct sense for a word. However, real-world lexical ambiguity is often graded; multiple senses can be plausible simultaneously. SemEval 2026 Task 5 (Gehring and Roth, 2025) addresses this by introducing a plausibility-rating task in English, requiring models to estimate how likely each sense is within a short narrative context. This is crucial because it moves NLU (Natural Language Understanding) toward capturing human-like nuances in semantic interpretation rather than binary labels.

Our system, jjerd, uses Large Language Models (LLMs) to generate plausibility estimates by treating raw model outputs as features for prediction. However, as prior work indicates a substantial gap between raw LLM predictions and human plausibility judgments (Gehring and Roth, 2025), we introduce a simple post-hoc calibration step. This involves training a regression model on the LLM-predicted scores from the training set to better align our model's outputs with human rating patterns. Furthermore, while traditional BERT-based WSD systems like GlossBERT (Huang et al., 2019) and BEM (Blevins and Zettlemoyer, 2020) excel at discrete sense selection, they are not optimized for graded scores. In this work, we adapt these encoder-based approaches to a regression setting to provide a robust baseline for plausibility-based evaluation. Trained on the official training set, our submitted Regression-Calibrated LLM system ranked 7th overall on the official SemEval test set.

Our experiments show that LLM-based models achieve higher correlations with human judgments than encoder-based baselines. Moreover, regression-based calibration further reduces the gap between model predictions and human plausibility scores.

2 Related work

Most Word Sense Disambiguation (WSD) benchmarks assume that a single word sense is correct in a given context, although lexical ambiguity is often graded rather than categorical. The AmbiStory dataset, introduced for SemEval 2026 Task 5 (Gehring and Roth, 2025), addresses this limitation by framing WSD as a plausibility-rating task, where human annotators assign ordinal plausibility scores (1–5) to competing senses in short narrative contexts. While experimental results show that large language models (LLMs) achieve high correlations with human judgments and strong accuracy within standard deviation, they still fall short of

*Corresponding author

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2723–2728

the human upper bound. This highlights both the strengths of LLMs in narrative understanding and the remaining gap in modeling graded semantic plausibility.

Earlier encoder-based approaches to WSD incorporate gloss information for discrete sense selection. GlossBERT (Huang et al., 2019) formulates WSD as a sentence-pair classification task by jointly encoding context–gloss pairs with a cross-encoder, achieving strong performance on all-words WSD benchmarks. Similarly, Blevins and Zettlemoyer (2020) (Blevins and Zettlemoyer, 2020) propose a gloss-informed bi-encoder model (BEM) that independently encodes contexts and sense glosses in a shared embedding space, substantially improving performance on rare and zero-shot senses. However, both GlossBERT and BEM are designed for categorical prediction and do not explicitly model graded plausibility or ordinal human judgments. In our work, we use these models as strong encoder-based baselines and focus instead on plausibility-rated WSD, where model outputs must align with continuous and ordinal human plausibility scores.

3 Our Approach

We frame SemEval 2026 Task 5 (Gehring and Roth, 2025) as a plausibility-rating problem rather than a traditional WSD classification task. The objective is to predict a continuous score between 1 and 5 that reflects how plausible a given sense is within the narrative context of a short story. The model must therefore learn to map contextual cues to graded plausibility rather than selecting a single "correct" sense.

We evaluate large language models (LLMs) using a prompting setup directly aligned with the instruction format of Gehring and Roth (2025). For each story, the model is provided with the pre-context, the ambiguous sentence, the optional ending, and a candidate sense definition, and is prompted to output a plausibility score on a discrete 1–5 scale. Although LLMs demonstrate strong contextual understanding and produce plausibility scores that are highly correlated with human judgments, their raw outputs are not explicitly optimized to match the ordinal structure of human annotations.To address this issue, we introduce a lightweight post-hoc calibration step using linear regression. Specifically, we treat the LLM-predicted plausibility score as a scalar feature and learn a mapping from this feature to human plausibility ratings using the training set. Let $x_i \in \{1, \ldots, 5\}$ denote the raw plausibility score predicted by the LLM for instance i, and let $y_i \in R$ be the corresponding average human-annotated plausibility rating. Using pairs (x_i, y_i) from the training data, we train a linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where β_1 is the learned weight, β_0 is the intercept (bias), and ϵ_i represents the residual error.The final calibrated plausibility score for a new instance is simply the continuous value predicted by the model. The trained regressor is applied to LLM predictions on the test split without any further fine-tuning of the LLM itself. This approach treats the LLM as a feature generator and uses linear regression to correct systematic bias and rescale predictions, preserving relative plausibility ordering while improving alignment with human judgments.

4 Experiments

4.1 Dataset

The dataset used in this work is derived from AmbiStory (Gehring and Roth, 2025), a collection of 4–5 sentence narratives constructed to study lexical ambiguity.

The data is provided in JSON format and distributed across the official train, development, and test splits. Each record contains several fields: the *homonym* (the ambiguous word), the corresponding *judged_meaning* (one candidate sense definition, derived from the SemEval-2017 Task 7 pun dataset), a *precontext* generated from GPT-4o based on the target sentence, the ambiguous *sentence* written by human annotators, and an optional *ending* written by humans. Human annotators provide five plausibility ratings from 1–5 for each sense–story pair; their mean is stored in *average* and their standard deviation in *stdev*. If a story contains two possible meanings and three possible endings (two alternative endings and one omitted), it yields six annotated samples in total.

After converting the JSON files into tabular format, we obtain 2,280 training samples, 588 development samples, and 930 test samples. Note that for our encoder inputs, we use the *precontext*, the *sentence*, and the optional *ending* as the combined context, and the *judged_meaning* as the sense definition.

4.2 Baselines

GlossBERT Regression. We use GlossBERT (Huang et al., 2019) as an encoder-based baseline. The model jointly encodes the context sentence containing the homonym and its candidate sense definition using the input format `[CLS] context [SEP] gloss [SEP]`, with the target word highlighted by quotation marks. Although GlossBERT was originally designed for discrete sense classification, we adapt it for plausibility scoring by replacing the classification head with a single linear regression layer on top of the `[CLS]` representation. We fine-tune a publicly available GlossBERT checkpoint trained on SemCor 3.0 using Mean Squared Error (MSE) loss. Training is performed for four epochs with Adam, a learning rate of 2×10^{-5}, weight decay of 0.01, and a batch size of 4.

Bi-encoder Regression (BEM). We adapt the bi-encoder architecture of Blevins and Zettlemoyer (2020) as an encoder-based baseline for plausibility scoring. The context sentence containing the homonym and the corresponding sense definition are encoded independently using two `bert-base-uncased` encoders. From the context encoder, we extract the representation of the target homonym token, while from the gloss encoder we use the `[CLS]` embedding as the sense representation. The dot product between these two vectors is used as a similarity score and passed to a regression head implemented as a two-layer feed-forward network:

$$\mathrm{RegHead}(x) = \mathrm{Linear}_2(\mathrm{ReLU}(\mathrm{Linear}_1(x))).$$

The model is trained using Mean Squared Error (MSE) loss and fine-tuned for four epochs with Adam, a learning rate of 2×10^{-5}, weight decay of 0.01, and a batch size of 4.

4.3 LLM Setup

We use DeepSeek-V3.2-Exp with its default model settings and adopt the same prompting format as Gehring and Roth (2025). Where the model is instructed to output a single plausibility score from 1 to 5 The full prompt is shown in Figure 1.

Plausibility Rating Prompt.
You will see a short text in which one sentence is marked with "**". That sentence contains a word that can take on multiple meanings depending on the context. One possible meaning is given.

Your task: Rate how plausible the given meaning is in the context using one of five scores:

- **1**: Not plausible at all given the context.
- **2**: Theoretically possible, but less plausible than other meanings.
- **3**: One of multiple similarly plausible interpretations.
- **4**: The most plausible interpretation, though others may still be possible.
- **5**: The only plausible meaning given the context.

There may be no single objectively correct answer. Always consider the full context when judging plausibility.

Examples:

The bat flew out of the cave.
Meaning: *a sports implement for hitting balls*
Correct answer: **1**

The letter specified where to meet him. *So after reading it, I went to the bank.*
Meaning: *a financial institution*
Correct answer: **3**

The composer often spontaneously had ideas for new melodies. *She writes notes on a sheet of paper.*
Meaning: *a brief written record; a memo*
Correct answer: **2**

Mr. Ellis walked to the town square with a big smile. He was getting ready to paint. *Whenever he sets up his easel in the town square, he always draws a crowd.*
Meaning: *to attract; direct toward itself*
Correct answer: **5**

Task Instance:
text
In this context, how plausible is it that the meaning of the word **homonym** is **judged_meaning**?

Return **only** the number **1, 2, 3, 4, or 5**.

Figure 1: Prompt used for plausibility ratings from large language models.

For post-calibration, we first apply the same LLM prompting setup to the entire training set to obtain raw plausibility predictions. We then train a linear regression model using these LLM predictions as features to map them to the corresponding human-annotated plausibility ratings. The trained regressor is subsequently applied to LLM predictions on the test split.

4.4 Evaluation Metrics

We follow the official evaluation protocol of the shared task and report two metrics.

Model	Dev		Test	
	Spearman	Acc. w/in SD	Spearman	Acc. w/in SD
Encoder-based baselines				
GlossBERT (regression)	0.44	0.72	0.52	0.71
Bi-encoder (BEM)	0.49	0.71	0.5	0.66
Large language models				
LLM (raw)	0.69	0.71	0.72	0.77
LLM + post-calibration	**0.69**	**0.84**	**0.72**	**0.83**

Table 1: Performance comparison on the official development set and test set.

Homonym	Precontext	Sentence	Ending	Human Avg.	Pred.
proof	Jared sat at the mahogany bar, eyeing the row of bottles lined up neatly on the shelf. The bartender smiled slyly as he poured a golden liquid into a crystal glass. "This one," he said confidently, "is our finest."	The best whiskey in town; the **proof** is in the drinking.	Look, it's written on the label.	2	5
element	John had always been fascinated by chemistry. In the lab, he surrounded himself with different substances, observing their reactions. He thrived in a world of scientific exploration and discovery.	John felt comfortable in his **element** during an experiment.	Working on experiments was John's idea of a dream job, and he got to do this every day now.	1.2	4
fly	Emily entered the bustling clothing store, hoping to find a new pair of jeans. She browsed through the racks, carefully examining each item. The shop was bright and lively, with colorful displays catching her eye.	When shopping in the clothes store, she noticed the **fly** on the trousers.	It quickly flew away, but she had already lost interest because of the filth.	4.2	1

Table 2: Data examples including homonym, precontext, target sentence, ending, human average plausibility score, and model prediction.

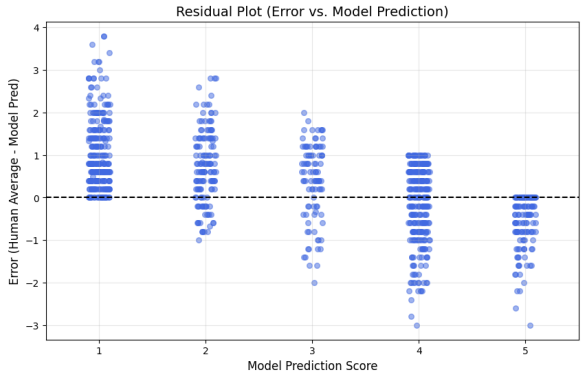

Figure 2: Residual plot of prediction errors (human average minus raw LLM score) versus raw LLM plausibility predictions.

Spearman Correlation. This metric measures rank-order agreement between model predictions and human plausibility scores.

Accuracy Within Standard Deviation. A prediction is considered correct if it falls within the interval $(\mu \pm \sigma)$, where μ is the mean of the five human plausibility ratings and σ is their standard deviation. Following the official task definition, if $\sigma < 1$, we set $\sigma = 1$ when computing this correctness range.

5 Results & Discussion

As shown in Table 1, the encoder-based baselines (GlossBERT and the bi-encoder) achieve Spearman correlations around 0.50–0.52 and accuracy within

standard deviation (Acc. w/in SD) ranging from 0.66 to 0.71 on the test set. The raw LLM substantially outperforms these baselines, achieving a Spearman correlation of 0.72 and an accuracy of 0.77.

Fitting this model on the 2,280 training samples yielded a highly significant relationship ($p < 0.001$), with an intercept (β_0) of 1.6302 and a slope (β_1) of 0.5443. The model explained 47% of the variance in human plausibility ratings ($R^2 = 0.470$, $F = 2022$) showing that the raw LLM scores are strong predictors of human judgment, but they need calibration. Applying post-calibration further improves performance, significantly increasing accuracy to 0.83. The increase in accuracy following calibration is driven by the correction of the model's boundary errors. As shown in Figure 2, the uncalibrated LLM exhibits a systematic bias at the extremes of the rating scale. Constrained by the prompt to output whole numbers (1–5), the model tends to be overconfident. For example, a raw prediction of 1 consistently underestimates the true human score, producing large positive errors that range from 0 up to 4. This occurs because a perfect ground-truth average of 1.0 is rare, requiring absolute unanimity among all five annotators. Similarly, a prediction of 5 typically overestimates the true score, resulting in negative errors extending from 0 down to -3, as natural variance in human judgment makes a flawless 5.0 nearly impossible. In contrast, when the model predicts a moderate score of 3, the errors are much more balanced and symmetrically distributed around zero (approximately ± 2), reflecting a lack of rigid boundary constraints. This discrepancy explains the effectiveness of our post-hoc calibration. The linear regression model (intercept $\beta_0 = 1.6302$, slope $\beta_1 = 0.5443$) naturally corrects this overconfidence. However, it is important to note that the Spearman correlation scores remain unchanged after the calibration because we apply the same linear transformation to all predictions. This kind of transformation only shifts and rescales the scores but does not change their order. Since Spearman correlation depends only on the ranking of the scores, not their exact values, it remains unchanged. In the official SemEval 2026 Task 5 test set, our submitted system ranked 7th overall.

The performance gap with the encoder baseline reflects the nature of the AmbiStory task, which requires understanding multi-sentence narratives rather than local context alone. BERT-based models focus on nearby words and often miss broader story-level cues, while LLMs better integrate information across the entire narrative, resulting in more accurate plausibility judgments.

Although LLMs are very good at understanding multi-sentence narratives, they can sometimes be misled by lengthy precontexts that skew the model's predictions. As shown in Table 2, in the first example for the word "proof" (judged meaning: twice the percentage of alcohol by volume), the precontext heavily emphasizes a bar and alcoholic setting. This causes the LLM to assign a maximum score of 5, missing the idiomatic usage in the shorter target sentence and ending, whereas human evaluators give it an average score of only 2. Similarly, for the word "element" (judged meaning: an indivisible chemical substance), the precontext focuses strongly on chemistry vocabulary; however, a careful reading of the target sentence and ending clearly indicates a figurative, non-chemical usage ("in his element"), yet the model still gets tricked into predicting a high score of 4. Finally, in the third example for the word "fly" (judged meaning: a two-winged insect), the precontext establishes a clothing store setting. This strong initial bias causes the LLM to assume the word refers to a zipper and assign a low score of 1, completely ignoring the crucial ending sentence which explicitly states that it "flew away". Therefore, developing prompting strategies or system modifications that force the LLM to place greater emphasis on the target sentence and ending could mitigate this priming bias and improve overall model performance.

Given the behavior of LLM outputs observed in Figure 2, more sophisticated calibration techniques beyond linear regression such as deep learning models, odd-degree polynomial regression, or other machine learning algorithms could potentially better capture the relationship between model predictions and human ratings. In our preliminary experiments, cubic polynomial regression and deep learning models did not yield significantly better results than linear regression. Future work could therefore explore alternative machine learning algorithms to further improve calibration performance.

6 Conclusion

In this work, we explored plausibility-rated word sense disambiguation using large language models. Our experiments show that LLMs produce plausibility scores that correlate well with human

judgments and outperform encoder-based baselines. The raw LLM achieved a Spearman correlation of 0.72 and accuracy within standard deviation of 0.77. Applying post-hoc regression calibration maintained the Spearman correlation at 0.72 while improving accuracy to 0.83. These results suggest that regression-calibrated LLMs provide an effective and practical approach for modeling human plausibility judgments.

References

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. 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.

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.