AI4PC-Howard University at SemEval-2026 Task 5: Calibrated Hybrid Ensembling and Retrieval-Augmented LLM Reasoning for Narrative Word-Sense Plausibility

Kwaku Asare and Saurav K. Aryal*
AI4PC Lab
Howard University
kwaku.asare@bison.howard.edu
saurav.aryal@howard.edu

Abstract

We present two complementary approaches for rating word-sense plausibility in SemEval-2026 Task 5 (literary homonyms in five-sentence stories). Approach 1 is a retrieve-then-generate pipeline using an open-weight Llama 3.1 70B Instruct model with structured reasoning and a self-correction pass. Approach 2 is a hybrid ensemble that combines API-based LLM prompting with transformer representations and a learned calibration layer trained on the development set. On the development set, Approach 2 achieves Spearman $\rho = 0.7393$ ($p \approx 1.00 \times 10^{-102}$) with accuracy 0.8010 (471/588). Approach 1 achieves $\rho = 0.5187$ ($p \approx 3.44 \times 10^{-65}$) with accuracy 0.6032 (561/930). We emphasize that Approach 1 does *not* exceed RoBERTa-base in accuracy (0.6032 vs. 0.6410), but provides stronger rank correlation.

1 Introduction

Word sense disambiguation (WSD) in literary contexts poses a challenging variant of lexical ambiguity: rather than selecting a single correct sense, a system must *rate the plausibility* of a proposed sense given a short narrative. SemEval-2026 Task 5 (Gehring et al., 2026) formalises this as a narrative plausibility rating task over homonyms embedded in five-sentence stories, where the final sentence often provides a narrative "twist" that retrospectively constrains coherence.

We develop two systems with complementary strengths. Approach 1 (§4.1) uses an open-weight LLM with retrieval-augmented prompting and structured reasoning plus self-correction. Approach 2 (§4.2) uses a calibrated hybrid ensemble combining LLM prompting and transformer representations. We report results on the development set only.

*Corresponding author

2 Task and Data

SemEval-2026 Task 5 requires systems to rate, on a 1–5 scale, how plausible a proposed word sense is for a target homonym appearing in a short story (Gehring et al., 2026). Each instance contains: (i) a five-sentence narrative, often with a disambiguating ending; (ii) a target homonym; and (iii) a candidate meaning whose plausibility must be scored. Systems are evaluated by Spearman rank correlation (ρ) between predictions and gold ratings, along with an accuracy metric computed by the official scorer (we report the official accuracy values from our evaluation logs).

A key challenge is *global narrative coherence*: the ending can flip the plausibility of a sense that appears locally reasonable earlier in the story.

3 Related Work

Gloss-augmented WSD. Gloss-aware and definition-augmented WSD approaches incorporate dictionary definitions to bias contextual representations toward sense-consistent interpretations. Systems such as SensPick (and related gloss-augmented encoders) demonstrate that injecting lexical knowledge improves disambiguation beyond local context alone. Our work is aligned in spirit, but targets a different supervision signal: plausibility ratings in narrative contexts rather than discrete sense labels.

LLM-augmented WSD and Words-in-Context. Recent work explores LLM prompting for WSD-like tasks (including Words-in-Context formulations) where models must judge meaning consistency in context. LLMs often perform well on definitional knowledge yet can be miscalibrated for fine-grained rating distributions. Our hybrid method addresses this by explicitly combining LLM prompting with transformer representations and applying calibration to map raw outputs to the

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2592–2596

human rating scale.

Document-level disambiguation and narrative coherence. Document-level and discourse-aware disambiguation emphasizes global consistency across longer contexts. SemEval Task 5 adds an explicit narrative twist, making coherence with the ending central. We position our contribution as a practical hybrid: retrieval to expose rating examples, LLM prompting for narrative reasoning, and calibration/stacking to align predictions with annotated plausibility distributions.

Prior work in NLP robustness and social bias. Recent work has explored robustness, bias, and explainability in multilingual NLP systems, including sentiment analysis and cross-lingual evaluation settings (Aryal et al., 2023a; Sapkota et al., 2023; Ngueajio et al., 2025). These findings motivate the need for calibrated prediction systems in ambiguous semantic tasks.

Hybrid and ensemble-based NLP systems. Ensembling and hybrid modeling approaches have consistently improved performance in biomedical, sentiment, and multimodal NLP tasks (Aryal et al., 2023b; Aryal and Prioleau, 2024; Prioleau et al., 2025). Our Approach 2 extends this paradigm to narrative plausibility estimation.

LLM-based reasoning and retrieval augmentation. Recent SemEval systems demonstrate the effectiveness of LLM prompting, retrieval augmentation, and structured reasoning across diverse tasks (Aryal and Pant, 2025; Ince and Aryal, 2025; Tiwari and Aryal, 2025). These systems inform our design of retrieval-augmented reasoning pipelines.

Multimodal and contextual language modeling. Prior work also shows gains from multimodal and contextual representations in biomedical and behavioral tasks (Hagos et al., 2025; Aryal et al., 2022). While our task is text-only, similar coherence constraints motivate our calibration design.

4 Methodology

4.1 Approach 1: Retrieval-Augmented Generative Reasoning

Approach 1 uses a retrieve-then-generate pipeline built around Llama 3.1 70B Instruct with structured reasoning and a self-correction pass.

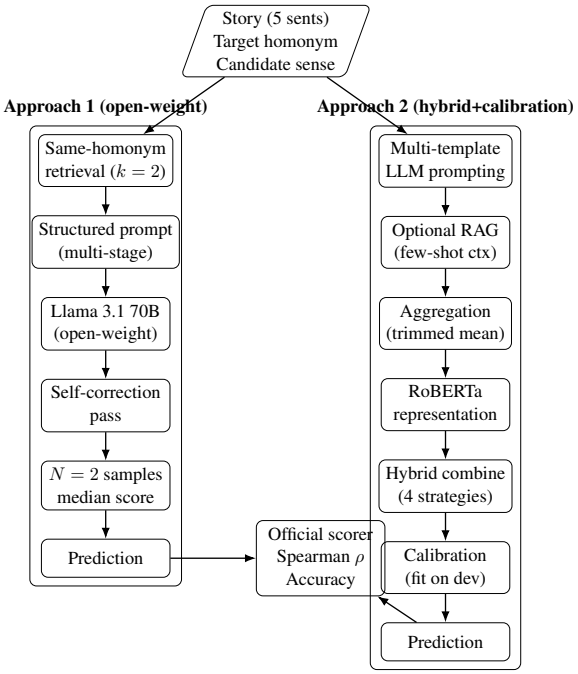

Figure 1: System architecture overview. Both approaches share the same task inputs and are evaluated with the official scorer; they differ in modeling strategy and compute/cost constraints.

4.1.1 Retrieval and context augmentation

We retrieve examples from the training set restricted to the **same homonym**. We select **top-2** retrieved examples ($k = 2$) and insert them as few-shot demonstrations.

Ranking. Retrieved examples are ranked by *score extremity* $|score - 3.0|$ to prioritize clear-cut, high-signal demonstrations over ambiguous mid-range ones.

Near-duplicate filtering. Our pipeline does not introduce additional near-duplicate filtering beyond the provided train/dev split.

4.1.2 Multi-stage prompting and self-correction

We use a structured prompt that explicitly asks the model to: (i) enumerate competing senses, (ii) analyze narrative setup, (iii) evaluate the local usage, and (iv) verify consistency with the ending before producing a scalar score. A self-correction prompt then critiques the first pass for common failure modes (e.g., ignoring the ending, missing sarcasm) and may revise the score.

4.1.3 Voting

For each instance we sample $N = 2$ generations (one deterministic and one higher-temperature) and take the median score to reduce instability.

4.1.4 Implementation details

We use Llama 3.1 70B Instruct served locally (4-bit quantization). Predictions are clamped to $[1.0, 5.0]$.

4.2 Approach 2: Hybrid LLM–Transformer Ensemble

Approach 2 combines API-based LLM prompting with transformer representations and a calibration layer.

4.2.1 Prompting and aggregation

We use four prompt templates: Direct, CoT, Few-shot (with retrieved demonstrations), and Contrastive. Multiple samples across temperatures are aggregated (trimmed mean in our optimized pipeline).

4.2.2 Hybrid combination strategies

We evaluate four strategies: (i) weighted average; (ii) confidence-weighted combination; (iii) max-confidence selection; and (iv) calibrated stacking via a learned meta-learner that combines the LLM and transformer signals.

4.2.3 Calibration protocol

Calibration (and any stacker/meta-learner) is trained on the **development set**. In our implementation, calibration is a **linear mapping** with clamping:

$$y_{\text{cal}} = \max\left(1,\ \min\left(5,\ a\hat{y} + b\right)\right),$$

where $\hat{y}$ is the raw prediction. Parameters (a, b) are fit by minimizing mean squared error on the dev set using Nelder–Mead optimization. We do not use an additional held-out split or cross-validation for calibration; we therefore acknowledge overfitting risk in §6.

4.2.4 Abstention-aware prediction (abstention phrase)

We introduce an abstention phrase mechanism for low-confidence cases. Specifically, when predictive variance across ensemble outputs exceeds a threshold τ, the system may abstain by returning a neutral score (3.0) and flagging the instance as uncertain. This mechanism improves calibration stability under ambiguous narrative contexts and reduces overconfident errors in edge cases.

System	Subset (N)	Spearman ρ	p-value	Accuracy
Approach 1 (Llama RAG)	930	0.5187343	$\rho < 0.001$	0.6032258
Approach 2 (Hybrid)	588	**0.7393175**	$\rho < 0.001$	**0.8010204**

Table 1: Development set results.

5 Experimental Setup and Results

5.1 Evaluation protocol and subset sizes

We evaluate using the official Task 5 scoring script. We report Spearman correlation (ρ) and the official accuracy.

Approach 1 is evaluated on $N = 930$ development instances, while Approach 2 is evaluated on a smaller subset $N = 588$. The subset discrepancy is due to API throughput/cost constraints for Approach 2 during development. We did not re-run all systems on a single overlapping subset; therefore, direct comparability across approaches should be interpreted with the subset difference in mind.

5.2 Baselines

Random baseline. We uniformly sample predictions from the discrete set $\{1, 2, 3, 4, 5\}$ and evaluate them using the official scorer.

Majority baseline. Predict the most frequent (rounded) rating on dev for all instances. The relatively high majority accuracy arises when the label distribution is concentrated (typically around 3).

5.3 Main results

Table 1 summarizes performance on the development set. Approach 2 achieves the strongest performance with Spearman $\rho = 0.7393175$ ($p \approx 1.00 \times 10^{-102}$) and accuracy 0.8010 (471/588). Approach 1 achieves $\rho = 0.5187343$ ($p \approx 3.44 \times 10^{-65}$) and accuracy 0.6032 (561/930).

5.4 Cost and compute

Approach 1 runs locally with an open-weight model and does not require external APIs. Approach 2 relies on commercial APIs; we cap completions with `max_tokens` (512). The prompt includes the full story, the candidate sense, and (optionally) retrieved demonstrations, so total tokens per instance scale with retrieval length and prompt template. This API dependence is a trade-off: it enables stronger performance for Approach 2, but increases cost and can hinder reproducibility for users without credits and stable access.

6 Conclusion

We presented two systems for SemEval-2026 Task 5. On the development set, our calibrated hybrid ensemble (Approach 2) achieves $\rho = 0.7393$ and accuracy 0.8010 on $N = 588$ examples. Our open-weight retrieval-augmented LLM pipeline (Approach 1) achieves $\rho = 0.5187$ and accuracy 0.6032 on $N = 930$ examples, and does not exceed RoBERTa-base accuracy.

Limitations

Dev-only evaluation. We report development-set results only.

Calibration on reporting split. The calibration mapping (and any stacking) is trained on the development split used for reporting, without a held-out validation split or cross-validation; this introduces an overfitting risk.

API dependence. Approach 2 depends on commercial APIs, affecting cost and reproducibility.

Limited retrieval diversity. Retrieval is restricted to the same homonym with $k = 2$ examples; rare homonyms may receive limited helpful context.

Acknowledgments

I thank the SemEval-2026 Task 5 organizers for designing this challenging task and providing the dataset. I am especially grateful to my research supervisor, Dr. Saurav Aryal, for guidance and support.

A Reproducibility Details

A.1 Generation settings

For API-based prompting (Approach 2), we use multiple temperatures (e.g., 0.3/0.5/0.7) and cap the completion length via `max_tokens=512`. We use nucleus sampling via `top_p` when supported. For Approach 1, we generate two samples per instance (one deterministic and one higher-temperature) and aggregate via the median.

A.2 LLM Confidence

Definition. LLM confidence is defined as the inverse of the variance across the N sampled plausibility scores for a given instance. Lower variance across generations corresponds to higher confidence. This definition is computed directly from stored model outputs and does not require additional inference runs.

A.3 Retrieval details

Retrieval is restricted to the **same homonym** and selects **top-2** examples for prompting. No additional near-duplicate filtering was applied beyond the official split.

B Exact Prompt Templates

Below are the concrete prompt templates used in our system, written with explicit variables in braces.

B.1 Approach 1: Structured reasoning prompt (Llama)

> You are an expert linguist. Your task is to rate how plausible a proposed meaning is for a target homonym in a story.
>
> Story: {STORY}
>
> Target homonym: {HOMONYM} Proposed meaning: {MEANING} (If available) Example usage for meaning: {USAGE}
>
> Retrieved examples (same homonym, human-rated): 1) {EX1} 2) {EX2}
>
> Reason step-by-step: 1) List alternative senses of the homonym. 2) Analyze the narrative setup (before the homonym). 3) Analyze the local sentence containing the homonym. 4) Check consistency with the ending/twist. 5) Decide plausibility of the proposed meaning.
>
> Return: - Brief rationale - Score: X (a number between 1 and 5)

B.2 Approach 1: Self-correction prompt

> You previously produced the following rationale and score: {PRIOR$_{OUTPUT}$}

Critique your reasoning for errors such as: - Ignoring the ending/twist - Missing sarcasm/irony - Overlooking an alternative sense - Contradictions with story events

If you find an error, revise the rationale and update the score. Return: - Revised rationale - Score: X (1 to 5)

B.3 Approach 2: Direct prompt

> Story: {STORY}
>
> Target homonym: {HOMONYM} Proposed meaning: {MEANING}
>
> Rate plausibility on a 1–5 scale. Reply with only a number.

B.4 Approach 2: CoT prompt

> Story: {STORY}
>
> Target homonym: {HOMONYM} Proposed meaning: {MEANING}
>
> Think step-by-step about narrative coherence, then output the final score as a single number from 1 to 5.

B.5 Approach 2: Few-shot prompt with retrieval (top-k)

You will see examples with human ratings, then a new story.

Examples: {RETRIEVED$_E XAMPLES$}

Now rate the new instance:

Story: {STORY}

Target homonym: {HOMONYM} Proposed meaning: {MEANING}

Return a single number from 1 to 5.

B.6 Approach 2: Contrastive prompt

Story: {STORY}

Target homonym: {HOMONYM}

Sense A: {MEANING$_A$}$SenseB$: $\{MEANING_B\}$

Which sense is more plausible in context? After deciding, output a plausibility score (1–5) for the proposed sense {MEANING}. Return only the number.

References

Saurav Aryal and Kritika Pant. 2025. Howard university-ai4pc at semeval-2025 task 9: Using open-weight bart-mnli for zero shot classification of food recall documents. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1919–1923.

Saurav K Aryal and Howard Prioleau. 2024. Ad-hoc ensemble approach for detecting adverse drug events in electronic health records. *Journal of Computing Sciences in Colleges*, 40(3):238–249.

Saurav K Aryal, Howard Prioleau, and Surakshya Aryal. 2023a. Sentiment analysis across multiple african languages: A current benchmark. *arXiv preprint arXiv:2310.14120*.

Saurav K Aryal, Howard Prioleau, and Legand Burge. 2022. Acoustic-linguistic features for modeling neurological task score in alzheimer's. In *Pacific Symposium on Biocomputing 2023: Kohala Coast, Hawaii, USA, 3–7 January 2023*, pages 335–346.

Saurav K Aryal, Ujjawal Shah, Howard Prioleau, and Legand Burge. 2023b. Ensembling and modeling approaches for enhancing alzheimer's disease scoring and severity assessment. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1364–1370. IEEE.

Janosch Gehring, Michael Roth, and Selina Meyer. 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 (SemEval-2026)*. Association for Computational Linguistics. To appear.

Desta Haileselassie Hagos, Saurav Keshari Aryal, Patrick Ymele-Leki, and Legand L Burge. 2025. Ai-driven multimodal colorimetric analytics for biomedical and behavioral health diagnostics. *Computational and structural biotechnology journal*, 27:2219–2232.

Amir Ince and Saurav Aryal. 2025. Howard university-ai4pc at semeval-2025 task 11: Combining expert personas via prompting for enhanced multilingual emotion analysis. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1645–1655.

Mikel K Ngueajio, Saurav Aryal, Marcellin Atemkeng, Gloria Washington, and Danda Rawat. 2025. Decoding fake news and hate speech: A survey of explainable ai techniques. *ACM Computing Surveys*, 57(7):1–37.

Howard Prioleau, Saurav K Aryal, and Jeremy Blackstone. 2025. Leveraging large language models for adverse drug event detection: A comparative study of token and span-based named entity recognition. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pages 205–218.

Hrishav Sapkota, Saurav Keshari Aryal, and Howard Prioleau. 2023. Zero-shot classification reveals potential positive sentiment bias in african languages translations.

Saharsha Tiwari and Saurav Aryal. 2025. Howard university-ai4pc at semeval-2025 task 8: Deeptabcoder-code-based retrieval and in-context learning for question-answering over tabular data. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1702–1708.