CiNet-Handai-Kyodai at SemEval-2026 Task 5: Combining LLM Prompting, Semantic Similarity, and Synthetic Gaze for Graded Sense Plausibility

Lis Kanashiro Pereira[1,2,3], **Fei Cheng**[4]
[1]Center for Information and Neural Networks, Japan
[2]National Institute of Information and Communications Technology, Japan
[3]The University of Osaka, Japan
[4]Kyoto University, Japan
liskanashiro@nict.go.jp, feicheng@i.kyoto-u.ac.jp

Abstract

We present a hybrid system for SemEval-2026 Task 5 on graded word-sense plausibility in narrative contexts. Our approach combines prompt-based large language model (LLM) scoring with three complementary features: semantic embedding similarity, story-conditioned definition generation, and a synthetic gaze signal based on predicted fixation time. We combine these signals using an ordinary least squares regressor. On the official test set, our best system achieves 90.10 $\text{Acc}_{\pm\text{SD}}$ and 79.19 Spearman correlation. The system surpasses the reported human reference score on Acc±SD, highlighting the value of combining LLM-based judgments with targeted linguistic and cognitive-inspired features.

1 Introduction

SemEval-2026 Task 5 (Gehring et al., 2026) focuses on the evaluation of graded word-sense plausibility in narratives, a setting where multiple senses may be simultaneously valid. Although instruction-tuned LLMs can estimate plausibility through direct prompting of the story alone, these scores are often biased toward dominant word senses and may fail to capture subtle contextual cues. To mitigate this, we complement prompt-based scoring with a suite of additional signals: embedding similarity, story-conditioned definition generation (Meconi et al., 2025), and synthetic gaze features. By fusing these components via an ordinary least squares (OLS) regressor, we demonstrate consistent performance gains across the official evaluation metrics.

2 Task Description

SemEval-2026 Task 5 (Gehring et al., 2026) focuses on **rating the plausibility of word senses** in lexically ambiguous sentences through narrative understanding. Unlike traditional Word Sense Disambiguation (WSD), which typically assumes a single "correct" sense, this task acknowledges that ambiguity and underspecification—alongside subjective factors such as individual linguistic experience—can render multiple senses simultaneously plausible in context. The task is built around the AmbiStory dataset (Gehring and Roth, 2025), which consists of short English stories with a fixed structure: three precontext sentences, one ambiguous target

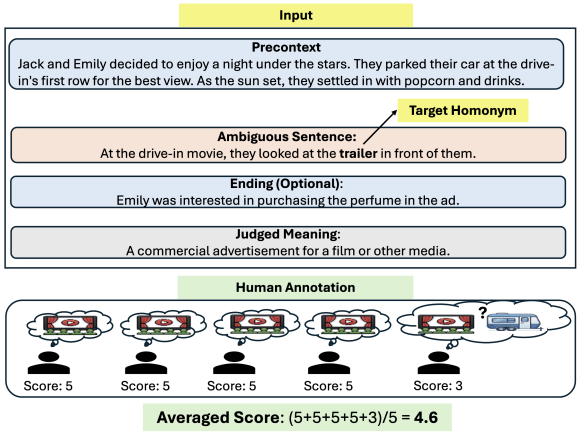

Figure 1: Overview of SemEval-2026 Task 5. (a) Instances consist of a short story with precontext sentences, an ambiguous sentence containing a target homonym, and an optional ending. Annotators score the plausibility (1-5) of a sense description, with the label being the average judgment. (b) Systems are evaluated using complementary measures: Spearman's ρ for ranking quality and Accuracy within Standard Deviation ($\text{Acc}_{\pm\text{SD}}$) with a minimum tolerance of ± 1. Together, these metrics capture both ranking and point-estimate quality relative to human variation.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1950–1956

sentence, and an optional ending sentence. The ambiguous sentence contains a **target homonym** that is ambiguous when read in isolation; surrounding sentences introduce narrative cues that manipulate the plausibility of different senses, so successful prediction requires higher-level story understanding. For each story, the dataset provides a **judged meaning** (i.e., a candidate sense description for the target homonym, accompanied by an example usage sentence), and systems must output a **graded plausibility score** on a 1–5 scale indicating how well this judged meaning fits the story context. An illustration of the task is shown in Figure 1.

AmbiStory contains 3,798 samples organized into 633 setups of six samples each, split into 2,280 / 588 / 930 instances for training, development, and test, respectively. Splits are stratified by homonym, ensuring that target words do not overlap across sets and thus preventing lexical leakage. On average, stories are approximately 49.77 words long (precontext: 31.5 words, ambiguous sentence: 9.24 words, ending: 13.5 words). To capture the graded nature of human sense perception, the dataset includes 19,049 annotations collected with at least five judgments per sample, from which an average plausibility score is computed for each instance.

Systems are evaluated with two complementary metrics. First, the correlation between a model's predicted plausibility scores and the human average is measured using **Spearman's** ρ (Spearman, 1904). Second, to account for the fact that some samples exhibit stronger annotator consensus than others, the task reports **Accuracy within Standard Deviation** ($\text{Acc}_{\pm\text{SD}}$), defined as the proportion of predictions that fall within one standard deviation of the annotators' mean score, with the tolerance floored at ± 1 even when the standard deviation is smaller. Together, these metrics evaluate both how well a system matches the relative ordering of human plausibility judgments (Spearman's ρ) and how often it produces scores that lie within the typical range of human disagreement for each sample ($\text{Acc}_{\pm\text{SD}}$).

3 Method

Our system is a hybrid of **prompt-based plausibility rating** and **auxiliary features** derived from embeddings and synthetic gaze. Specifically, we: (i) query an LLM with a prompt based on the AmbiStory prompt format (Gehring and Roth, 2025) to obtain a plausibility score on the 1–5 scale (§3.1); (ii) compute additional feature(s) from the same input, including embedding-based similarity (§3.2), story-conditioned definition generation (§3.3), and synthetic gaze prediction (§3.4), (iii) and fuse the prompt score and feature values using a **linear regressor** (specifically ordinary least squares, §3.5) trained on the training set to predict the final plausibility score.

3.1 Prompt-based plausibility scoring

Given a 4-5 sentence narrative (with the ambiguous sentence explicitly marked) and a candidate meaning description for the target word, the LLM predicts a graded plausibility score on the 1–5 scale. We evaluate both **zero-shot** and **few-shot** regimes (Table 1); the few-shot setting prepends a small set of labeled examples formatted identically to the target instance.

Chain-of-thought (CoT) variant. For the CoT setting, we augment the base prompt with the instruction: "Think step-by-step about how the given meaning fits the context. Then provide your score." This encourages explicit reasoning about meaning–context compatibility before outputting the final 1–5 plausibility score (reported as `GPT-4.1`$_{\text{CoT}}$ and `GPT-5.1`$_{\text{CoT}}$ in Table 1).

3.2 Embedding-based Similarity Features (EmbSim)

We use the OpenAI `text-embedding-3-large` model to obtain dense semantic representations of text inputs and derive auxiliary similarity-based features for plausibility prediction. Unlike prompt-based scoring, these features provide continuous estimates of semantic compatibility between the narrative context and the candidate meaning.

For each instance, we first construct the full story context by concatenating the precontext, ambiguous sentence, and optional ending. We also encode the judged meaning text provided in the dataset. Based on these representations, we compute two complementary similarity features.

Delta Similarity. To estimate the contribution of the target homonym to the candidate meaning, we compare the judged meaning against (i) the full story and (ii) a modified version of the story in which the target word is replaced with a placeholder token (. . .). Let e_{story}, e_{blank}, and e_{judged} denote the corresponding embeddings. We define:

$$\Delta_{sim} = \cos(e_{story}, e_{judged}) - \cos(e_{blank}, e_{judged}) \quad (1)$$

A larger value indicates that the presence of the target word increases alignment between the story and the candidate sense.

Story-Definition Compatibility (PromptEOL). We additionally construct an augmented text by appending the candidate meaning to the original story using a natural-language template:

```
[story] The word [target] here means
[definition].
```

We then compute cosine similarity between the embedding of the original story and that of the augmented text:

$$Compat = \cos(e_{story}, e_{augmented}) \quad (2)$$

where $e_{augmented}$ denotes the embedding of the template-augmented story. If the candidate meaning is compatible with the narrative context, appending the definition should preserve the overall semantic representation more closely than an incompatible meaning.

The resulting similarity features are used as auxiliary inputs to the linear regressor, corresponding to the Prompt + EmbSim configuration in Table 1.

3.3 Sense definition generation (Def.Gen.)

Sense descriptions can be terse or stylistically mismatched with narrative language, which may weaken both LLM scoring and embedding-based matching. Inspired by definition-generation settings for probing sense understanding in LLMs (Meconi et al., 2025), we introduce a **definition generation** step (Def.Gen.). For each instance, we make one additional LLM call to generate a short, dictionary-style definition of how the target word is used in the given story, using the following prompt:

> **Story:**
> `[story]`
> **Target word:** `[homonym]`
>
> *Write a short dictionary-style definition phrase for how the target word is used in this story. Output only the definition phrase, nothing else.*

After generating the story-specific definition phrase, we embed it and compute cosine similarity to the judged meaning text from the dataset. This definition–meaning similarity is added as an auxiliary feature and combined with the prompt-based score via the linear regressor (§3.5). This configuration corresponds to the Prompt + Def.Gen. + EmbSim setting in Table 1.

3.4 Synthetic gaze feature

Cognitive signals such as gaze capture language processing as it unfolds, providing information about attention allocation and processing difficulty that complements post-hoc plausibility judgments. Since the SemEval data does not include eye-tracking recordings, we use **synthetic gaze** predicted from text. Concretely, we use the gaze-prediction model of Li and Rudzicz (2021), a RoBERTa-based model trained on human eye-tracking corpora, which predicts multiple token-level gaze measures (e.g., number of fixations, first fixation duration, go-past time, total reading time, and fixation proportion). Following prior work using fixation-based measures in NLP (Hollenstein et al., 2021), we use only the **total fixation time** feature (TFT) for the target word as a single scalar gaze signal (Table 1). The intuition is that longer predicted fixation time on the ambiguous word may reflect increased processing difficulty or uncertainty during reading, which can be associated with greater ambiguity and hence provide complementary information for plausibility scoring. This configuration corresponds to the Prompt + Def.Gen. + EmbSim + TFT setting in Table 1.

3.5 Linear fusion with a regressor

Except for the Prompt only setting (which directly uses the LLM score), all other settings combine the prompt score with auxiliary features via a linear model. We fit the fusion model using **ordinary least squares** (OLS) regression on the training set by estimating coefficients that minimize the squared error with respect to the gold mean human rating μ_i for each instance. We use OLS for simplicity, interpretability, and robustness under small-data settings. At inference time, the regressor outputs a real-valued prediction, which we clip to the valid score range $[1, 5]$.

4 Experimental Settings

4.1 Hyperparameter Settings

For all embedding-based features, we compute vector representations using the OpenAI `text-embedding-3-large` model and cosine similarity. For prompt-based inference, we sweep the decoding temperature over

$\{0, 0.1, 0.2, 0.3, 0.7, 0.8\}$ and report results for the best-performing setting on the development set (unless otherwise noted).

4.2 Fine-Tuned Baselines

In addition to prompting-based systems, we train supervised baselines by fine-tuning Transformer encoders to directly predict plausibility scores from the story and the candidate *judged meaning*.

DeBERTa$_{\text{V3-LARGE}}$ regression. We fine-tune `microsoft/deberta-v3-large` with a single regression output. Each instance is formatted as a pair of texts: the first sequence contains the *judged meaning* description (concatenated with the provided example usage sentence), and the second sequence contains the full story context. To make the target word explicit, we mark the homonym occurrence in the ambiguous sentence with special tokens `<tgt>` and `</tgt>` (added to the tokenizer vocabulary). Concretely, the model input is:

- A: `judged_meaning (+ Example: example_sentence)`,
- B: `precontext + <tgt> sentence </tgt> + ending`.

We train the model to regress to the gold mean human rating μ_i. Following common practice for stable regression training, we linearly normalize targets from the 1–5 scale to $[0, 1]$ during training and map predictions back to $[1, 5]$ at inference time (clipping to the valid range). We optimize mean squared error and select the best checkpoint by development Spearman correlation.

Optimization details. We use a maximum input length of 768 tokens, batch size 16, learning rate 1×10^{-5}, and train for 15 epochs with a warmup ratio of 0.1 and weight decay 0.05. To improve stability, we apply gradient checkpointing and mixed-precision training. We additionally use layer-wise learning rate decay with a decay factor of 0.9.

`gpt-oss-20b` instruction fine-tuning. We additionally fine-tune `gpt-oss-20b` using an instruction-following format that mirrors our few-shot prompting setup. Concretely, each training example is converted into a few-shot prompt that includes (i) the task rubric and a small set of in-context demonstrations (story with marked ambiguous sentence, target word, judged meaning, and gold score), followed by (ii) the target instance with its story and judged meaning. The model is trained to output a single plausibility score on the 1–5 scale, supervised with the gold mean human rating μ_i (mapped back to the discrete 1–5 range at inference time). This baseline provides a direct comparison between in-context prompting and parameter-updated instruction tuning under the same prompt format.

5 Results

Tables 1 and 2 summarize our development and test performance in terms of accuracy and Spearman correlation (Acc / Spr). We compare prompt-only scoring against progressively richer feature-augmented variants, and include two fine-tuned baselines for reference.

5.1 Development results

On the development set (Table 1), adding embedding-based similarity (EmbSim) consistently improves performance over the prompt-only setting. For instance, in the GPT-4.1 zero-shot setting, Prompt + EmbSim raises accuracy from 77.04 to 87.24 while Spearman correlation remains comparable (75.83 $\rightarrow$ 74.10). Incorporating story-conditioned definition generation (Def.Gen.) further yields the best overall development results among our prompting-based systems: GPT-4.1 zero-shot with Prompt + Def.Gen. + EmbSim reaches 89.11 / 77.44, and adding the synthetic gaze feature (TFT) achieves our strongest development accuracy (90.47) with competitive correlation (77.36).

While few-shot prompting significantly improves performance in the prompt-only setting—for instance, increasing GPT-4.1 accuracy from 77.04 to 79.25 —the gains from in-context demonstrations are more limited in hybrid settings. This is expected, as the linear regressor already leverages the full training set to fuse auxiliary signals, potentially saturating the benefit of few-shot examples. Nevertheless, feature augmentation remains beneficial even in the few-shot regime: GPT-4.1$_{\text{CoT}}$ with Prompt + Def.Gen. + EmbSim achieves 89.45 / 76.89, and adding the synthetic gaze feature (TFT) attains the highest development Spearman correlation (77.73) among all few-shot runs. These results suggest that (i) explicitly grounding candidate meanings via similarity features and (ii) generating story-specific definitions both improve alignment with human plausibility

LLM	Regime	Prompt (Acc / Spr)	Prompt + EmbSim (Acc / Spr)	Prompt + Def.Gen. + EmbSim (Acc / Spr)	Prompt + Def.Gen. + EmbSim + TFT (Acc / Spr)
GPT-4.1	Zero-shot	77.04 / 75.83	87.24 / 74.10	89.11 / 77.44	**90.47** / 77.36
GPT-5.1	Zero-shot	80.61 / 75.55	85.88 / 73.69	88.94 / 77.51	85.03 / 75.47
GPT-4.1	Few-shot	79.25 / 74.74	83.67 / 72.72	89.11 / 76.91	84.86 / 75.90
GPT-4.1$_{\text{CoT}}$	Few-shot	79.76 / 76.93	85.20 / 74.08	89.45 / 76.89	87.75 / **77.73**
GPT-5.1	Few-shot	81.29 / 74.62	85.20 / 74.84	88.26 / 77.51	–
GPT-5.1$_{\text{CoT}}$	Few-shot	81.97 / 75.09	85.37 / 74.45	86.90 / 77.57	85.54 / 75.33
Fine-Tuning					
DeBERTa$_{\text{LARGE}}$	Fine-Tuning	66.03 / 66.03			
gpt-oss-20b	Fine-Tuning	73.00 / 65.00			

Table 1: Development-set results (accuracy / Spearman). Underlined scores correspond to models submitted to the leaderboard. Bold scores indicate the best accuracy and Spearman scores.

Model	Regime	Setting	Dev (Acc / Spr)	Test (Acc / Spr)
Human	-	Upper bound Scoring	-	89.20 / **83.40**
GPT-4.1	Zero-shot	Prompt + EmbSim	87.24 / 74.10	89.46 / 76.09
GPT-4.1	Zero-shot	Prompt + Def.Gen. + EmbSim	89.11 / **77.44**	89.89 / 78.87
GPT-4.1	Zero-shot	Prompt + Def.Gen. + EmbSim + TFT	**90.47** / 77.36	**90.10** / 79.19

Table 2: Submitted models only (underlined in Table 1), plus the human baseline. We report development and test results (accuracy / Spearman). Test results are obtained by ensembling three runs per model with sampling temperature 0.8. Bold indicates the best accuracy and Spearman scores on each split.

judgments beyond what simple in-context learning provides.

Finally, the fine-tuned baselines underperform the prompting-based LLM systems on the development set. In particular, DeBERTa$_{\text{LARGE}}$ reaches 66.03 / 66.03 and `gpt-oss-20b` reaches 73.00 / 65.00, highlighting the advantage of leveraging large instruction-tuned LLMs with auxiliary semantic and cognitive signals for this task.

5.2 Submitted systems and test results

Table 2 reports the test performance of our submitted systems together with the human upper-bound baseline. We focus on the three GPT-4.1 zero-shot submissions, which progressively incorporate EmbSim, Def.Gen., and TFT.

On the test set, we observe consistent improvements as additional signals are introduced. Prompt + EmbSim achieves 89.46 / 76.09, while adding definition generation yields a substantial gain in correlation to 89.89 / 78.87. Our best system, Prompt + Def.Gen. + EmbSim + TFT, attains 90.10 / 79.19, which is the strongest overall submitted result.

Notably, **all submitted configurations outperform the human baseline on accuracy** (ranging from 89.46 to 90.10 vs. 89.20), while remaining below the human upper bound in Spearman correlation (79.19 vs. 83.40). This pattern suggests that our feature-augmented approach produces pointwise scores that fall within the task's acceptance window ($Acc_{\pm SD}$) more often than individual human judgments, while humans still provide more consistent global ranking of plausibility.

All submitted test results are obtained by **ensembling three runs per model** with sampling temperature 0.8 (Table 2), which reduces variance and yields more stable plausibility predictions. Overall, the test results indicate that combining prompt-based scoring with semantic similarity, story-specific definition generation, and a lightweight gaze-derived signal provides complementary benefits, improving both accuracy and rank correlation.

6 Analysis

To understand where the synthetic gaze feature helps, we compare our full system Prompt + Def.Gen. + EmbSim + TFT against the same system without TFT (Prompt + Def.Gen. + EmbSim)

on the development set: 90.47 / 77.36 versus 89.11 / 77.44 in accuracy and Spearman correlation (Table 2). The two systems produce very similar outputs overall; on a representative dev run, they are scored identically by the official metric on 543 of 588 items, so the 1.4-point accuracy gap is decided on the remaining 45 items.

The gains from TFT are not uniform across the plausibility range. Grouping items by their gold rating, the TFT variant is $+5.5$ points on items rated below 2 and $+4.0$ points on items rated between 2 and 3, essentially tied on items between 3 and 4, and -3.1 points on items rated 4 or above. In short, TFT helps on the implausible end of the scale and hurts on the most plausible end.

The reason is a small but consistent shift in calibration. The no-TFT variant tends to over-predict: on average, its score is 0.06 points above the gold rating. Adding TFT pulls predictions downward and almost eliminates this bias (average error 0.02). Pushing scores down is the right direction on items where the correct answer is low—where the no-TFT variant over-predicts—but occasionally overshoots on items where the correct answer is high.

A representative case is the homonym *tips*. The candidate meaning was the monetary sense ("a relatively small amount of money given for services rendered"), which annotators rated at a low plausibility of 1.8:

> **Story:** Anna walked into the small cafe. . . She noticed the barista arranging flowers on the tables. Nearby, a group of kids was crafting paper airplanes, giggling as they carefully folded the edges. Those are some small **tips** over there.

Without TFT, the system predicted 3.55, likely because *cafe* and *barista* in the preceding context evoke the monetary sense even though the target sentence refers to the folded tips of paper airplanes. With TFT, the prediction dropped to 2.35, much closer to the gold score. The same pattern appears for *rare* (gold 2.0; no-TFT 3.54; TFT 2.83), where the pre-context features a butcher shop and a steak "so unique it could only be found once in a blue moon"—priming the infrequent sense—but the target sentence *"this particular steak has to be rare"* uses the cooking-doneness sense. A similar case arises for *draw* (gold 2.0; no-TFT 4.14; TFT 3.34), where cigarettes in the pre-context prime the inhale sense but the sentence ultimately refers to drawing on a sketchpad. In each case, a salient topical cue activates the wrong sense of the homonym, and the gaze signal produces a better-calibrated score than the no-TFT variant. These examples are consistent with the hypothesis that longer predicted reading time on the target word reflects processing difficulty when the prompted sense is a poor fit, helping the model discount misleading topical associations.

The same mechanism works against TFT on idiomatic high-plausibility items. On *"He got to the **point**"* with the essential-meaning sense (gold 4.00), the no-TFT variant predicts 3.76 while the TFT variant predicts 2.51—an under-prediction of 1.25 that drops the item outside the metric's tolerance. Cases of this kind, where the target sense is licensed by a conventional collocation rather than by the surrounding topic, account for most of the TFT variant's losses on items with gold ≥ 4 and point toward stronger ordinal calibration (mentioned as future work in §7) as a natural next step.

7 Conclusion

We proposed a hybrid system for graded sense plausibility that augments prompt-based LLM scoring with embedding similarity, story-conditioned definition generation, and a target-word synthetic gaze feature, fused via ordinary least squares regression. On the test set, our best model achieves 90.10 Acc±SD and 79.19 Spearman correlation, exceeding the reported human reference score on Acc±SD. Future work will explore stronger calibration for ordinal scoring and richer cognitive features.

Acknowledgments

References

Janosch Gehring, Selina Meyer, and Michael Roth. 2026. SemEval-2026 task 5: Rating plausibility of word senses in ambiguous stories through narrative understanding. In Proceedings of the 20th International Workshop on Semantic Evaluation, San Diego, California. Association for Computational Linguistics.

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

Nora Hollenstein, Emmanuele Chersoni, Cassandra L Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021. Cmcl 2021 shared task on eye-tracking prediction. In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 72–78.

Bai Li and Frank Rudzicz. 2021. TorontoCL at CMCL 2021 shared task: RoBERTa with multi-stage fine-tuning for eye-tracking prediction. In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 85–89, Online. Association for Computational Linguistics.

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, pages 33897–33916, Suzhou, China. Association for Computational Linguistics.

Charles Spearman. 1904. The proof and measurement of association between two things. The American Journal of Psychology, 15(1):72–101.