VerbaNexAI at SemEval-2026 Task 5: Few-Shot Chain-of-Thought with Selective Self-Consistency and Isotonic Calibration for Word Sense Plausibility Rating

Daniel Peña Gnecco, Jairo Serrano, Edwin Puertas, Juan Carlos Martinez-Santos
Universidad Tecnológica de Bolívar, Cartagena, Colombia
{dgnecco, jserrano, epuerta,jcmartinezs}@utb.edu.co

Abstract

We present a system for rating word sense plausibility in ambiguous narrative contexts for SemEval-2026 Task 5. Our approach ensembles three large language models (Llama-3.1 70B, Qwen-2.5 32B, and Gemma-2 27B) using a computationally efficient, uncertainty-aware pipeline. We combine few-shot chain-of-thought prompting with *selective self-consistency*, which applies stochastic multiple sampling exclusively to items identified as inherently ambiguous. This targeted strategy reduces inference costs by approximately 45% while maintaining robustness in predictions. To correct the systematic bias of LLMs toward extreme ratings, we apply isotonic regression to shift the output distribution toward patterns of human judgment. Our system achieves a Spearman correlation of 0.67 and an accuracy within 0.76 standard deviations, ranking 34th out of 79 participating teams (top 43% without task-specific fine-tuning). Detailed error analysis reveals that while our system performs strongly on clear contexts ($\rho = 0.78$), current prompting paradigms struggle significantly to model multimodal human disagreement in genuinely ambiguous cases ($\rho = 0.58$), highlighting an important challenge for future work on subjective semantic tasks.

1 Introduction

Traditional Word Sense Disambiguation (WSD) tasks typically assume a single correct sense per instance. However, real-world language often exhibits genuine ambiguity, with multiple interpretations remaining plausible to varying degrees due to insufficient context. SemEval-2026 Task 5 (Gehring et al., 2026) formalizes this challenge by requiring systems to predict ordinal plausibility ratings (1–5) for word senses in narrative contexts. The difficulty is compounded by the fact that human annotators demonstrate substantial disagreement, reflecting genuine linguistic uncertainty rather than mere annotation noise.

To address this, we propose an uncertainty-aware inference pipeline that leverages the reasoning capabilities of Large Language Models (LLMs) while mitigating their known limitations in ordinal calibration and computational cost. Our system ensembles three diverse open-weight models through a four-stage process: (1) few-shot chain-of-thought (CoT) prompting to structure plausibility reasoning, (2) selective self-consistency sampling to dynamically allocate compute to ambiguous instances, (3) arithmetic ensemble aggregation, and (4) isotonic regression to calibrate raw predictions against empirical human rating distributions.

We make three key contributions:

- *Selective self-consistency*: a heuristic-driven inference strategy that reduces computational overhead by ≈45% while improving Spearman correlation on high-variance items (+0.05).
- Post-hoc isotonic calibration corrects the LLM bias toward extreme predictions, providing a substantial gain at negligible cost (+0.017 Spearman on the training set).
- A granular variance analysis exposing a fundamental limitation of current LLM prompting: while effective for low-variance items ($\rho = 0.78$), averaging stochastic samples fails to capture multimodal disagreement in genuinely ambiguous narratives ($\rho = 0.58$ for $\sigma \geq 1.5$).

Although our system ranks in the middle tier ($\rho = 0.67$), it provides a computationally efficient baseline that clarifies the boundaries of unadapted LLMs on subjective semantic tasks.

2 Task Description

SemEval-2026 Task 5 requires systems to predict ordinal plausibility ratings (1–5) for specific word

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1469–1476

senses in ambiguous stories, given a precontext, target sentence, target sense definition, example sentence, and an optional story ending. The dataset includes 2,280 training and 588 test items.

The primary challenges are fine-grained discrimination between competing senses, variable context lengths (endings are absent in $\approx 33\%$ of cases), and high human subjectivity (σ up to 2.0). Systems are ranked using the arithmetic mean of Spearman's rank correlation (ρ) and accuracy within one standard deviation of the human mean.

3 Related Work

Our approach intersects four areas in NLP: LLM semantic evaluation, reasoning, adaptive computation, and calibration.

3.1 LLMs and Word Sense Disambiguation

Traditional WSD has transitioned from dedicated supervised encoders to generative prompting. In previous SemEval challenges, static architectures (e.g., CNNs/LSTMs or fine-tuned seq2seq models) consistently underperformed LLM-based approaches (Morillo et al., 2024; Peña Gnecco et al., 2025). Recent evaluations corroborate this trend, though LLMs still struggle with low-frequency senses (Meconi et al., 2025) and hallucinate meanings when context is insufficient (Basile et al., 2025). We address these limitations by ensembling diverse open-weight architectures (Llama-3, Qwen-2.5, Gemma-2) to capture complementary semantic priors.

3.2 Chain-of-Thought and Reasoning

LLMs often fail to self-correct reasoning without external feedback (Huang et al., 2024). We use Chain-of-Thought (CoT) prompting (Wei et al., 2023) combined with Self-Consistency (Wang et al., 2023) to marginalize over reasoning paths at inference time.

3.3 Adaptive and Efficient Inference

Self-consistency scales computational cost linearly. To align with "Green AI" (Schwartz et al., 2019), recent work proposes adaptive stopping based on statistical confidence (Aggarwal et al., 2023; Li et al., 2024). We introduce *selective self-consistency*, using a lightweight linguistic heuristic (narrative ending length) to bypass multiple sampling for unambiguous items.

3.4 Uncertainty Quantification and Calibration

LLMs remain miscalibrated on ordinal tasks (Kim et al., 2025; Joshi et al., 2025). We operationalize insights from recent calibration research (Kuhn et al., 2023; Lin et al., 2024) via post-hoc isotonic regression, a computationally inexpensive method that maps overconfident predictions to human-like distributions.

4 System Overview

Our four-stage pipeline balances LLM reasoning capabilities with computational efficiency and calibrated outputs (Figure 1). The core design principle is *selective complexity*: applying expensive stochastic techniques exclusively to high-ambiguity instances.

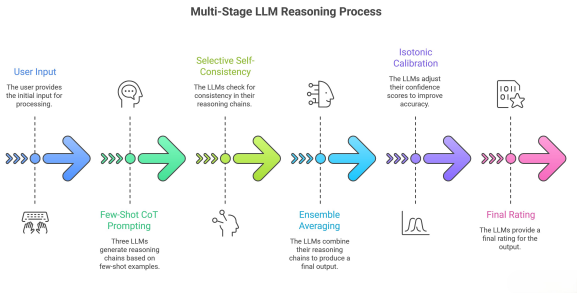

Figure 1: Overview of the proposed four-stage inference pipeline: (1) few-shot chain-of-thought (CoT) prompting with multiple LLMs, (2) selective self-consistency based on input ambiguity, (3) ensemble averaging, and (4) isotonic calibration to produce the final plausibility rating (1–5).

4.1 Base Architecture and Model Selection

We construct an ensemble of three open-weight LLMs accessed via Ollama (Ollama Inc., 2026): Llama-3.1 70B (Grattafiori et al., 2024), Qwen-2.5 32B (Qwen et al., 2025), and Gemma-2 27B (Team et al., 2024). Inter-model agreement on the development set is 75% for items with explicit endings and 55% for items without endings, confirming complementary semantic priors. We aggregate predictions via simple arithmetic averaging, which proved more robust against overfitting on limited training data than weighted alternatives (Jia et al., 2024).

4.2 Few-Shot Chain-of-Thought Prompting

We use few-shot CoT prompting (Wei et al., 2023) with three curated examples anchoring the rating scale: a perfect fit (Rating 5) with explicit lexical confirmation, an implausible fit (Rating 2) with contradictory evidence, and an ambiguous case (Rating 3) lacking a definitive ending. The full prompt structure is provided in Appendix A.

We chose few-shot over zero-shot prompting after observing that zero-shot runs systematically collapsed predictions toward the scale extremes (ratings 1 or 5), failing to produce calibrated mid-range scores. Providing three anchoring examples with explicit reasoning chains mitigated this tendency and yielded more uniform coverage of the 1–5 scale before calibration.

The CoT structure forces the model to sequentially evaluate ending support, precontext expectations, and annotator variability before outputting a single integer. This normalizes uncertainty in genuinely ambiguous cases and counteracts the LLM tendency toward confident, extreme predictions.

4.3 Selective Self-Consistency

Ambiguity Heuristic: An item x is classified as ambiguous if $\mathcal{A}(x) = \mathbb{1}[|\text{ending}(x)| = 0] \vee \mathbb{1}[|\mathcal{W}(\text{ending}(x))| < \tau]$, where $\mathcal{W}(\cdot)$ extracts the set of words and $\tau = 6$, selected after comparing $\tau \in \{3, 6, 9\}$ on a held-out validation split. This heuristic operationalizes the observation that absent or brief endings lack sufficient disambiguating context: a narrative ending provides the strongest signal for resolving sense plausibility, so its absence or brevity is a reliable, label-free, O(1) proxy for expected annotator disagreement. Items classified as ambiguous exhibit significantly higher human annotation variance ($\sigma = 1.42$) compared to clear items ($\sigma = 0.91$, $p < 0.001$).

Sampling Strategy: For ambiguous items ($\approx 33\%$ of the dataset), we sample $k = 3$ predictions per model at $T_{\text{amb}} = 0.5$ and average the ratings. For clear items, we generate a single deterministic prediction at $T_{\text{clear}} = 0.1$. This reduces total inference calls by $\approx$33% compared to uniform sampling while improving Spearman correlation on high-variance items (+0.05). The aggregate effect on overall Spearman is neutral ($\Delta = 0.000$; see Section 6), as the benefit concentrates in the high-variance subset, but the efficiency gain is global.

4.4 Isotonic Regression Calibration

To correct the LLM bias toward extreme ordinal ratings, we apply isotonic regression (Wang, 2025), learning a mapping $\hat{y}_{\text{cal}} = f_{\text{iso}}(\hat{y}_{\text{raw}})$ where $f_{\text{iso}} : \mathbb{R} \to [1, 5]$ is a piecewise constant, non-decreasing function fitted by minimizing squared error under the monotonicity constraint. We fit f_{iso} exclusively on out-of-sample predictions via 5-fold cross-validation to prevent data leakage, using the same fold assignments as the ablation study (detailed in Section 6). At inference, predictions are clipped to $[1, 5]$ and rounded to the nearest integer.

5 Experimental Setup

We validate our approach through ablation studies and hyperparameter optimization on the training set, using 5-fold cross-validation throughout.

5.1 Implementation and Infrastructure

We implemented the pipeline in Python 3.10 using Ollama (Ollama Inc., 2026) for local LLM inference via 4-bit quantization (accessed through its REST API via `requests` 2.32.5), scikit-learn 1.7.2 (Pedregosa et al., 2011) for isotonic regression, numpy 2.3.5 and scipy 1.16.3 for numerical operations, and MLflow 3.6.0 (Zaharia et al., 2018) for experiment tracking, running on a single NVIDIA A100 GPU (40GB VRAM). We used no task-specific fine-tuning libraries (e.g., transformers or langchain); all LLM inference is handled through Ollama's API. The full source code is publicly available at `https://github.com/VerbaNexAI/SemEval2026`.

As shown in Table 1, *selective self-consistency* substantially reduces the computational footprint by generating ≈ 1.66 samples per item on average instead of 3, cutting total inference calls from 25,812 to 14,280 and GPU runtime from over 7 hours to approximately 4 hours across the full dataset.

5.2 Hyperparameter Configuration

We optimized hyperparameters on a held-out validation split (Table 2). For clear items, $T_{\text{clear}} = 0.1$ was selected to minimize decoding variance while avoiding the repetition artifacts of fully greedy ($T = 0$) generation. For ambiguous items, $T_{\text{amb}} = 0.5$ was chosen over higher values ($T \in \{0.7, 1.0\}$) because it provided sufficient output diversity for self-consistency aggregation without producing incoherent or out-of-range ratings. Sampling $k =$

Metric	Uniform SC ($k = 3$)	Selective SC	Savings
LLM Inference Calls	25,812	14,280	$\approx$ **44.6%**
Est. Output Tokens	$\sim$129,000	$\sim$71,400	$\approx$ **44.6%**
Total GPU Time	$\sim$7.2 hours	$\approx$ 4.0 hours	$\sim$**3.2 hours**

Table 1: Computational cost comparison. Selective SC ($k = 1$ for 67%, $k = 3$ for 33% of items) yields substantial savings on a single NVIDIA A100 (40GB) without degrading performance.

Parameter	Value
Models	Llama-3.1 70B, Qwen-2.5 32B, Gemma-2 27B
Inference	Ollama (4-bit quantization)
T_{clear}	0.1
T_{amb}	0.5
Samples (k)	3 (ambiguous items only)
Ambiguity Threshold (τ)	6 tokens
Context Window	4096 tokens
Calibration	Isotonic Regression (5-fold CV)

Table 2: Hyperparameter configuration optimized on the validation set.

3 predictions per ambiguous item gave the best correlation-to-cost trade-off, with only marginal gains for $k = 5$ (+0.004 Spearman) at substantially higher inference cost.

5.3 Baseline and Ablation Configurations

We conducted an ablation study via 5-fold cross-validation on the training set, evaluating five sequential configurations: (1) single model (Llama-3.1 70B, zero-shot), (2) 3-model ensemble (zero-shot), (3) + few-shot CoT, (4) + selective self-consistency ($k = 3$), and (5) + isotonic calibration (full system). We additionally benchmarked uniform SC and selective SC with $k = 5$ to validate our efficiency heuristics, and compared test-set results against the official competition baselines (random assignment and majority class).

5.4 Evaluation Metrics

The primary metric is Spearman's rank correlation (ρ); the secondary is *accuracy within standard deviation* (Accuracy w/in SD), i.e., the percentage of predictions falling within one standard deviation of the mean human rating. Systems are ranked by the combined average of both.

6 Results

Our system achieves a Spearman correlation of 0.67, an accuracy within standard deviation of 0.76, and a combined average of 0.72 on the test set (Table 3), placing 34th out of 79 teams without any task-specific fine-tuning (top 43%). The system significantly outperforms the random (Spearman 0.02) and majority class (Spearman 0.00) baselines, though a substantial gap remains relative to the top-performing systems (top-5 average Spearman 0.84)[1], suggesting a performance ceiling for few-shot prompting approaches versus supervised fine-tuning.

6.1 Ablation Study

Table 4 quantifies each component's contribution via 5-fold cross-validation on the training set.

Ensembling provided the largest single gain (+0.040), followed by isotonic calibration (+0.017). The apparent CoT drop (-0.017) is a methodological confound: CoT and the higher sampling temperature ($T_{\text{amb}} = 0.5$) are coupled in this ablation design and cannot be disentangled; test set results confirm the generalization benefit of the full CoT pipeline. Decoupling these effects (CoT at $T = 0.1$) is left for future work. Uniform self-consistency offered no advantage over selective SC, and $k = 5$ yielded only marginal gains (+0.004), confirming diminishing returns.

6.2 Impact of Human Annotation Variance

Performance degrades sharply with annotation variance (Table 5), revealing a fundamental limitation: averaging stochastic LLM samples implicitly assumes unimodal uncertainty, effectively capturing minor model variation but failing to model the multimodal disagreement (e.g., a polarized split between ratings 2 and 5) characteristic of high-variance instances.

6.3 Validation of Ambiguity Heuristic

Stratifying performance by ending presence validates our ambiguity heuristic (Table 6). Items with explicit endings exhibit lower human variance ($\sigma = 1.08$) and higher performance (Spearman 0.72), while missing endings correlate with higher

[1] Scores from the official SemEval-2026 Task 5 leaderboard (Gehring et al., 2026).

System	Spearman (ρ)	Acc. w/in SD	Combined
Top-1 System	0.86	0.93	0.89
Top-5 Average	0.84	0.92	0.88
Our System	**0.67**	**0.76**	**0.72**
Rank	*34 / 79 (top 43%, no fine-tuning)*		
Majority Class	0.00	0.31	0.16
Random Baseline	0.02	0.20	0.11

Table 3: Official test set performance comparison against benchmarks and top-tier systems (scores from the official leaderboard).

Configuration	Spearman	Δ
Single Model (Llama-3.1)	0.633	–
Ensemble (Zero-shot)	0.673	+0.040
+ Few-shot CoT†	0.656	-0.017
+ Selective SC ($k = 3$)	0.656	+0.000
+ Isotonic Calibration	**0.673**	**+0.017**

Table 4: Ablation results (5-fold CV on training set). †The apparent drop when adding CoT is a confound: CoT is introduced alongside higher temperature ($T_{\text{amb}} = 0.5$), increasing variance before calibration. The net benefit of CoT is confirmed on the test set. $\Delta = 0.000$ for Selective SC reflects a global aggregate; the local improvement on high-variance items is +0.05 (Section 6.2).

Variance Group (σ)	Spearman	Accuracy
Low ($\sigma < 1.0$)	0.78	0.82
Medium ($1.0 \leq \sigma < 1.5$)	0.69	0.72
High ($\sigma \geq 1.5$)	0.58	0.58

Table 5: Performance degradation by annotation variance group.

uncertainty ($\sigma = 1.35$), lower performance (Spearman 0.61), and near-doubled inter-model disagreement (25% to 45%). Items with very brief endings (≤ 6 words) patterned similarly to those without, justifying threshold $\tau = 6$, which achieved the best coverage–precision trade-off across $\tau \in \{3, 6, 9\}$.

Ending Type	Spearman	Disagree (%)
Explicit Ending (> 6 words)	0.72	25%
Brief Ending (≤ 6 words)	0.63	40%
No Ending	0.61	45%

Table 6: Ambiguity heuristic validation by ending type.

6.4 Robustness and Generalization

Two alternative test-set configurations—*Balanced* ($T_{\text{SC}} = 0.6$) and *Ensemble* ($k = 5, T_{\text{SC}} = 0.65$)—both achieved Spearman 0.67 ± 0.001, confirming robustness to moderate hyperparameter perturbations.

7 Discussion

Our results validate two critical design choices. Isotonic calibration delivered the most reliable gain (+0.017 ρ) at negligible cost by compressing the output distribution to correct the inherent LLM bias toward extreme predictions. Selective self-consistency successfully operationalized adaptive computation: concentrating multiple sampling on ambiguous items reduced runtime by ≈45% while improving performance on the hardest cases (+0.05 Spearman). Neither expanding to $k = 5$ nor applying complex ensemble weighting outperformed simple arithmetic averaging, aligning with ensemble literature for constrained datasets (Jia et al., 2024).

The $\rho = 0.17$ gap between our system and the leading approaches suggests a ceiling for unadapted open-weight LLMs on this task. Closing it likely requires supervised fine-tuning, heterogeneous ensembles combining generative and discriminative models, or integration of external lexical resources for task-specific semantics.

8 Future Work

Three directions address the identified limitations in high-variance items:

- **Distributional Prediction:** Modeling $P(y|x)$ using evidential deep learning (Li et al., 2025) or Monte Carlo dropout (Gao et al., 2025) to capture multimodal human disagreement.
- **Stratified Calibration:** Training distinct isotonic calibrators conditioned on predicted input variance for targeted bias correction.
- **Meta-Learning via Stacking:** Employing a meta-model (Ridoy et al., 2024) to predict final ratings from base-model distributions, disagreement statistics, and linguistic features.

9 Conclusion

We presented an efficient pipeline for word sense plausibility rating combining few-shot CoT prompting, a diverse LLM ensemble, selective self-consistency, and isotonic calibration. On SemEval-2026 Task 5, our system achieves Spearman $\rho = 0.67$ and accuracy within SD of 0.76, ranking 34th out of 79 teams without fine-tuning. The key takeaways are: isotonic calibration is important for our approach, aligning LLM outputs with human ordinal distributions; selective self-consistency enables dynamic inference budget allocation, cutting costs by ≈45% while improving performance on the hardest items; and averaging over stochastic samples is insufficient for modeling the multimodal disagreement of genuinely ambiguous narratives ($\rho = 0.58$ for $\sigma \geq 1.5$). We hope this work serves as a transparent baseline and motivates future research into distributional prediction for subjective semantic tasks.

Acknowledgments

To the master's degree scholarship program in engineering at the Universidad Tecnologica de Bolivar (UTB) in Cartagena, Colombia. We express our heartfelt gratitude to the VerbaNex AI Lab team for their unwavering dedication, collaboration, and continuous support throughout our research journey. We also extend our sincere thanks to the organizers of SemEval-2026 Task 5 for designing such a challenging and thought-provoking task, and to the anonymous reviewers for their constructive and insightful feedback.

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. 2023. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12375–12396, Singapore. Association for Computational Linguistics.

Pierpaolo Basile, Lucia Siciliani, Elio Musacchio, and Giovanni Semeraro. 2025. Exploring the word sense disambiguation capabilities of large language models. *Preprint*, arXiv:2503.08662.

Shiqi Gao, Tianxiang Gong, Zijie Lin, Runhua Xu, Haoyi Zhou, and Jianxin Li. 2025. Flue: Streamlined uncertainty estimation for large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):16745–16753.

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.

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.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. *Preprint*, arXiv:2310.01798.

Jianguo Jia, Wen Liang, and Youzhi Liang. 2024. A review of hybrid and ensemble in deep learning for natural language processing. *Preprint*, arXiv:2312.05589.

Abhinav Joshi, Areeb Ahmad, and Ashutosh Modi. 2025. Calibration across layers: Understanding calibration evolution in llms. *Preprint*, arXiv:2511.00280.

Daehwan Kim, Haejun Chung, and Ikbeom Jang. 2025. Calibration of ordinal regression networks. *Preprint*, arXiv:2410.15658.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

Yawei Li, David Rügamer, Bernd Bischl, and Mina Rezaei. 2025. Calibrating llms with information-theoretic evidential deep learning. *Preprint*, arXiv:2502.06351.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. 2024. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. *Preprint*, arXiv:2401.10480.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Domenico Meconi, Simone Stirpe, Federico Martelli, Leonardo Lavalle, and Roberto Navigli. 2025. Do large language models understand word senses? *Preprint*, arXiv:2509.13905.

Anderson Morillo, Daniel Peña, Juan Carlos Martinez-Santos, and Edwin Puertas. 2024. VerbaNexAI lab at

SemEval-2024 task 1: A multilayer artificial intelligence model for semantic relationship detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1344–1350. Association for Computational Linguistics.

Ollama Inc. 2026. Ollama: Run large language models locally. https://ollama.com. Accessed: 2026-02-27.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Daniel Peña Gnecco, Juan Carlos Martinez Santos, and Edwin Puertas. 2025. VerbaNexAI at SemEval-2025 task 2: Enhancing entity-aware translation with Wikidata-enriched MarianMT. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1255–1262. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Shahriyar Zaman Ridoy, Md. Shazzad Hossain Shaon, Alfredo Cuzzocrea, and Mst Shapna Akter. 2024. Enstack: An ensemble stacking framework of large language models for enhanced vulnerability detection in source code. In *2024 IEEE International Conference on Big Data (BigData)*, pages 6356–6364.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green ai. *Preprint*, arXiv:1907.10597.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Cheng Wang. 2025. Calibration in deep learning: A survey of the state-of-the-art. *Preprint*, arXiv:2308.01222.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Matei A. Zaharia, Andrew Chen, Aaron Davidson, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, Fen Xie, and Corey Zumar. 2018. Accelerating the Machine Learning Lifecycle with MLflow. *IEEE Data Eng. Bull.*, 41:39–45.

A Few-Shot Chain-of-Thought Prompt Structure

The prompt follows a fixed structure consisting of four components, filled at inference time from each instance's fields.

(1) Role and task instruction. A brief system-level description establishing the model as a word sense disambiguation expert, defining the ordinal rating scale (1–5) with a one-line gloss for each value.

(2) Three in-context examples. Each example is drawn from the training set and presents the full instance fields (story precontext, ambiguous sentence, ending if available, target sense definition, and example sentence), followed by a step-by-step *Analysis* field and a final *Rating*. The three examples anchor the scale extremes and midpoint:

- **Example 1 (Rating 5):** an instance with explicit lexical confirmation of the target sense in the story ending.
- **Example 2 (Rating 2):** an instance where the ending directly contradicts the target sense.
- **Example 3 (Rating 3):** an instance with no ending, where precontext provides only inconclusive evidence.

(3) Chain-of-thought reasoning scaffold. Before outputting the rating, the model reasons through three sequential questions: (i) Does the ending, if present, support, contradict, or ignore the target sense? (ii) Does the precontext establish strong expectations toward this sense? (iii) Considering the full narrative and plausible annotator variability, what is the most defensible rating?

(4) Target instance and output constraint. Instance fields are inserted via template variables (`{precontext}`, `{sentence}`, `{meaning}`, `{example}`). The `{ending_part}` field renders as `Ending: {ending}` when present and is omitted otherwise. The prompt concludes with an explicit

instruction to output only a single digit (1–5), enforcing the integer generation constraint described in Section 5.

For ambiguous items (Section 4.3), this prompt is issued $k = 3$ times at $T_{\text{amb}} = 0.5$ and the resulting ratings are averaged; for clear items it is issued once at $T_{\text{clear}} = 0.1$.