CUCLASIC at SemEval-2026 Task 5: LLM Prompting Strategies for Rating Ambiguous Word Senses

Federico Ortega Riba, Jasper Wilkerson, Kelsey LaFreniere Adams
University of Colorado Boulder, Departments of Computer Science and Linguistics
{federico.ortegariba, jasper.wilkerson, kelsey.lafreniereadams}@colorado.edu

Abstract

Word sense disambiguation has been a foundational task in computational semantics since the 1990s, but remains an unsolved problem when it comes to bridging human and computational evaluation of ambiguity. The SemEval-2026 Task 5 attempts to address this gap. We test six Large Language Models (LLMs) from the Llama and Gemini families in order to evaluate LLMs' ratings of ambiguous textual excerpts, experimenting with zero- and few-shot variants of prompts and analyzing how simple linguistic cues improve performance. We propose a methodology of eliciting human-like ratings from language models by using examples with low and high standard deviations between human ratings. We further evaluate and compare the prediction patterns of different models and how they align with the human generated ratings. Our best model (Gemini 3-Flash) achieves a 75% score combining Spearman correlation and accuracy within one standard deviation.

1 Introduction

Word sense disambiguation (WSD) has been a foundational task in computational semantics since the 1990s (Veronis and Ide, 1990; Krovetz and Croft, 1989). While not widely pursued at present, WSD remains an open challenge (Basile et al., 2025), largely because word senses can be highly ambiguous and context-dependent.

Word sense disambiguation has historically been treated as a binary classification task. Previous benchmarks like Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 Task 1 (Mihalcea et al., 2004), SemEval-2007 Task 17 (Pradhan et al., 2007), and multilingual expansions like SemEval-2013 Task 12 (Navigli et al., 2013), SemEval-2015 Task 13 (Moro and Navigli, 2015), and SemEval-2021 Task 2 (Martelli et al., 2021) have generally operated on an "either/or" assumption. The most recent visual WSD challenge in SemEval 2023 Task 1 (Raganato et al., 2023) maintained this discrete categorization. However, this dichotomy is often far from reality in natural language understanding. In many instances, there is even a disagreement among human raters (Mihalcea et al., 2004), or environments in which either sense of the word is equally plausible. Therefore, it is necessary that effective WSD models take into account subjective evaluations of word sense disambiguation tasks, without the assumption that only one sense can be correct.

The SemEval-2026 Task 5 (Gehring et al., 2026) attempts to address this issue by introducing a dataset focused primarily on ambiguous word senses in the context of longer narratives in English. Unlike prior binary WSD tasks, this challenge requires systems to predict the human rated plausibility of word senses, particularly in narratives where two senses of the words are semantically plausible.

In this system description paper, we address the narrative-dependence of word sense by proposing prompting techniques with several LLMs from the Llama and Gemini families. Our goal is to analyze which linguistic cues might help increase accuracy and alignment with actual human Likert scores. To study this, we first conduct a systematic evaluation on the development set with different prompting settings that include either edge case examples or almost-perfect agreement examples that vary in homonym and ambiguity level, and we deploy the best prompting strategy using the test set. Overall, our findings suggest that LLMs show two distinct trends: overconfidence with scores that favor 1's and 5's, or moderation with scores that favor 2's and 4's. We conclude that examples with better human inter-annotator agreement are more helpful than edge cases in in-context learning, and we hypothesize that LLMs struggle with uncertainty, which makes them default to boundary scores. In the final task ranking, we reached a score of 75% with an overall Spearman correlation of 73% and an accuracy of 77%. Full prompts and code can be

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 886–893

found on our GitHub.[1]

2 Background

2.1 Task Description

All of our data comes from the AmbiStory dataset, published by the task organizers. The narratives in the data consist of three parts: a precontext with three sentences that ground the story, an ambiguous sentence containing a homonym that causes it to have two different plausible interpretations, and optionally one of two endings, which often imply a specific word sense of the homonym.

The narratives are accompanied by a judged meaning (e.g., for the word "driving", a judged meaning may be "operate or control a vehicle"), and a clear example sentence illustrating the sense of the word. Each entry also has 5 human ratings on a 1-5 Likert scale as to how "plausible" each definition is for the given narrative, along with accompanying average and standard deviation statistics. Score distributions (Figure 1) follow a bounded discrete ordinal distribution.

The task uses two metrics: accuracy (within one standard deviation) and Spearman correlation. These metrics are averaged to get the overall evaluation metric.

2.2 Related Work

Since 2001 with the first shared task on WSD, a wide range of approaches has been proposed, from more traditional machine learning frameworks to modern LLM architectures. For example, Basile et al. (2025), aiming to revive the use of LLMs in WSD tasks, robustly evaluate LLM performance in a zero-shot environment compared with finetuning. Importantly, their paper serves as a baseline of expectations for our own choice of initial zero-shot performance. While Basile et al. (2025) also focus heavily on open LLMs' performance on a multilingual corpus dataset, we limit our own work to English for simplicity and adherence to the SemEval task. The novelty of our work lies in incorporating two other prompting scenarios for WSD, including 3- and 5-shot prompts with examples from the training data, which was not studied by Basile et al. (2025).

Ming et al. (2025), acknowledging a tendency for LLMs to perform poorly on WSD when the disambiguation in question requires higher semantic knowledge and precision, approach the use of LLMs for word sense disambiguation from a knowledge distillation framework. They leverage (several versions of) Llama 3.1 to generate a silver-standard dataset on which to train a model capable of WSD. Because the SemEval Task 5 dataset was not designed to maximize semantic precision, we aim for a simpler approach.

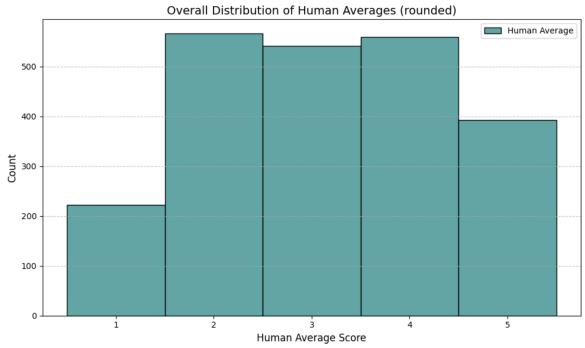

Figure 1: The distribution of human average Likert scores in the AmbiStory training set (2280 examples), rounded to the nearest whole number.

Sumanathilaka et al. (2024) attempt to evaluate numerous proprietary and open-source LLMs' performance in WSD, exploring one-shot, few-shot, and chain-of-thought prompting. Similarly to their approach, our study evaluates several open-source LLMs with zero- and few-shot prompting, though we focus on Likert score correlation and not part-of-speech tagging. Additionally, their comparison of various Llama models (2-70B and 3-70B) versus GPT (4 Turbo, 3.5 Turbo) and Gemini models serve as an interesting backdrop for our own findings of current Llama and Gemini models' performance.

Over the years, as more SemEval tasks have been published, researchers have realized that ambiguity has more than one dimension: ambiguity is both contextual and lexical. In order to address both, Moro and Navigli (2015) disambiguate between interpretations of a whole sequence. For this task, they annotate each sense on a Likert scale, similar to our current task. The authors point out that the inter-annotator agreement is poor, revealing that more context would be needed for consistent interpretations even across human annotators. They use the Jaccard index for sense labels of an instance and other cluster comparison metrics (Fuzzy NMI and Fuzzy B-Cubed). Our paper differs from this in its ranking strategy, since we use Spearman and accuracy within one standard deviation.

[1] https://github.com/fede-ortega/SemEval-2026-SharedTask-5

3 System Overview

Our prompts were designed following Google's Gemini for Google Workspace Prompting Guide 101,[2] which emphasizes specifying a persona, an explicit task, sufficient context, and a constrained output format for the response. Specifically, we cast the model as an expert linguist for WSD and provided the minimum task-relevant inputs (`homonym`, `context`, `ending`, `sentence` and `judged_meaning`), while constraining the output format to a single Likert-style integer in the expected .json format. An example of what this base user prompt looks like is the following:

```
Give a Likert score from 1 to 5 for how well
the homonym '{homonym}' matches the judged
meaning '{judged_meaning}' in this
sentence: '{sentence}'.
Base your answer on this ending as well: {ending}.
Return ONLY this JSON object (no extra
text): '{{"id": "{id_}", "prediction":
<1-5 integer>}}'
```

We also evaluated three prompting scenarios with two LLM families: Llama and Gemini. Because Llama is an open-weights model, it allows for remarkable reproducibility. This also allows for deeper customization (fine-tuning or adapters), although our evaluation only focuses on optimal prompting strategies with linguistic cues and does not cover fine-tuning. On the other hand, Gemini has a strong managed performance and rapid iterations, which set it apart as one of the most powerful LLMs available today (Minaee et al., 2025). Our LLM choice was also motivated by the cost of reproducing the same experiments. In this sense, our evaluation also assesses how cost-efficient a model with a freely accessible API – such as Llama – is, compared to another model with a higher API cost, such as Gemini. Per-token pricing for the Gemini models is public on their model card's website. Anecdotally, Gemini 3-Flash had the highest end-to-end latency in our runs, which we attribute to its internal reasoning trace, while `Gemini 3.0-Pro` returned slightly faster despite being the larger model. However, we did not log per-call token counts systematically.

In our three prompting strategies, the zero-shot prompt served as a baseline: it only instructed the model to output a Likert score from 1–5, corresponding to the degree of fit between a judged meaning and a sentence with in-context usage, as in the example above. This initial evaluation reflects how well the model can infer the task from instructions alone.

To leverage in-context learning, we then tested 3-shot and 5-shot variants, consistent with evidence that a small number of well-chosen demonstrations can teach response patterns and stabilize formatting (Brown et al., 2020) and with Google's guidance that few-shot examples often improve reliability and enforce a desired output pattern.

The 3-shot prompt included one example for the scores that had lower precision with respect to the gold standard, which were scores 1 to 3. Once that was defined, two different 3-shot versions were created, with examples that showed either higher or lower standard deviations from the training set. The intuition behind this, for cases where standard deviation was high, was to expose the model to edge cases to better define decision boundaries; however, the evaluation suggested that examples with the lowest standard deviations (or higher agreement between annotators) helped the model disambiguate better. Furthermore, the 5-shot prompt provided one example for each Likert score (1–5), explicitly establishing the entire scale so that each ordinal value had a concrete interpretation and clear instructions of how Likert scores should be measured in the system prompt. This choice also reflects Kibria et al. (2024)'s practical prompting advice to keep examples consistently structured and to experiment with the number of demonstrations while avoiding excessive examples that may overfit the prompt pattern.

4 Experimental Setup

LLMs Used. Our Gemini models are:

- `gemini-2.5-pro`
- `gemini-3-pro-preview`
- `gemini-3-flash-preview`

Our Llama4 and Llama3 models are:

- `Maverick-17B-128E-Instruct-FP8`
- `Scout-17B-16E-Instruct-FP8`
- `3.3-70B-Instruct`

Implementation. Our runs were made using the respective models' APIs. All prompts were run with the full three-sentence precontext, the ambiguous sentence, and (when available) the ending.

[2] https://workspace.google.com/learning/content/gemini-prompt-guide (Accessed May 1, 2026)

Temperature was set to 0.0 across all runs to enforce deterministic outputs and isolate the effect of the prompting strategy. The behavior of these models at $T > 0$ on a Likert task is therefore not characterized here.

System and user prompts are defined and can be found on our GitHub. The system prompts define the persona, general WSD task definition and Likert score scales as follows:

```
SYSTEM_PROMPT = You are an expert linguist in
word sense disambiguation. Word sense disambigua-
tion tasks commonly assume one word sense to be
the 'correct'one, but that is not necessarily
reflective of reality.
Ambiguities, underspecification and personal
opinions can influence which word senses one finds
plausible in a given context, and there is a
difference between the intuition of humans and the
predictions of computational models.
Our stories consist of three parts: A precontext,
consisting of three sentences that ground the story,
an ambiguous sentence, containing a homonym that
causes it to have two widely different plausible
interpretations, and optionally one of two endings,
which often imply a specific word sense of the
homonym.
The Likert score scale is as follows:
1 = completely unrelated / wrong sense
2 = weakly related / unlikely
3 = plausible but uncertain
4 = strongly related / likely
5 = clearly the intended meaning
```

The user prompts defined the specific WSD style for this shared task as well as the context (few-shots). The homonyms present in the few-shot examples were *drive*, *inflation*, *foggiest*, *heated* and *count*. Structured output was enforced to match the required task format. The rest of the hyperparameters were set to default.

5 Results

The initial evaluation with the development set can be found in Table 1 below. We report the task metrics and specify the best prompt strategy in the `Shot` column. The complete breakdown across all zero-, 3-, and 5-shot variants is provided in Appendix A.

For experiments with the test set, we used the best prompt strategy overall: 5-shot. Results for this set are shown in Table 2.

6 Prediction Analysis

An examination of the prediction distributions reveals major differences between human rater and model behavior. While the human-generated scores followed a bounded discrete ordinal distribution

Model	Shot	Spearman	Acc.
Llama			
`Maverick`	0	62.36	66.33
`Scout`	5	54.48	60.88
`3.3 70B`	0	61.48	60.03
Gemini			
`2.5-pro`	5	64.80	61.39
`3.0-Pro`	5	68.38	65.48
`3-Flash`	**5**	**70.82**	**75.85**

Table 1: SemEval evaluation metrics for the best-performing prompting strategy for each model tested in the dev set (588 examples).

Model	Shot	Spearman	Acc.
Llama			
`Maverick`	5	57.32	57.63
`Scout`	5	55.79	61.29
`3.3 70B`	5	61.49	63.65
Gemini			
`2.5-pro`	5	69.96	67.31
`3.0-Pro`	5	67.37	66.55
`3-Flash`	**5**	**72.7**	**76.77**

Table 2: SemEval evaluation metrics for the best-performing prompting strategy for each model tested in the test set (930 examples).

(Figure 1), the evaluated models failed to reproduce this distribution. One of the most consistent patterns replicated across model families was the near collapse of predictions of a score of 3. This hesitation to predict neutral values was the primary contributor to the relatively poor performance of the tested models on this task.

Both model families tested overwhelmingly skewed right, indicating a systemic bias towards overconfidence. This aligns with recent research demonstrating that instruction-tuned language models exhibit systemic overconfidence (Sun et al., 2025) and a positivity or sycophancy bias, frequently defaulting to affirmative (Sharma et al., 2023). While Llama family models defaulted to predictions of 5's across all tested models, Gemini family models predicted a generally bimodal distribution of 1's or 5's. The only exception was the leading model, Gemini 3-Flash, whose predictions approximated a distribution much closer to the human baseline (Figure 2), although still suffering from the aforementioned deficit in "neutral" predictions. This is better exhibited when looking at samples with high standard deviations. For these

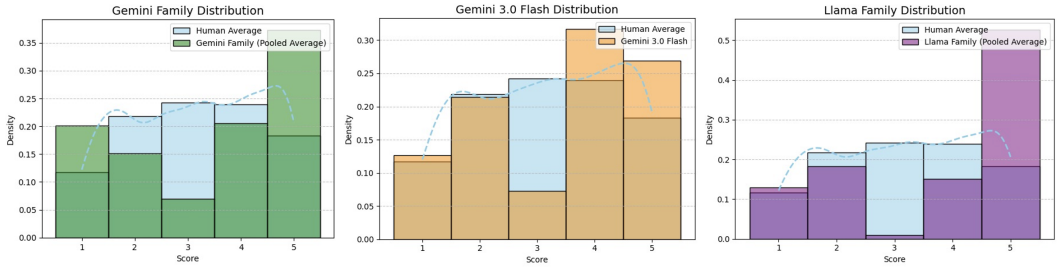

Figure 2: Prediction distributions for the Gemini family (left), our best-performing model (center) and the Llama family (right).

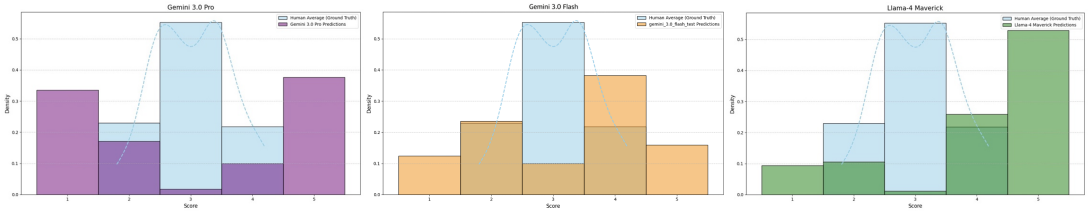

Figure 3: Standard deviations for Gemini 3.0 Pro (left), our best-performing model (center) and Llama4 Maverick (right).

"contentious" examples, one might expect a model to take a middling stance. However, as seen in Figure 3, these models were reluctant to predict middle values, and accuracy drops accordingly on high-SD items across all six models (Figure 10). By contrast, on low-SD items (<0.5), all models track the human distribution more closely, though the Llama family retains its skew toward "5" (Figures 4-9).

The Llama models almost always predicted "5", even in examples with high human disagreement. The Gemini models (with the exception of 3-Flash) tended to score near endpoints, predicting 1's and 5's for samples with high disagreement. Only the Gemini 3-Flash model managed to pick up on the subjectivity of some of the questions and predicted middle values like 2 and 4. Even still, Gemini 3-Flash was hesitant to predict 3's and still chose a side, but with slightly less confidence than other Gemini family models. This more conservative approach enabled the Gemini 3-Flash model to correctly predict more examples where the human average was ≈3, which happened to be most of the samples in the dataset.

The breakdown of task metrics in Table 3 reveals that even the few 3 predictions that the model *did* make were less accurate than its predictions on the extremes of the scale. However, the increased accuracy in the predictions of 1 and 2 are sufficient to balance this deficiency. While perhaps initially unintuitive, it is not surprising that the 3.0-Pro model performs worse than the 3-Flash model. Degradation in performance from smaller to larger models has been well documented, known as inverse scaling (Stringli et al., 2025). Models with more reinforcement learning from human feedback (RLHF) are often punished more for uncertainty and learn overconfidence. In a task requiring models to give a large amount of uncertain answers, we hypothesize that this overconfidence greatly impacts the Pro models and that it is the Flash model's willingness to give answers of 2, 3, and 4 that allows it to succeed in this task.

Guesses	Gemini 3-Flash Accuracy
1	0.81
2	0.80
3	0.69
4	0.75
5	0.76

Table 3: The accuracy of predictions by Gemini 3-Flash, within one standard deviation.

7 Conclusion

This research evaluated the performance of LLMs on the SemEval-2026 Task 5, a continuous scaled WSD task. Using zero- and few-shot prompting strategies across the Llama and Gemini model families, we assessed the capability of LLMs to rate subjective, ambiguous narratives in alignment with

human annotations. We established a methodology for eliciting ratings for ambiguous narratives by using examples of low standard deviations as guiding examples for the LLMs and developing sufficiently detailed user prompts to outline this complex task. We experimented with examples demonstrating edge cases in tandem with examples with almost perfect agreement to assess the role of linguistic cues in WSD.

Our results suggest that current LLMs are consistently overconfident in their predictions; only Gemini 3-Flash's predictions approached a distribution closer to the human ratings, achieving an overall evaluation score of 0.75 combining Spearman correlation and accuracy within one standard deviation.

Limitations and Future Work

Prompt component ablations. We did not run ablations removing individual prompt fields such as `judged_meaning` or `homonym`, leaving open the question of which components carry the most signal. To our knowledge, prompt-component ablation for WSD is not addressed in prior work, so the effect of removing these fields on Likert calibration is underexplored.

Why low-SD demonstrations help. We hypothesize that low-SD examples help because their labels are less noisy and provide a cleaner mapping from narrative cues to Likert scores, whereas high-SD examples conflate ambiguity with annotator disagreement. We did not, however, conduct a controlled quantitative comparison matching low-SD and high-SD demonstrations on ambiguity level (e.g., homonym frequency, sense distance). A follow-up study would clarify whether the gain comes from label noise, item difficulty, or both.

Calibration toward neutral scores. Our analysis shows models systematically avoid predicting 3, which is the modal human label. Future work could mitigate this by (i) adjusting in-context examples to over-represent score-3 cases or (ii) adding explicit decision-rule instructions tying high inter-annotator disagreement to a score of 3.

Per-category analysis. Tables 2 and 4 report aggregate metrics only. A finer-grained breakdown by gold Likert score, by homonym, or by part-of-speech is left for future work.

Acknowledgments

The research that led to this publication was conducted with the support of a US-Spain Fulbright grant.

References

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France. 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.

Raihan Kibria, Sheikh Intiser Uddin Dipta, and Muhammad Abdullah Adnan. 2024. On functional competence of LLMs for linguistic disambiguation. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 143–160, Miami, FL, USA. Association for Computational Linguistics.

Robert Krovetz and W Bruce Croft. 1989. Word sense disambiguation using machine-readable dictionaries. In *Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 127–136.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The senseval-3 English lexical sample task. In *Proceedings of SENSEVAL-3, the Third*

International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pages 25–28, Barcelona, Spain. Association for Computational Linguistics.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large language models: A survey. *Preprint*, arXiv:2402.06196.

Liqiang Ming, Sheng-hua Zhong, and Luncong Li. 2025. Towards general-domain word sense disambiguation: Distilling large language model into compact disambiguator. In *Conference on Empirical Methods in Natural Language Processing*, pages 884–897. Association for Computational Linguistics.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. SemEval-2023 task 1: Visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2227–2234, Toronto, Canada. Association for Computational Linguistics.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Elena Stringli, Maria Lymperaiou, Giorgos Filandrianos, Athanasios Voulodimos, and Giorgos Stamou. 2025. Pitfalls of scale: Investigating the inverse task of redefinition in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9445–9469.

T.G.D.K Sumanathilaka, Nicholas Micallef, and Julian Hough. 2024. Can llms assist with ambiguity? a quantitative evaluation of various large language models on word sense disambiguation. In *1st International Conference on NLP AI for Cyber Security*, pages 97–108. Association for Computational Linguistics.

Fengfei Sun, Ningke Li, Kailong Wang, and Lorenz Goette. 2025. Large language models are overconfident and amplify human bias. *arXiv preprint arXiv:2505.02151*.

Jean Veronis and Nancy Ide. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

A Complete Development Set Results

Table 4 reports the full breakdown of zero-, 3-, and 5-shot prompting strategies on the development set across all evaluated models. These results motivate our selection of 5-shot as the prompting strategy for the test set submission: 5-shot achieves the highest combined score on the strongest models, while Llama models show variable behavior across shots. `Gemini 3.0-Pro` was only tested with the 5-shot template due to API query limits.

Model	Shot	Spearman	Acc.
Llama			
`Maverick`	0	62.36	66.33
`Maverick`	3	57.36	65.31
`Maverick`	5	59.77	61.06
`Scout`	0	50.73	57.82
`Scout`	3	50.53	62.59
`Scout`	5	54.48	60.88
`3.3 70B`	0	61.48	60.03
`3.3 70B`	3	56.10	60.54
`3.3 70B`	5	55.70	59.52
Gemini			
`2.5-pro`	0	60.18	53.74
`2.5-pro`	3	63.86	60.03
`2.5-pro`	5	64.80	61.39
`3.0-Pro`	5	68.38	65.48
`3-Flash`	0	68.47	60.37
`3-Flash`	3	68.11	66.67
`3-Flash`	**5**	**70.82**	**75.85**

Table 4: Complete results across all prompting strategies for each model on the dev set (588 examples).

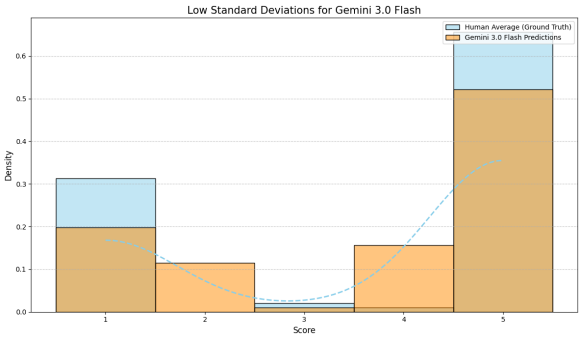

Figure 4: Low Standard Deviation (<0.5) Examples for Gemini 3-Flash

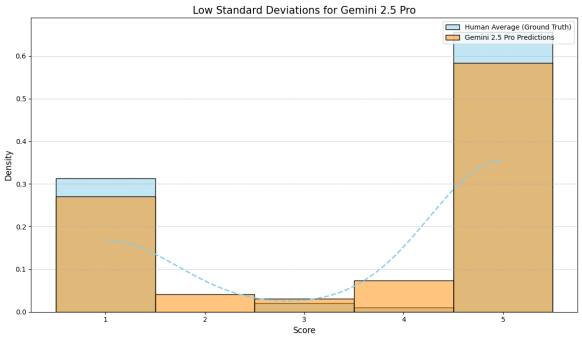

Figure 5: Low Standard Deviation (<0.5) Examples for Gemini 2.5-Pro

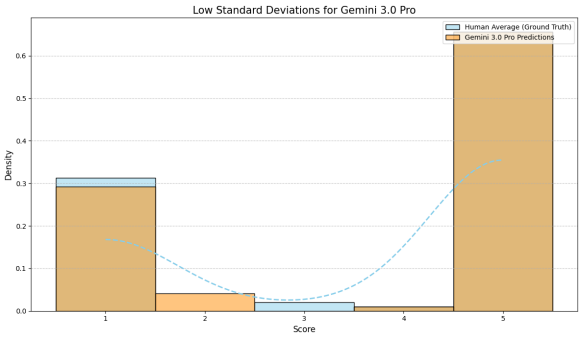

Figure 6: Low Standard Deviation (<0.5) Examples for Gemini 3.0-Pro

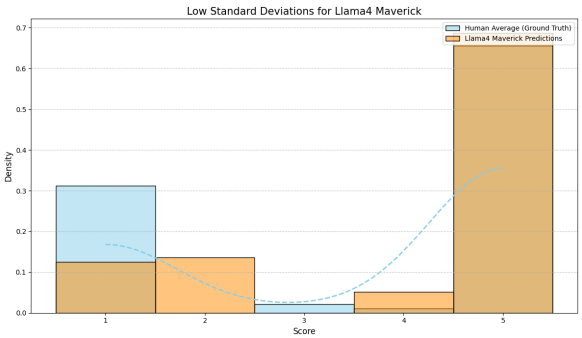

Figure 7: Low Standard Deviation (<0.5) Examples for Llama4 Maverick

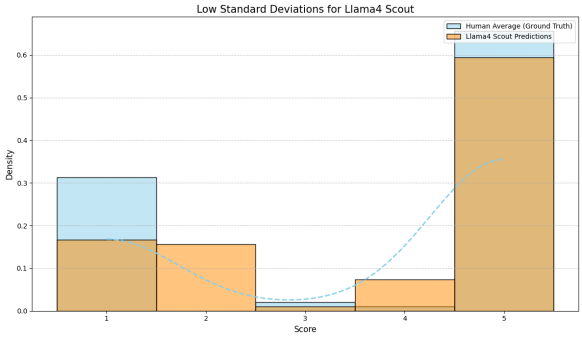

Figure 8: Low Standard Deviation (<0.5) Examples for Llama4 Scout

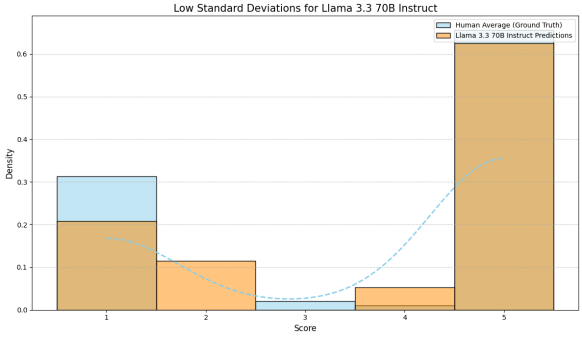

Figure 9: Low Standard Deviation (<0.5) Examples for Llama 3.3 70B

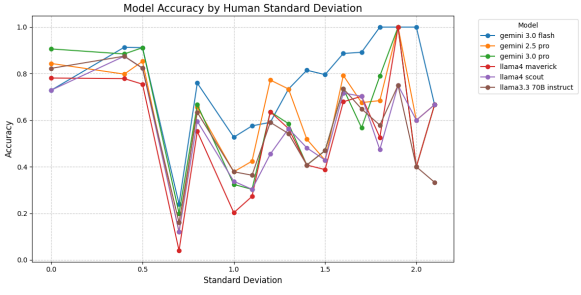

Figure 10: Accuracy by Standard Deviation (Binned by 0.1)