Team p1 at SemEval-2026 Task 5: Semantic Bridge - Augmented Encoding for Word Sense Plausibility

Pawan Kumar Rajpoot
pawan.rajpoot2411@gmail.com

Abstract

We present a hybrid system for SemEval 2026 Task 5: Rating Plausibility of Word Senses in Ambiguous Stories. Our approach reframes LLMs as feature generators rather than direct predictors. We combine two subsystems: one that appends LLM-generated hints to the input context and trains an encoder-based regression model, and another that feeds structured hints from multiple LLM configurations into a lightweight regression ensemble. We generate multilingual enrichments to probe LLMs for complementary signals and take advantage of the fact that translation into certain languages implicitly disambiguates word senses, making the encoder more robust. The 50/50 ensemble achieves 859/930 (92.37%) accuracy with Spearman $\rho = 0.8384$ on the test set, exceeding the estimated human ceiling of 89.2%. The same LLM enrichments, processed through fundamentally different paradigms (tabular regression vs. full-text encoding), produce complementary errors that cancel under ensembling. Notably, simple 50/50 averaging captures this gain without any learned combiner, suggesting that representation diversity is the primary driver.

1 Introduction

Word sense disambiguation (WSD) is one of the oldest tasks in natural language processing (Navigli, 2009), dating back to the earliest work in machine translation. SemEval 2026 Task 5 (Gehring et al., 2026) introduces a graded formulation of WSD for English, building on the insight that word senses are not discrete categories but admit degrees of applicability (Erk and McCarthy, 2009): given a short story containing an ambiguous word, rate how plausible a proposed word sense is on a continuous 1–5 scale. This is more challenging than traditional WSD classification because a proposed sense may be partially consistent with the story context, or consistent only under certain interpretations. Five human annotators independently rate each story–sense pair, and a system's prediction is considered correct if it falls within one standard deviation of the mean (or within 1.0, whichever is larger).

Two observations motivate our approach. First, sense distinctions that are ambiguous in one language are often resolved naturally in another through differences in lexical choice or verb semantics. We act on this by prompting an LLM to generate translations and sense-relevant paraphrases in Hindi, using the cross-lingual signal as an additional feature. Second, large language models are excellent reasoners but poor calibrators (Guo et al., 2017). Rather than relying on an LLM to directly output numeric ratings, we feed LLM-generated features, including cross-lingual cues, into downstream regression models (System A), letting the regressor handle numeric calibration. We further combine this tabular approach with a fine-tuned cross-encoder (System B) that reads the full story and enrichment text. Because the two systems operate on entirely different representations (8 numeric features vs. raw text), their errors tend to be complementary, which benefits the final ensemble (Figure 2).

Our ensemble achieves 92.37% accuracy on the test set, well above the human ceiling. The benchmark reports DeepSeek-V3 achieving 79.0% (zero-shot) and 81.6% (4-shot) accuracy; we were unable to replicate the zero-shot result, obtaining 71.29% with our prompt. The regression model alone reaches 89.78%, showing that decoupling reasoning from calibration is effective even without text input. Our code and annotated data are publicly available.[1,2]

[1]Code and data: https://github.com/pawan2411/semeval26-task5

[2]Fine-tuned model: https://huggingface.co/pawan2411/semstage7b

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3095–3102

2 Background

2.1 Task Formulation

Each sample in SemEval 2026 Task 5 consists of a short story (precontext + ambiguous sentence + ending) and a proposed word sense (meaning + example sentence) for a target homonym. Five human annotators independently rate how well the proposed sense fits the story on a 1–5 scale. The gold label is the average of these five ratings. A system prediction is correct if it falls within max(SD, 1.0) of the gold average, where SD is the annotator standard deviation.

2.2 Dataset Statistics

A critical property of the dataset is that homonyms are disjoint across splits: train, dev, and test share virtually no homonyms (overlap = 0–1). This means the system must generalize to unseen homonyms, not just unseen stories for known homonyms. Table 1 summarizes key statistics. Each homonym has 2–5 distinct senses (mean = 2.2), with multiple story–sense pairs per homonym.

Split	Samples	Homonyms	Avg Rating	Avg SD	Senses/Hom
Train	2,280	220	3.14	0.95	2.2
Dev	588	55	3.12	0.95	2.2
Test	930	87	3.15	0.94	2.2
Total	3,798	361	3.14	0.95	2.2

Table 1: Dataset statistics. Per-split homonym counts sum to 362 because one form appears in two splits; the unique total is 361.

2.3 Rating Distribution

The rating distribution (Figure 1) is roughly symmetric around 3.0 but bimodal, with peaks near 1–2 (senses that clearly do not fit) and 4–5 (senses that clearly fit). The intermediate range (2.5–3.5) contains the hardest cases.

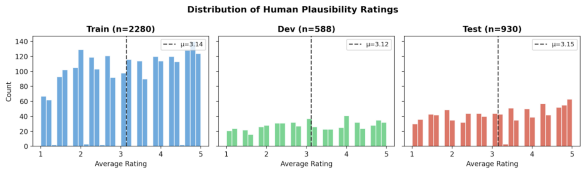

Figure 1: Distribution of Human Plausibility Ratings across Train, Dev, and Test splits.

2.4 Annotator Agreement

Annotator standard deviation averages 0.95 across all splits. For 62% of samples SD falls below 1.0, so the effective tolerance is exactly 1.0; for the remaining 38%, disagreement widens it. Disagreement is highest for mid-range ratings (2.5–3.5) and lowest for extremes.

3 System Overview

3.1 Semantic Bridging

Semantic bridging connects a candidate word sense to its story context through structured reasoning. We implement this as a Chain-of-Thought (Wei et al., 2022) enrichment pipeline, following prior work in sense-context matching (Navigli, 2009; Loureiro et al., 2021; Huang et al., 2019). Each enrichment produces a discrete label: **POOR** (sense does not fit), **PARTIAL**[3] (ambiguous fit), or **STRONG** (sense aligns well).

For DeepSeek (CoT extraction): For each story–sense pair, we prompted DeepSeek-Chat-V3 (DeepSeek-AI, 2024) to: (1) identify the target homonym and its proposed sense, (2) locate supporting or contradicting evidence, (3) reason step-by-step about semantic fit, and (4) output a POOR, PARTIAL, or STRONG judgment. We ran this independently in English, Hindi, and Korean (cot_e, cot_h, cot_k), with prompts designed for native reasoning in each language to surface distinct disambiguation signals.

For Qwen fine-tuning (SFT extraction): We fine-tuned Qwen-2.5-7B (Yang et al., 2024) using LoRA (Hu et al., 2022) to distill (Hinton et al., 2015) the DeepSeek CoT outputs into a smaller model. The DeepSeek-generated semantic bridges on the training set served as silver-standard targets. We trained three separate LoRA adapters (one per language: English, Hindi, Korean) using English input in all cases, producing enrichments sft_e, sft_h, and sft_k. We distill into Qwen rather than SFT DeepSeek directly for compute tractability.

3.2 Translation

We translated all story fields into Hindi and Korean using DeepSeek-Chat-V3 (DeepSeek-AI, 2024), batching five text fields per API call (Appendix A.4). Translation naturally resolves ambiguity because different senses of a polysemous word are typically lexicalized as distinct words in the target language (Diab and Resnik, 2002; Lefever and Hoste, 2010; Banea et al., 2011). We chose Hindi and Korean because they span distinct syntactic structures (SVO vs. SOV) and unrelated language

[3]The CoT prompt template (Appendix A.1) uses "MODERATE FIT"; we rename it to PARTIAL for clarity.

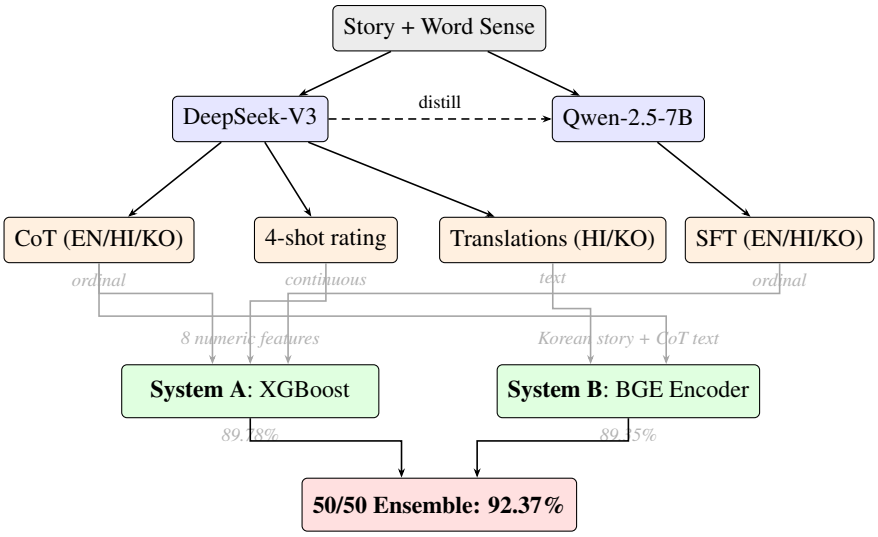

Figure 2: System pipeline. DeepSeek-V3 generates CoT enrichments, 4-shot ratings, and translations; Qwen-2.5-7B (distilled from DeepSeek) produces SFT labels. System A uses only numeric features; System B reads full text. Dashed arrow = knowledge distillation.

families, so they tend to disambiguate different sets of words. Translation here is not a WSD solution but a *forcing function* that surfaces the LLM's sense commitment as a discrete lexical choice.

3.3 4-Shot Rating Prediction

We also prompted DeepSeek-Chat-V3 in a 4-shot setting to directly predict a 1–5 plausibility rating, following the benchmark baseline (Gehring et al., 2026). The prompt included four exemplars (scores 1, 2, 3, and 5). The 4-shot prediction alone achieves 81.6% on the test set, better than our zero-shot baseline (71.29%, vs. 79.0% reported by Gehring et al. (2026)) but far below our final system. This continuous score (`deepseek_4shot`) is used alongside the discrete POOR/PARTIAL/STRONG labels as input to System A.

3.4 Final Augmented Data

Feature	Source Model	Method	Output
cot_e	DeepSeek-Chat-V3	CoT English	POOR / PARTIAL / STRONG
cot_h	DeepSeek-Chat-V3	CoT Hindi	POOR / PARTIAL / STRONG
cot_k	DeepSeek-Chat-V3	CoT Korean	POOR / PARTIAL / STRONG
sft_e	Qwen-2.5-7B-LoRA	SFT English	POOR / PARTIAL / STRONG
sft_h	Qwen-2.5-7B-LoRA	SFT Hindi	POOR / PARTIAL / STRONG
sft_k	Qwen-2.5-7B-LoRA	SFT Korean	POOR / PARTIAL / STRONG
deepseek_4shot	DeepSeek-Chat-V3	4-shot prompting	Continuous (1–5)
deepseek_rating	DeepSeek-Chat-V3	CoT raw rating	Continuous (1–5)

Table 2: Complete feature set. The first six features are discrete plausibility labels from the semantic bridge; the last two are continuous numeric predictions.

3.5 System A: Regression Ensemble

System A encodes the six discrete labels as ordinal features (POOR → 1, PARTIAL → 2, STRONG → 3) and combines them with two continuous features: deepseek_rating and deepseek_4shot (Section 3.3), yielding 8 inputs total. The point is to let LLMs handle qualitative reasoning while a simple regressor learns the mapping to calibrated numeric scores. The input is purely numeric: no story text, no enrichment prose, no embeddings.

3.6 System B: Fine-Tuned Cross-Encoder

System B uses BAAI/bge-reranker-v2-m3 (Chen et al., 2024), a multilingual cross-encoder based on XLM-RoBERTa (Conneau et al., 2020), fine-tuned as a regressor. The input concatenates the Korean-translated story with the sense definition and full-text CoT enrichments in English and Korean:

```
{korean_story} [SEP] The word '{homonym}' means: {meaning} (e.g., {example}) [ANALYSIS] {enrichment_e} {enrichment_k}
```

Unlike System A, System B reads the full CoT reasoning text, attending to semantic cues lost in ordinal reduction. We use Korean translations as the primary story input because translation forces the translator to commit to a specific sense.

4 Experimental Setup

4.1 System A Training

Parameter	Value
Framework	XGBoost
Objective	Huber loss ($\delta = 1.0$)
Number of features	8 (6 ordinal + 2 continuous)
Ordinal encoding	POOR=1, PARTIAL=2, STRONG=3
Ensemble strategy	Average of 3-feature and 8-feature models
Test accuracy	89.78%

Table 3: System A configuration (Table 2 lists all features).

System A (Table 3) uses XGBoost (Chen and Guestrin, 2016) with Huber loss, ensembling a 3-feature model with the full 8-feature model.

4.2 System B Training

Parameter	Value
Base model	BAAI/bge-reranker-v2-m3 (XLM-RoBERTa)
Learning rate	1e-5
Batch size	4
Max sequence length	512 tokens
Training epochs	15 (early stopping on dev loss)
Enrichment languages	English + Korean (concatenated)
Precision	BF16
Dev accuracy (before retrain)	88.27%

Table 4: System B training configuration.

System B (Table 4) fine-tunes BAAI/bge-reranker-v2-m3 as a regression head. Training uses early stopping on dev loss, giving 88.27% dev accuracy (Table 7); for the final submission the same configuration was retrained on combined train+dev (2,868 samples), yielding 89.35% test accuracy (Table 5).

4.3 Ensemble Configuration

The final prediction is a simple 50/50 average of System A and System B outputs. The ensemble works because the two systems see very different inputs (Dietterich, 2000): System A sees only 8 numeric features while System B reads full text but lacks the 4-shot prediction, Hindi/Korean SFT signals, and continuous DeepSeek rating. This difference pushes the ensemble to 92.37%, exceeding the human ceiling of 89.2%.

5 Results

5.1 Overall System Comparison

Table 5 compares all systems on the test set (930 samples). The ensemble achieves 92.37% accuracy with Spearman $\rho = 0.8384$ (Pearson $r = 0.8378$, MAE = 0.53), exceeding the human ceiling by over 3 points. Both individual systems already surpass the human ceiling, but the ensemble adds another 2.5+ points, which is hard to explain without assuming that the two systems fail on different samples.

System	Correct	Accuracy	Δ vs Human
DeepSeek-V3 zero-shot (benchmark)	—	79.00%*	−10.20%
DeepSeek-V3 zero-shot (ours)	663	71.29%	−17.91%
DeepSeek-V3 4-shot (benchmark)	—	81.60%	−7.60%
Estimated human ceiling	—	89.20%	—
System B (BGE cross-encoder)	831	89.35%	+0.15%
System A (Huber regression ensemble)	835	89.78%	+0.58%
Ensemble (A + B, 50/50)	**859**	**92.37%**	**+3.17%**

Table 5: Overall system comparison on the test set (930 samples). Δ vs Human is relative to the estimated human ceiling of 89.20%. *Reported by Gehring et al. (2026); we were unable to replicate this result with our prompt, obtaining 71.29% instead. The System B row (89.35%) uses the retrained train+dev model; the 88.27% in Table 7 is the train-only model on dev.

5.2 System A Feature Ablation

Table 6 shows incremental feature additions. The raw 4-shot prediction achieves 81.6% (Section 3.3); as the sole Huber regressor input, accuracy drops to 76.88% because regression shrinks predictions toward the mean. Adding ordinal features and ensembling the 3f + 8f models reaches 89.78%.

Feature Set	Correct	Accuracy
fourshot_pred only	715	76.88%
cot_h + cot_e (best 2-ordinal)	784	84.30%
All 6 ordinal features	795	85.48%
All 6 ordinal + fourshot_pred	817	87.85%
System A ensemble (Huber 3f + Huber 8f)	**835**	**89.78%**

Table 6: System A feature ablation on the test set. Each row adds features incrementally.

5.3 System B Enrichment Ablation

Table 7 ablates enrichment languages. Without enrichment the cross-encoder achieves 71.43%; English alone yields 86.90%. The best configuration is English + Korean (88.27%), outperforming all three languages (85.88%), suggesting Hindi introduces noise for the encoder.

Enrichment Configuration	Dev Accuracy	Δ vs None
No enrichment (Korean story only)	71.43%	—
Hindi only	74.49%	+3.06%
Korean only	74.49%	+3.06%
Hindi + Korean	80.78%	+9.35%
English only	86.90%	+15.47%
English + Hindi	86.73%	+15.30%
English + Hindi + Korean	85.88%	+14.45%
English + Korean (selected)	**88.27%**	**+16.84%**

Table 7: System B enrichment language ablation, on the *dev* set (train-only model). Δ vs None is relative to the no-enrichment baseline (Korean story only). The selected row (English + Korean) was retrained on train+dev for the final submission, yielding 89.35% on test (Table 5).

6 Analysis

6.1 Error Characterization

The ensemble misclassifies 71 of 930 test samples. Table 8 shows that errors are almost exclusively *regression toward the mean*. For samples with gold ratings 4.0–5.0, all 31 errors are under-predictions; for gold ratings below 3.0, all 27 errors are over-predictions. These one-directional errors account for 82% (58/71) of all failures. The system hedges toward the center of the scale, consistent with the Huber loss objective, which penalizes large residuals less aggressively than MSE.

Gold Range	Over-pred	Under-pred	Errors / Total	Error %
1.0–2.0	15	0	15 / 201	7.5%
2.0–3.0	12	0	12 / 195	6.2%
3.0–4.0	4	9	13 / 226	5.8%
4.0–5.0	0	31	31 / 308	10.1%

Table 8: Error direction by gold rating range. The system systematically over-predicts for low-gold samples and under-predicts for high-gold samples.

6.2 Tolerance and Near-Misses

The evaluation metric's tolerance is max(SD, 1.0), meaning samples with high annotator agreement (SD < 1.0) receive the strictest threshold. The error rate under strict tolerance (SD < 1.0, tolerance = 1.0) is 9.7% (52/537), roughly double the 4.8% (19/393) for samples where annotator disagreement widens the band. Of the 71 errors, 54% (38/71) are near-misses, falling within 0.3 points of the tolerance boundary; only 7% (5/71) exceed it by more than 1.0 point.

At first glance, samples with bimodal annotator splits (choice spread ≥ 3) have a *lower* error rate (4.6%) than non-bimodal samples (9.6%). This is a tolerance-formula artifact rather than a genuine ease effect: high annotator disagreement inflates SD, widening the max(SD, 1.0) tolerance window, so the same absolute prediction error is more likely to fall inside the acceptance band.

6.3 Error Concentration

Errors cluster on a few homonyms: 46 of 87 (53%) have zero errors, while the top 13 account for over half of all failures. The worst is *disappear* (4/6 wrong, 67%), where the system consistently under-predicts (μ_{pred} = 3.27 vs. μ_{gold} = 4.87).

6.4 Qualitative Failure Modes

We manually inspected the worst errors and found two recurring patterns:

Indirect evidence requiring inference. The system fails when the ending confirms the proposed sense through synonyms or entailment rather than lexical overlap. For example, *"stop"* with meaning "come to a halt, stop moving" receives gold = 5.0 (SD = 0.0, unanimous agreement) but pred = 2.99. The ending ("She needed this noise to end") uses "end" rather than "stop," requiring the model to infer semantic equivalence.

Figurative–literal confusion. For *"holes"* with meaning "a fault," the system predicts 4.28 against a gold of 1.8 (SD = 1.30) when the ending mentions "paintings full of these flaws." The high SD indicates annotator disagreement (one rater gave 4), yet the system commits strongly to the wrong reading. Similarly, *"bang"* meaning "sudden loud noise" (gold = 4.0) is under-predicted at 1.97 because the ending uses "burst with excitement" with no auditory cue.

Limitations

Regression-to-mean bias. The dominant failure mode (82% of errors, Section 6.1) is systematic hedging toward the scale center, a direct consequence of Huber loss down-weighting large residuals. A natural remedy is *quantile regression* or *asymmetric loss* that penalizes under-prediction of high-gold and over-prediction of low-gold samples more heavily.

Near-miss fragility. Since 54% of errors are near-misses (within 0.3 of the tolerance boundary), post-hoc calibration (e.g., Platt scaling (Platt, 1999) or isotonic regression) could push many predictions inside the acceptance window without retraining.

Feature gaps. Errors on homonyms like *holes* and *bang* stem from conflating figurative and literal readings; WordNet supersense tags (Miller,

1995) could help distinguish these. The *stop*/"end" failure (Section 6.4) suggests adding an NLI feature (Bowman et al., 2015) to capture synonymy beyond lexical overlap. Our pivot languages (Hindi, Korean) were guided by SVO/SOV contrast but not systematically optimized; Hindi is kept as a System A feature but dropped from System B's text input (Table 7), suggesting typological distance from English matters more than word order alone. Future work could explore Russian, German, or VSO languages like Tagalog.

References

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2011. Word sense disambiguation with multilingual features. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Alexis Conneau, Karttikeya Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

DeepSeek-AI. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 255–262. Association for Computational Linguistics.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449. Association for Computational Linguistics.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

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

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514. Association for Computational Linguistics.

Els Lefever and Véronique Hoste. 2010. SemEval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20. Association for Computational Linguistics.

Daniel Loureiro, Jose Camacho-Collados, and Luis Espinosa-Anke. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Wang, Bowen Zheng, Chengyuan Yu, Dayiheng Li, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

A Prompt Templates

A.1 Semantic Bridge CoT Prompt (DeepSeek Inference)

This prompt is used with DeepSeek-Chat-V3 to generate chain-of-thought enrichments at inference time for train, dev, and test sets. It is run independently in English, Hindi, and Korean.

System Prompt:

```
You are an expert linguistic analyst specializing
in word sense disambiguation and semantic fit
analysis.

Your task is to evaluate how well a proposed
meaning of a word fits within a given story
context.

For each story, you will:
1. Understand the story context and narrative
2. Analyze how the target word is used in the
   ending
3. Check if the proposed meaning is confirmed or
   contradicted by evidence in the story
4. Identify any wordplay, puns, or double meanings

Then classify the fit into one of three categories:

- STRONG FIT: The proposed meaning clearly and
  unambiguously matches the story context. The
  ending provides strong evidence confirming this
   specific meaning.

- MODERATE FIT: The fit is AMBIGUOUS or CONFUSING.
   This happens when:
  * The word is complex with multiple plausible
     interpretations in context
  * There is wordplay, puns, or intentional double
     meanings
  * The context sends mixed signals
  * Reasonable people would DISAGREE about whether
     the meaning fits

- POOR FIT: The proposed meaning is clearly
  contradicted by the story context. The ending or
   context provides evidence that a DIFFERENT
   meaning is intended.

Your response must include:
1. Step-by-step reasoning analyzing the story and
   word usage
2. A final assessment in this exact format:
   [CLASS - brief explanation]
```

User Input Template:

```
Story: {story}

Target Word: "{homonym}"
Proposed Meaning: "{meaning}"
Example of this meaning: "{example}"

Analyze how well this meaning fits the story
context.
```

A.2 Label-Guided Prompt (SFT Training Data)

This prompt is used with DeepSeek-Chat-V3 to generate silver-standard annotations for fine-tuning Qwen-2.5-7B. Unlike the inference prompt, it exposes the gold human score to guide the reasoning, producing high-quality training targets.

```
You are a linguistic expert. You will analyze
whether a word meaning fits a story context.

IMPORTANT: The correct human judgment score is
provided. Your job is to EXPLAIN WHY this score
is correct.

---

Story: {story}

Target Word: "{homonym}"
Proposed Meaning: "{meaning}"
Example of this meaning: "{example}"

TRUE HUMAN SCORE: {score}/5
(1 = doesn't fit at all, 5 = fits perfectly)

---

Your task: Explain WHY the score is {score}.
Use this reasoning:

Step 1 - Story Understanding:
What happens in this story? (1 sentence)

Step 2 - Ending Analysis:
What does the ending reveal about the word's
meaning?

Step 3 - Evidence Check:
- If score is HIGH (4-5): What phrases
  CONFIRM this meaning fits?
- If score is LOW (1-2): What phrases
  CONTRADICT this meaning?
- If score is MEDIUM (3): What creates the
  ambiguity?

Step 4 - Wordplay/Idiom Check:
Is there a pun, wordplay, or idiom? How does it
affect the fit?

---

OUTPUT FORMAT:

<cot>
[Step-by-step reasoning - 3-5 sentences]
</cot>

<analysis>
[Analysis: FIT_LEVEL - key evidence from story]
</analysis>

Where FIT_LEVEL matches the score:
- Score 1-2 -> POOR FIT
```

```
- Score 3   -> MODERATE FIT
- Score 4-5 -> STRONG FIT

Keep <analysis> under 20 words.
```

A.3 4-Shot Rating Prediction Prompt

This prompt is used with DeepSeek-Chat-V3 to obtain a direct continuous plausibility rating on the 1–5 scale.

System Prompt:

```
You will read a short text. One of the sentences
is highlighted in bold. That sentence contains a
word that can typically take on multiple different
meanings, depending on the context. One of those
meanings is given to you.

Your task is simple: Annotate how plausible a
meaning of a word is in the context of the short
text using one of five scores:

* 1: The displayed meaning is not plausible at
     all given the context.
* 2: The displayed meaning is theoretically
    conceivable, but less plausible than other
      meanings.
* 3: The displayed meaning represents one of
    multiple, similarly plausible interpretations.
* 4: The displayed meaning represents the most
    plausible interpretation; other meanings may
      still be conceivable.
* 5: The displayed meaning is the only plausible
     meaning given the context.

There will be times where there is no objectively
correct answer. Whatever the case, always look at
all of the sentences and carefully think about how
plausible each meaning would be.
```

Few-Shot Exemplars:

```
Take a look at the following examples.
--
**The bat flew out of the cave.**
In this context, how plausible is it that the
meaning of the word "bat" is "A sports implement
for hitting balls (e.g. in baseball)"?
Correct answer: 1
--
The letter specified where to meet him. **So after
reading it, I went to the bank.**
In this context, how plausible is it that the
meaning of the word "bank" is "a financial
institution"?
Correct answer: 3
--
The composer often spontaneously had ideas for new
melodies. **She writes notes on a sheet of paper.**
She can later turn these into a piece.
In this context, how plausible is it that the
meaning of the word "notes" is "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.** His painting of a
flower looked really realistic!
In this context, how plausible is it that the
meaning of the word "draws" is "to attract;
direct towards itself"?
Correct answer: 5
--
```

User Query Template:

```
{few_shot}

Now take a look at the following text:
{precontext} **{sentence}** {ending}

In this context, how plausible is it that the
meaning of the word "{word}" is "{word_sense}"
(as in: "{example}")?

Return only the numbered score (1, 2, 3, 4 or 5).
Do not return anything else!
```

A.4 Translation Prompt (DeepSeek)

This prompt is used with DeepSeek-Chat-V3 to translate all story fields into Hindi or Korean. Each sample is translated in a single batched API call.

System Prompt:

```
You are a translator. Output only valid JSON,
no markdown.
```

User Prompt Template:

```
Translate these English texts to {language}.
Return ONLY a valid JSON object with the same
keys and {language} translations.
No markdown code blocks, no explanation, just
the raw JSON.

{
  "precontext": "...",
  "sentence": "...",
  "ending": "...",
  "judged_meaning": "...",
  "example_sentence": "..."
}
```

The model returns a JSON object with identical keys containing the translated text. Non-text fields (homonym, numeric scores) are preserved from the original English data. We used `temperature=0.3` for all translation calls.