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

Janosch Gehring **Selina Meyer** **Michael Roth**
University of Technology Nuremberg
Natural Language Understanding Lab
{firstname.lastname}@utn.de

Abstract

We introduce SemEval-2026 Task 5 on *Rating Plausibility of Word Senses in Ambiguous Stories through Narrative Understanding*. The dataset for this task consists of 4–5 sentence English short stories. In each story, one sentence includes a lexical ambiguity and different senses are to be judged in terms of plausibility on a Likert scale. The task is intentionally constructed to be challenging because stories only providing sparse contextual cues. We give an overview of well-performing, frequent and interesting approaches used by participating systems. From a total of 175 registered participants and 27 submitted system description papers, the best system achieved an "accuracy within standard deviation" score of 93.3%.

1 Introduction

Lexical ambiguity, the phenomenon of a word having multiple word senses, is highly common in many languages and a fundamental challenge of natural language processing. As an example, the word *key* has dozens of word senses (Oxford University Press, 2025), such as in *door key*, *keyboard key*, and *piano key*, among others. Even so, humans are generally efficient at resolving these ambiguities and correctly interpreting a piece of text, especially if the surrounding context strongly implies one word sense (Rayner and Duffy, 1986).

Word Sense Disambiguation (WSD), i.e. identifying the intended sense of a polysemous word in a given context, is a notorious challenge in the field of computational linguistics. Computational models such as large language models (LLMs) need to correctly understand language in order to be usable in practical applications such as question-answering or machine translation, so their proficiency at WSD is crucial. When a model's interpretation of a polysemous expression is misaligned with human understanding, it can result in misunderstandings and reduced reliability.

Although WSD has been extensively studied in computational linguistics—particularly through popular datasets such as Word-in-Context (WiC; Pilehvar and Camacho-Collados, 2019)—these benchmarks often fail to account for the full complexity of the task. In real-world scenarios, the context may be too sparse for a single definitive interpretation, disambiguating cues may be found beyond the sentence-level, or multiple interpretations may be similarly plausible. Such nuances are typically absent from existing datasets, yet addressing them is essential in order to develop models with a robust understanding of language.

We introduce this shared task as a step towards enhancing the effectiveness of computational models in such challenging scenarios. Each sample in our dataset consists of a 4–5 sentence short story, including a polysemous target word with multiple potentially applicable senses. The narratives are constructed in such a manner that the immediate context surrounding the lexical ambiguity typically provides insufficient information for disambiguation. As the extent to which a story is resolvable is often subjective, we assess word sense plausibility by multiple annotators via a 5-point Likert scale rather than through binary labels.

In the following, we first summarize the data (§2) and describe the general task setup (§3), before discussing the results in terms of the final leaderboard as well as insights from participating systems (§4) and a qualitative follow-up analysis (§5).

2 Data

Each story consists of a four-sentence *setup* that includes a polysemous word (e.g. 'pressure') with two WordNet senses (Miller, 1994) that are compatible with the setup. For every setup, two possible *endings* provide more contextual clues, biasing the interpretation towards one of the senses. For each combination of setup, ending (or no ending) and

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3691–3703

Story Setup	The garage was bustling that Saturday morning. Dave, the tire mechanic, had a long list of cars to work on. Each customer seemed to be in a hurry, demanding quick service. The tire mechanic was feeling the **pressure**.		
Story Endings	He wanted to be able to please everyone and do a good job.	The wheel he was touching seemed to still be a bit low.	[No ending]
Human ratings for sense 'mental [...] distress'	3, 4, 5, 5, 5	2, 5, 1, 3, 3	3, 5, 5, 4, 5
Human ratings for sense 'force [...] in pascals'	1, 1, 2, 2, 2, 1	3, 3, 4, 4, 3	2, 1, 2, 2, 1

Table 1: An example for a story setup with three different endings. The annotated human ratings for each word sense with respect to each ending are listed.

WordNet sense, multiple annotators provide plausibility ratings on a 5-point Likert scale, representing the gold standard. An example from the dataset is shown in Table 1.

In the following, we summarize the data collection process for the competition dataset, *AmbiStory*. For more details, refer to Gehring and Roth (2025).

2.1 Story Collection

Stories in our dataset consist of three parts: the setup, consisting of an *ambiguous sentence* and a *precontext*, as well as an optional *ending*. Each part is collected in a separate step as described below.

Ambiguous Sentences The ambiguous sentences are the centerpieces of each story, being located between the precontext and ending and including the lexical ambiguity. We first collect a set of ambiguous words from the SemEval-2017 Task 7 pun dataset (Miller et al., 2017), which contains polysemes that were used in puns. For each word, the two word senses the pun centers on are detailed. We extract these word sense pairs and automatically filter ones where the word senses have different parts of speech or are part of longer expressions, as that limits sentence composition. In total, our final pool contains 729 word sense pairs.

During data collection on Prolific, we display a sense pair for a random word to crowdworkers and task them with writing an ambiguous sentence where the polysemous word is used such that both of these sense glosses are applicable. They may also choose to skip sense pairs.

Precontext The precontexts are the first three sentences of the story and are intended to introduce the general premise. We generate these automatically using GPT-4o (OpenAI et al., 2024) and manually check for harmful content. Given the ambiguous sentence, the model is prompted to generate a three-sentence beginning that precedes the ambiguous sentence and does not resolve the lexical ambiguity. The precontext and ambiguous sentence together form the *setup*.

Endings Each setup in our dataset is split into three stories, which differ in the ending. The first story is 'open-ended', i.e. lacking an ending sentence. The other two endings are composed by humans. For each setup, we employ two Prolific crowdworkers to write an ending given one of the applicable word senses. Their task is to end the story in such a manner that the given word sense is more plausibly the one intended by the author of the story. Since different word senses are shown to different crowdworkers, this results in distinct endings that may differently affect story understanding.

2.2 Plausibility Annotation

In the final step of data collection, annotators are given a story and one of the target word senses. Their task is to judge the plausibility of the word sense given the story on a Likert scale from 1 to 5, where 1 signifies that the sense is inconceivable given the context whereas a 5 denotes that the sense is clearly the correct interpretation, and scores in-between denote various degrees of perceived plausibility. Note that annotators never see the word's entire sense inventory, nor do they receive more than one sample based on a specific setup. As such, annotations for each sense are independent from one another and the full WordNet sense inventory. Annotators are additionally encouraged to mark noisy stories as 'nonsensical', although they are still required to rate the plausibility to the best of their ability. Each sense with respect to each story is rated by at least 5 annotators.

2.3 Final Data and Split

The dataset is split into training, validation and test set using a lexical split of roughly 60%/15%/25%, so ambiguous words in the test set are not seen during training and validation. In total, the dataset contains 19,049 annotations over 3,798 samples (i.e. unique combinations of setup, ending and word sense) and 411 distinct word sense pairs.

3 Task Description

Our SemEval task revolves around predicting plausibility of a word sense in a given story, based on human annotations with values ranging from 1 (inconceivable) to 5 (clearly correct). The task could be tackled as a classification or regression task.

3.1 Task Organization

We organized the task in two phases using the online platform Codabench. The development phase ran between October 8th, 2025, and January 10th, 2026. During this phase, the training and validation set were shared with participants, including gold labels. Participants could freely experiment with the data and upload their scores on the validation set to the leaderboard. Participants were also provided with a starting kit including a sample prediction file, scoring script, and format checker.

The evaluation phase ran from January 10th to February 3rd, 2026. At the start of this phase, we released the test set without gold labels. Participants were limited to one submission per day. During the evaluation phase, participants could only see the scores of their own submissions, and could choose a submission to add to the leaderboard.

The leaderboards can be viewed on the competition page.[1] New submissions can still be added to the *post-evaluation phase* leaderboard.

3.2 Evaluation Metrics

The shared task uses two primary evaluation metrics: *Spearman's rank correlation coefficient* and *Accuracy within Standard Deviation*. The leaderboard is sorted by the average of both metrics, so participants are required to optimize both equally. The two metrics are chosen to optimize for directional trends in plausibility ratings while taking into account human label variation.

[1] https://www.codabench.org/competitions/10877/#/results-tab

Rank Correlation Coefficient We use Spearman's ρ (Spearman, 1904) to measure the correlation between the averaged human labels and the model predictions.

Accuracy within Standard Deviation This metric measures the relative proportion of model predictions that are within the standard deviation (or at least 1) of the human-annotated mean. The motivation for this measure is to enable samples with higher standard deviation to also have a larger range of "correct" prediction values.

3.3 Baselines

We include implementations for two simple baselines in the code for our competition: *Majority*, which always picks the most common integer prediction (on our full dataset: 4), and *Random*, which selects an integer between 1 and 5 randomly.

In addition, we report zero-shot baselines based on GPT-4o and Llama-3.1 8B to serve as an indicator of out-of-the-box capabilities of modern LLMs on this task. While most systems are generally expected to beat the random and majority baseline, the LLM-based baselines are harder to outperform. More details on these baselines, including hyperparameter settings and prompts, are discussed by Gehring and Roth (2025).

3.4 Human Performance

We approximate human performance by calculating the score each human would obtain on the test set when taking the rest of the annotations as gold data. We then average the scores of the best-performing human in each group of annotators receiving the same samples. The human performance estimated this way lies at an Accuracy of 0.892 and Spearman ρ of 0.834, resulting in an average score of 0.864.

4 Participant Systems and Results

4.1 Official Ranking

In total, the 175 participants who registered for our task on Codabench made 845 submissions. 78 participants submitted a score to the evaluation phase leaderboard. Of those, 27 teams submitted system description papers. The leaderboard for the evaluation phase is shown in Table 2.

4.2 Systems Overview

We provide an overview of common approaches to the shared task, and highlight particularly successful or interesting systems.

#	Team Name	Acc. w/in SD (%)	ρ	Avg.
1	SRCB	93.3	.856	.895
2	UAlberta	92.5	.840	.882
3	Team p1	92.4	.838	.881
	Human Performance			.864
4	COGNAC	88.4	.835	.859
5	Sabanci_group4	90.0	.805	.853
6	CiNet_Handai_Kyodai	90.1	.792	.847
7	ChulaNLP	82.9	.719	.774
8	JCT 2026	82.0	.712	.766
9	NCL-UoR	79.4	.731	.762
10	SwanNLP	79.7	.723	.760
	GPT-4o Baseline			.756
11	CUCLASIC	76.8	.727	.747
12	ConText	77.6	.698	.736
13	Guys_LLM	78.8	.679	.734
14	SU NLP 29	78.4	.682	.733
15	VerbaNexAI	75.9	.673	.716
16	Habib Disambiguators	74.1	.562	.652
17	SemTechLab	70.0	.576	.638
18	YNU-HPCC	67.6	.583	.630
19	SVNIT_CSE_AI	68.2	.546	.614
20	PuerAI	68.8	.533	.611
21	Ambirig	66.6	.490	.578
22	blue	64.2	.508	.575
	Llama-3.1 8B Baseline			.563
23	AI4PC-Howard Univ.	60.3	.519	.561
24	ZCY	56.8	.202	.385
25	Narrative Team	54.2	.169	.356
26	UWB-NLP	54.5	.132	.338
	Majority Baseline			.279
27	Paradise	54.2	-.038	.252
	Random Baseline			.227

Table 2: Final leaderboard for the task's evaluation phase, sorted by the average of Acc. w/in SD and ρ.

4.2.1 Generative LLM-based methods

The best-performing systems on our benchmark relied on the usage of generative (decoder-based) LLMs. In total, 14 teams incorporated such LLMs into their best system. Commonly used generative LLMs include GPT-5 (Singh et al., 2025), GPT-4o (OpenAI et al., 2024), Qwen3 (Yang et al., 2025), DeepSeekV3.2 (DeepSeek-AI et al., 2025), Gemini-2.5 (Comanici et al., 2025), and Llama-3 (Grattafiori et al., 2024), among others.

The teams using these LLMs explored a broad range of prompting strategies. For instance, CUCLASIC (Riba et al., 2026) designed 0-, 3- and 5-shot prompts, UAlberta (Basil et al., 2026) decomposed the task into a set of binary questions and prompted the LLM to answer each separately, and several systems employed chain-of-thought (CoT) prompting.

4.2.2 Encoder-only based approaches

The best approach by 12 teams is based purely on encoder-only language models (single models or combinations). The most commonly used encoder-only models across all systems are DeBERTa (He et al., 2023) and RoBERTa (Liu et al., 2019), which are used by 10 and 3 teams, respectively.

4.2.3 Traditional approaches

Two systems used only traditional machine learning and no transformer-based methods, ranking last on the leaderboard. Features included similarity and text complexity, as well as annotation variance and presence of the ending of a story.

4.2.4 Hybrid / Ensembling approaches

Many teams decided to combine different base models: 3 teams used hybrid systems, ensembling both generative LLMs and encoder-only language models. Ensembling was also used in 6 of the 10 purely generative LLM-based systems and in both of the traditional machine learning systems. Only 1 of the 12 purely encoder-only approaches used ensembling in their best-performing system. In their system descriptions, some teams point out that model ensembling is an effective method of simulating the multi-annotator setup used for collecting the dataset. For instance, the team COGNAC (Islam and Erana, 2026) has found that LLM ensembling greatly enhances the alignment to the gold data: Even ensembles of comparatively small models were shown to rival the best individual model performances.

4.2.5 Performance by Model Size

Figure 1 illustrates the relationship between leaderboard score and size of the largest model component used for prediction reported by each system. While generative LLM-based approaches tend to outperform encoder-only models, leaderboard rank does not necessarily improve with increasing model size. Notably, very large models (e.g. GPT-5) do not consistently beat smaller LLMs such as Qwen3-8B or Qwen3-14B, which in some cases achieve

Team Name	System Type	Model Components	Additional Keywords
SRCB	Gen.	Qwen3-14B/30B	Multi-Target, Data Augmentation
UAlberta	Hybrid	GPT-4o/5.2, DeepSeek-V3.2, ...	Task Decomposition, Ensemble
Team p1	Gen.	Qwen2.5-7B, DeepSeek-Chat-V3	CoT, Translation
COGNAC	Gen.	GPT-5-nano, ...	Comparative Prompting, Ensemble
Sabanci_group4	Gen.	Qwen3-8B, DeepSeek-V3.2	Student-Teacher
CiNet_Handai_Kyodai	Hybrid	GPT-4.1, RoBERTa-base, ...	Prompting, Gaze Features
ChulaNLP	Gen.	DeepSeek-V3.2-Exp	Prompting, Calibration
JCT 2026	Hybrid	Llama-3-8B, DeBERTa-v3-large	LoRA, NLI
NCL-UoR	Gen.	GPT-4o	Prompting
SwanNLP	Gen.	GPT-4o, DeepSeek-V3, ...	Prompting, Ensemble
CUCLASIC	Gen.	Gemini-3-pro, Gemini-3-flash, ...	Prompting
ConText	Enc.	DeBERTa-xLarge	Curriculum Training, Data Augmentation
Guys_LLM	Enc.	DeBERTa-v3-large	NLI
SU NLP 29	Enc.	DeBERTa-v3-large	LoRA
VerbaNexAI	Gen.	Llama-3.1-70B, Qwen-2.5 32B, ...	CoT, Ensemble
Habib Disambiguators	Enc.	DeBERTa-v3-base	Distribution Prediction
SemTechLab	Enc.	nli-mpnet-base-v2	Span-Specific Features
YNU-HPCC	Enc.	DeBERTa-v3-large	MSE Regression
SVNIT_CSE_AI	Enc.	DeBERTa-v3-large	Ordinal Regression
PuerAI	Enc.	DeBERTa-v3-large	Contrastive Regression
Ambirig	Enc.	DeBERTa-v3-base	GlossBERT-style, EMD loss
blue	Enc.	BERT-base	Cross-Encoder
AI4PC-Howard Univ.	Gen.	Llama-3.1-70B, ...	Prompting, Calibration
ZCY	Enc.	SBERT, MPNet, ...	Semantic Similarity, Classification
Narrative Team	Enc.	DistilBERT	Fine-Tuning
UWB-NLP	Trad. ML	Several Features	Semantic Similarity, Page-Rank
Paradise	Trad. ML	Several Features	Feature Engineering

Table 3: Summary of the submitted systems in their ranked order on the leaderboard. The additional keywords are hand-picked from keywords provided by each team to best represent their highest-scoring system.

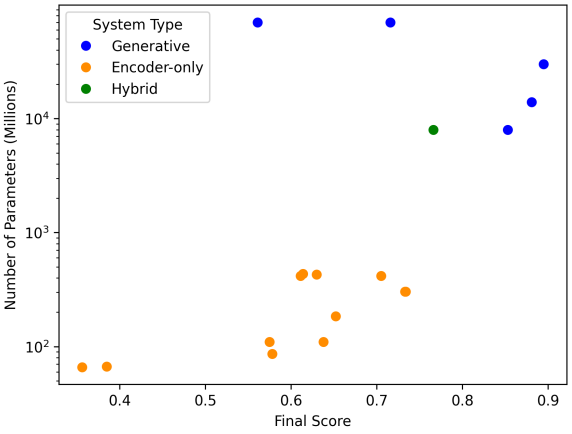

Figure 1: A comparison of best-performing model sizes according to the descriptions provided by the participating teams, sorted by final score. Systems involving models with unknown sizes (such as GPT models) and traditional feature-based methods are not displayed.

competitive or superior results. In fact, the highest-ranking system by SRCB (Zhang et al., 2026) reportedly performs similarly for both Qwen3-14B and Qwen3-30B. This indicates that the task cannot be solved through brute-force parameter scaling, but instead requires targeted strategic approaches.

4.2.6 Usage of Training Data

In total, 23 out of 27 teams relied on our training data to some extent in their best-performing system. The systems not using it are COGNAC, NCL-UoR (Wu et al., 2026), CUCLASIC, and AI4PC-Howard University (Asare and Aryal, 2026), all of which rely on prompting-based generative models. Notably, COGNAC's ensembling method achieves particularly high scores despite not fine-tuning on our task. Still, several systems that did fine-tune on our training data are able to achieve comparable performance despite being much smaller than those systems, suggesting that fine-tuning on our dataset is practical for this task.

4.2.7 Usage of External Resources

This competition permitted the use of any external resources. Still, 17 out of 27 teams elected to tackle the task without incorporating external data (aside from knowledge stored in model-intrinsic parameters). Interestingly, although the dataset itself is based on WordNet glosses, only 4 teams reported leveraging WordNet as an external resource. For instance, UWB-NLP (Taylor, 2026) used a PageRank algorithm (Page et al., 1999) on the WordNet graph to find sets of words biasing a story towards a particular word sense.

Of the 10 teams that used external data, 6 enriched the samples in the dataset with additional information using pretrained language models (see Section 4.2.8 for examples). Only one augmented the training set with additional samples: ConText (Faisal et al., 2026) included samples from WiC and a Graded Word Sense Disambiguation dataset (Cassotti and Tahmasebi, 2025), aligning them with the AmbiStory data format using Gemini 2.5 Flash. They further augmented the data using techniques such as back-translation and synonym replacement. Their approach outscores all other encoder-only models.

4.2.8 Other Highlights

Table 3 provides a concise overview of submitted system types, model components and approaches. Some additional highlights include: Team p1 (Rajpoot, 2026) used DeepSeek-Chat-V3 to generate features in the form of story translations to Hindi and Korean. As many polysemous words lose their ambiguity when translated to another language, these translations provided beneficial signals that were successfully incorporated into their ensemble model.

Team CiNet_Handai_Kyodai (Pereira and Cheng, 2026) incorporated eye-tracking into their approach by generating synthetic gaze features on the provided dataset. They simulate the total fixation time on the target polyseme, using it as an indicator of the word's ambiguity. The intuition here is that ambiguously used words would take readers longer to process, so conversely, the processing time of readers may be indicative of a word's degree of ambiguity. Indeed, they report a light performance increase on the test set when including total fixation time in their ensemble of LLM-based methods.

Two teams approached the task as a Natural Language Inference (NLI) problem. In the system by the Guys_LLM team (Antonelli-Dziri et al., 2026), the story context is used as the premise and the sense definition as hypothesis. Their approach is the highest-scoring system that does not use multi-billion parameter LLMs in any capacity, indicating the potential of NLI systems in parsing context for WSD tasks. Team JCT 2026 (Laufer et al., 2026) developed a DeBERTa-v3-large based NLI model, achieving similar results to the Guys_LLM team on the development set. They also combined this NLI model with a LoRA-finetuned Llama-3 model in

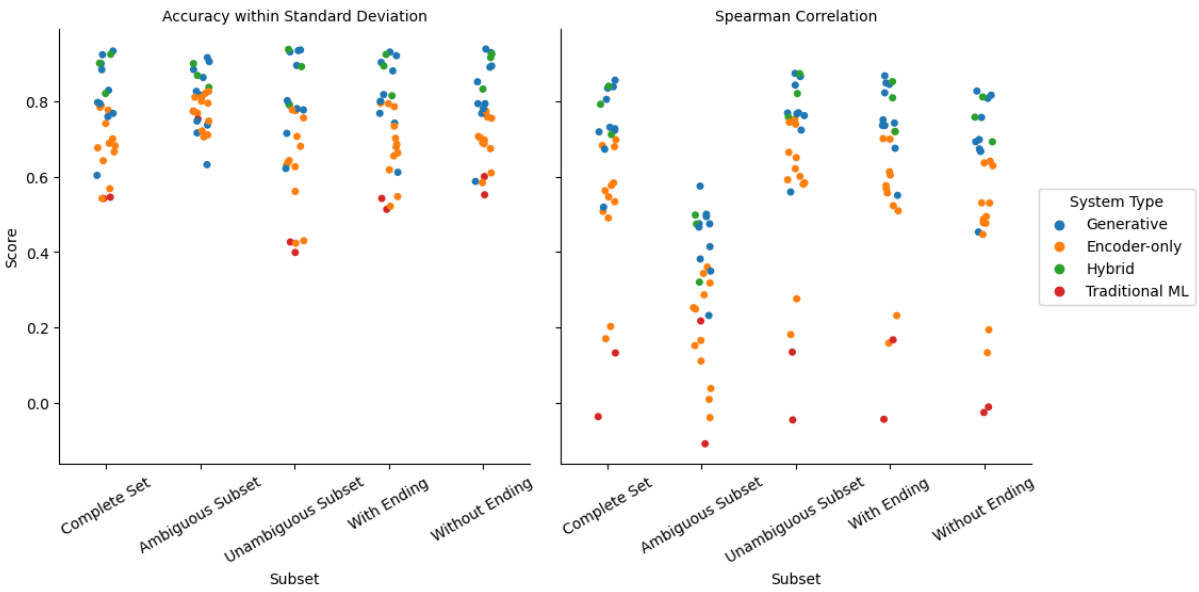

Figure 2: Accuracy within Standard Deviation and Spearman's ρ on different subsets of the test data.

an ensemble, enhancing the system's performance even further.

Finally, team SRCB enriched the training samples by adding two types of LLM-generated explanations for their scores. One explanation is comparative: It groups samples with the same setup and word sense but different endings, and explains why those endings lead to different plausibility scores. The other explanation provides a rationale for each sample's score, justifying the gold plausibility score. Combined with other strategies, such as multiple training objectives, their system achieves 1st place in our competition, achieving both the best Spearman ρ and Accuracy within Standard Deviation.

4.3 Additional Analysis

To identify strengths and shortcomings of different system types and individual approaches, we analyze system performances on specific subsets of the test data. With these analyses, we provide insights into how the systems handle distinct types of ambiguity stemming from underspecification and highly ambiguous word senses.

4.3.1 Performance on Open-Ended Stories

As described in Section 2.1, each setup in the dataset has two different endings that are intended to increase the plausibility of one word sense. In addition, each setup also has a variant without an ending, leaving the intended word sense more open. In our previous work (Gehring and Roth, 2025), we found that LLMs' zero- and few-shot performance on open-ended stories is generally lower than on stories with endings. Thus, we are interested in which participating systems are more robust to this than out-of-the-box LLMs and whether there is a trend in performance across system types.

We find that 23 out of the 27 systems perform worse on open-ended stories than on stories with an ending, although the difference is comparatively small (average scores decreasing between 0.01 and 0.08), However, we also identify individual systems that perform better on open-ended stories: VerbaNexAI (Gnecco et al., 2026), blue (Singhal et al., 2026), Narrative Team (Istrate et al., 2026), and Paradise (Goyal et al., 2026) all achieve higher scores on open-ended stories than on ended stories, with modest score increases ranging between 0.01 and 0.05. Out of these systems, VerbaNexAI's system is the only one with a generative component, whereas the others fully rely on encoder-only or traditional machine learning architectures. This reinforces our previous findings that open-ended stories pose difficulties for LLM-based systems, even those that exhibit a high overall performance. VerbaNexAI's ensemble-based approach specifically accounts for higher inter-model disagreement, which they observe for stories with no or very short endings, by employing selective self-consistency: For such data points, the system samples three predictions per model at higher temperature settings than for points with low inter-model disagreement, which are then averaged before ensembling.

ID	Story	Word Sense	Avg. Rating	✓
20	**Precontext:** *Jack had been waiting for this day for months. The exam was finally over, and all he could do now was wait. [...]* **Sentence:** *One could see the **sweat** on his face as he waited for the results.* **Ending:** *He need not have worried; he passed with an A.*	salty fluid secreted by sweat glands	$5.0^{\pm 0.0}$	10
21		agitation resulting from active worry	$3.2^{\pm 1.1}$	20
696	**Precontext:** *The family gathered in the park for the town's annual celebration. Everyone was excited for the evening's main event. As the sky darkened, the anticipation grew among the crowd.* **Sentence:** *The kids got a real **bang** out of the fireworks show.* **Ending:** *They couldn't have been more excited and full of energy from the show.*	the swift release of a store of affective force	$1.4^{\pm 0.5}$	0
697		a sudden very loud noise	$2.0^{\pm 0.7}$	13
800	**Precontext:** *Max received a strange package in the mail. Inside, he found a peculiar-looking fruit and a mysterious letter. The letter claimed the fruit held extraordinary powers.* **Sentence:** *It was not easy to **swallow**.* **Ending:** *Max did not believe the note with the fruit.*	pass through the esophagus as part of eating or drinking	$1.8^{\pm 0.8}$	8
801		believe or accept without questioning or challenge	$3.4^{\pm 1.3}$	26

Table 4: Examples for stories and word senses along with average plausibility ratings and number of systems (denoted by ✓) that correctly predicted plausibility within standard deviation.

4.3.2 Performance on Ambiguous Stories

For each story (and polysemous word) in our dataset, we collected annotations for two distinct word senses. In some cases, both word senses were perceived as highly plausible by annotators. Such cases differ from samples where both word senses predominantly received middle scores; picking high scores for both indicates that annotators were having difficulties perceiving another plausible sense for the given polyseme. This high level of certainty about the correctness of a given word sense in the story context is often indicative of a high similarity between the two word senses (such as literal and metaphorical senses portraying similar concepts—see the samples with ID 20 and 21 in Table 4 for an example). We expect this to pose a challenge to participating systems. To explore this, we compare system performance on the subset of stories in the test data for which both word senses received an average plausibility rating higher than 3 (190 samples) with performance on the remaining data.

On this ambiguous subset of the data, 23 systems perform worse, with decreases in performance compared to the rest of the data being more severe than in the case of open-ended stories: Among the systems with lower performance, the mean decrease in average score is $0.18^{\pm 0.06}$, with one system's performance decreasing by 0.29. The four systems with *higher* performance on the ambiguous subset are Paradise and Narrative Team, who also saw an increase on open-ended stories, as well as UWB-NLP and ZCY (Zhou et al., 2026). None of these systems include a generative component.

Examining the metrics separately (see Figure 2), it becomes apparent that the drop in performance stems from considerable decreases in Spearman correlation compared to unambiguous samples for most systems and across all approaches, with the exception of traditional ML, where one system (UWB-NLP) sees an increase in correlation by 0.08. Decreases in correlation range between 0.07 (Paradise) and 0.5 (blue), with a median of 0.32. On the other hand, accuracy generally increases for ambiguous samples, presumably because lower plausibility scores are less common.

4.3.3 Prediction Value Distributions of Systems Compared to Ground Truth

Figure 3 shows the range of predicted values by each system on the test set, compared with the ground truth. While the ground truth includes all values between 1 and 5 and is relatively evenly distributed across the scale, many of the systems predicted a much narrower range. This is especially apparent for the two traditional machine learning based systems submitted by UWB-NLP and Paradise, but can also be observed in many encoder-only, generative and hybrid systems. Interestingly, many of the top-performing systems did not predict the full range of possible values. This is likely a result of optimizing the 'Accuracy within Standard Deviation' metric, as avoiding extreme values increases the probability that predictions fall within the standard deviation range.

Overall, encoder-only-based systems show a ten-

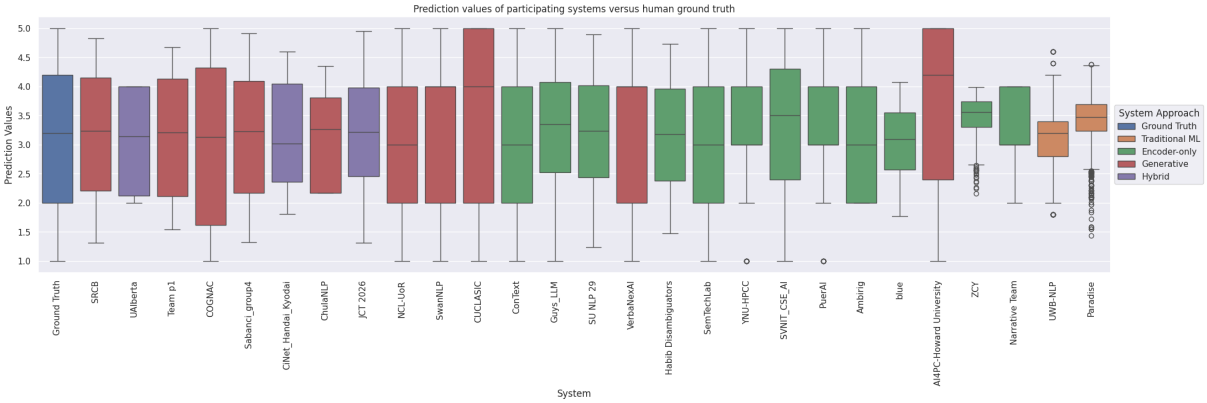

Figure 3: Range of predicted values by individual systems on test data, compared to average plausibility scores given by human annotators.

Gold Label (Plausibility)	Average Prediction	Average Difference	Correct Systems
$1.0 - 1.9$ (Inconceivable)	$2.37^{\pm 0.6}$	$0.95^{\pm 0.7}$	$16.55^{\pm 5.4}$
$2.0 - 2.9$ (Rather implausible)	$2.92^{\pm 1.3}$	$0.79^{\pm 0.6}$	$21.36^{\pm 5.5}$
$3.0 - 3.9$ (Rather plausible)	$3.32^{\pm 1.3}$	$0.65^{\pm 0.5}$	$\mathbf{23.73}^{\pm 3.5}$
$4.0 - 5.0$ (Clearly correct)	$3.70^{\pm 0.7}$	$0.87^{\pm 0.6}$	$19.19^{\pm 5.5}$

Table 5: Average predictions of systems for different ranges of plausibility ratings in the annotated test data, along with mean absolute distances between system prediction and gold label, and average number of systems which correctly predicted plausibility within standard deviation. Highlighted in boldface is the average share of correct systems (23.73 out of 27) at the plausibility range that is predicted best ($3.0 - 3.9$).

dency to predict mostly middle values, whereas some generative systems (such as CUCLASIC and AI4PC-Howard University) also predict extreme values, with a general bias towards higher values. This is also indicated by prediction statistics per average plausibility rating, as shown in Table 5.

For values lower than two, the average distance between system-predicted values and gold labels is highest: on average, only 16.55 out of 27 systems predict values within one standard deviation of the human average score, with the rest predicting plausibility scores that are too high. The second-highest difference between system-predicted and gold label values concerns the other extreme (i.e. scores higher than or equal to 4): Here, slightly more systems predicted a result within standard deviation (on average, 19.19 out of 27), but plausibility is underestimated by other systems. This finding is especially interesting in the context of generative models: One might assume that LLMs encode enough background knowledge to disambiguate clear cases without problems. However, the relatively low performance of most systems on such samples indicates that knowledge learned from an LLM's training data alone is often insufficient to predict when a word sense is clearly plausible or inconceivable. We outline potential reasons for this below, drawing on qualitative insights.

5 Qualitative Insights

Through a qualitative analysis of the results of participating systems, we find the following aspects to be particularly challenging.

Distinction of Word Senses As already discussed in Gehring and Roth (2025), humans' distinction between word senses does not necessarily align with that of models. In Table 4, the story with IDs 696/697 shows the word 'bang' being used as part of the phrase *'to get a bang out of something'*. Although the intended reading may appear obvious, humans consistently scored the word sense 'the swift release of a store of affective force' low, presumably due to interpreting the gloss more literally. In spite of this, all participating systems scored this word sense too highly, and WordNet itself also categorizes this usage as an instance of this sense. We theorize that models, particularly those with access to sense inventories such as WordNet during pretraining, are influenced by those resources and their predefined word sense inventories, while most humans are not—thus interpreting glosses

differently.

Reasoning By design, our dataset contains stories with sparse contextual cues and often even superficially contradictory information. For instance, the usage of the word 'bang' in the previously discussed example (ID 696/697) suggests that the bang is not part of fireworks, even though the story revolves around fireworks. In such cases, models must rely on reasoning capabilities to understand the narratives on a deeper level. In the story with IDs 800/801 (see Table 4), '*not easy to swallow*' may refer to either a fruit or a letter, both of which play an equally important role in the story's plot. Perhaps due to a lack of superficial evidence in favor of either word sense, many systems struggle predicting plausibility scores similar to humans.

6 Conclusion

We presented the data, participating systems and results of our shared task on ambiguous short stories and summarized the results of 27 participants.

Most systems used generative LLMs or BERT-based encoder-only models for their systems, with generative LLMs in particular emerging as victorious on the leaderboard. Indeed, multiple systems even exceeded human performance. All of the systems high up in the ranking used a combination of LLMs and other elaborate techniques, such as ensembles, data augmentation and machine translation. Many other systems relied mostly on BERT-based models and other comparatively inexpensive features. While those systems are for the most part outperformed by the generative LLM approaches in this competition, many are competitive with out-of-the-box LLMs (as seen with our baselines and Gehring and Roth, 2025) despite being much less costly. Some of the best-performing systems in this category are those relying on NLI or fine-tuning for regression.

The results show that complex systems fine-tuned on task-specific data are generally effective at disambiguating in this setting. Especially for LLM-based approaches, it is possible that the WordNet senses used in the dataset are encoded into the LLMs' training data in some way. Future work may explore the extent to which such previous knowledge yields advantages in LLM-based plausibility rating. In the future, we also aim to improve the task dataset, especially regarding its restrictions in size, language and variation.

Limitations

Our dataset contains some notable limitations that we plan to address in future work. First, all stories follow a fixed 4–5 sentence structure, where the ambiguous word is always in the fourth sentence and the disambiguating information, when present, usually appears in the fifth sentence. This predictable structure may allow systems to exploit positional cues that are unlikely to generalize to more naturalistic real-world text. Furthermore, as parts of our stories are composed or reviewed with the assistance of generative models, the dataset may reflect linguistic biases often found in AI-generated text. Finally, the dataset is limited to English and the stories were written and annotated predominantly by native speakers living in the UK. Consequently, the plausibility judgments may not fully represent how English speakers from other regions or non-native English speakers would interpret the same stories.

References

Niccoló Antonelli-Dziri, Sixtine Marcotte, Emanuele Rosapepe, Gabriele Santona, Omar Wafaay, Lorenzo Vaiani, Riccardo Coppola, and Flavio Giobergia. 2026. Guys_LLM at SemEval-2026 Task 5: Nli-informed regression for graded word-sense plausibility in narrative contexts. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Kwaku Asare and Saurav K. Aryal. 2026. AI4PC-Howard University at SemEval-2026 Task 5: Calibrated hybrid ensembling and retrieval-augmented LLM reasoning for narrative word-sense plausibility. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Zohaib Aslam, Ahsan Siddiqui, and Ayesha Enayet. 2026. Team Habib Disambiguators at SemEval-2026 Task 5: Assessing semantic plausibility using regularized transformer fine-tuning. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Karlo Babić, Ana Meštrović, and Slobodan Beliga. 2026. SemTechLab at SemEval-2026 Task 5: Context-aware homonym disambiguation via span-specific interaction features. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Mingyu Bai, Jin Wang, and Xuejie Zhang. 2026. YNU-HPCC at 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.

David Basil, Junhyeon Cho, Chirooth Girigowda, Guoqing Luo, Sahir Momin, Sevryn Robinson, Ning Shi, and Grzegorz Kondrak. 2026. UAlberta at SemEval-2026 Task 5: Disambiguating stories via task decomposition. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Salih Numan Büyükbaş, Doruk Benli, Osman Kara, and Dilara Keküllüoğlu. 2026. Sabanci_group4 at SemEval-2026 Task 5: Uncertainty-aware semantic plausibility scoring via GNLL regression and LLM rationales. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3279 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv*, abs/2507.06261.

Jiaxu Dao, Zhuoying Li, Hangchao Ma, Jinli Tong, Xiaoli Lan, Yifan Lu, and Zhanji Yang. 2026. PuerAI at SemEval-2026 Task 5: Homograph appropriateness assessment via deBERTa contrastive regression and contextual grouping. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. DeepSeek-V3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Fakeha Faisal, Rubab Shah, Syeda Yashfeen Zehra Zaidi, Azkaa Nasir, Sandesh Kumar, and abdul samad. 2026. ConTexT at 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.

Daniel Arturo Peña Gnecco, Edwin Puertas, Juan Carlos Martinez Santos, and Jairo Serrano. 2026. VerbaNexAI at SemEval-2026 Task 5: Few-shot chain-of-thought with selective self-consistency and isotonic calibration for word sense plausibility rating. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Dhruv Goyal, Ishita Gupta, and Jatin Bedi. 2026. Paradise at SemEval-2026 Task 5: On the limitations of surface-level features for graded word sense plausibility prediction. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Azwad Anjum Islam and Tisa Islam Erana. 2026. COGNAC at SemEval-2026 Task 5: LLM ensembles for human-level word sense plausibility rating in challenging narratives. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Valentin Istrate, Mocanu Octavian, and Tatiana Khaidukova. 2026. Narrative Team at SemEval-2026 Task 5: Rating plausibility of word senses in ambiguous sentences through narrative understanding. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Naina Ramesh Jain, Nidhi Arora, Pal Thakkar, and Siba Sankar Sahu. 2026. SVNIT_CSE_AI at SemEval-2026 Task 5: Rating plausibility of word senses in ambiguous sentences using multi-architecture analysis. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Musab Ahmed Khan. 2026. SU NLP 29 at SemEval-2026 Task 5: DynaOrd - hybrid dynamic ordinal regression with LoRA-fine-tuned DeBERTa-v3. In *Proceedings of the 20th International Workshop on*

Semantic Evaluation, San Diego, California. Association for Computational Linguistics.

Chava Laufer, Batel Sara Turjeman, and Chaya Liebeskind. 2026. Team "JCT 2026" at SemEval-2026 Task 5: AmbiStory navigating narrative consistency through a hybrid LLM-NLI ensemble. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Wayu Limsuwan and Attapol Rutherford. 2026. ChulaNLP at SemEval-2026 Task 5: Regression-calibrated LLM for word-sense scoring. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

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.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 task 7: Detection and interpretation of English puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Oxford University Press. 2025. *Oxford English Dictionary Online*. Oxford University Press. Accessed 19 February 2026.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Lis Kanashiro Pereira and Fei Cheng. 2026. CiNet-Handai-Kyodai at SemEval-2026 Task 5: Combining llm prompting, semantic similarity, and synthetic gaze for graded sense plausibility. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. *Preprint*, arXiv:1808.09121.

Pawan Kumar Rajpoot. 2026. Team p1 at SemEval-2026 Task 5: Semantic bridge - augmented encoding for word sense plausibility. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Keith Rayner and Susan A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3):191–201.

Federico Ortega Riba, Jasper Wilkerson, and Kelsey LaFreniere Adams. 2026. CUCLASIC at SemEval-2026 Task 5: LLM prompting strategies for rating ambiguous word senses. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Soumyajit Roy. 2026. Ambirig at SemEval-2026 Task 5: Distributional ordinal modelling for ambiguous word senses in narrative contexts. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Rhea Singhal, Krish Sharma, Lakksh Sharma, and Jatin Bedi. 2026. blue at SemEval-2026 Task 5: NarrBERT : Narrative-aware BERT for word sense disambiguation. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

C. Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103.

Deshan Koshala Sumanathilaka, Nicholas Micallef, Julian Hough, and Saman Jayasinghe. 2026. SwanNLP at SemEval-2026 Task 5: An LLM-based framework for plausibility scoring in narrative word sense disambiguation. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Stephen Taylor. 2026. UWB-NLP at SemEval-2026 Task 5: Synsets and their contexts. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Tong Wu, Thanet Markchom, and Huizhi(Elly) Liang. 2026. NCL-UoR at SemEval-2026 Task 5: Embedding-based methods, fine-tuning, and LLMs for word sense plausibility rating. In *Proceedings of*

the 20th International Workshop on Semantic Evaluation, San Diego, California. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yuming Zhang, Junyu Zhou, Hongyu Li, Yongwei Zhang, Shanshan Jiang, and Bin Dong. 2026. SRCB at SemEval-2026 Task 5: A multi-target finetuning framework for large language models with joint regression and text generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Sunny Zhou, Jordan Youner, and Dean Cahill. 2026. ZYC at SemEval-2026 Task 5: Application of BERT-based contextual embeddings similarity for WSD. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.