Calibrated Hybrid Ensembling and Retrieval-Augmented LLM Reasoning Narrative Word-Sense Plausibility

Kwaku Asare, Saurav K. Aryal, PhD

AI4PC Lab, Institute for Human Centered Artificial Intelligence

INTRODUCTION

Word sense disambiguation in literary text is challenging because meaning depends not only on the target word, but also on the full narrative context.

- In SemEval-2026 Task 5, systems must rate how plausible a proposed meaning is for a homonym in a five-sentence story, where the final sentence may introduce a twist that retrospectively changes interpretation. Ratings are given on a 1–5 scale and evaluated by Spearman rank correlation and accuracy.
- To address this, we developed two complementary approaches: (1) a retrieval-augmented open-weight LLM pipeline with structured reasoning and self-correction, and (2) a calibrated hybrid ensemble combining LLM prompting, transformer representations, and a learned calibration layer.

Our work explores how narrative reasoning, retrieval, and calibration can improve plausibility scoring in ambiguous literary contexts, and demonstrates that combining these techniques yields more reliable predictions than any single method alone. Importantly, **our best system achieves Spearman ρ = 0.7393 and 80.1% accuracy on the development set, outperforming a strong RoBERTa-base baseline.**

Task & Dataset

- **Input**: 5-sentence narrative + target homonym + candidate sense
- **Output**: Plausibility score (1–5 scale)
- **Eval**: Spearman ρ + official accuracy

Task Example

STORY

Sarah practiced every day for the big event.
Her fingers moved quickly across the **keys**.
The audience waited in silence.
She took a deep breath and began to play.
The piano filled the room with music.

LOW PLAUSIBILITY
Candidate sense A
Metal objects used to open locks
Score: 1.3 / 5
Plausible before sentence 4, but the ending rules it out.

HIGH PLAUSIBILITY
Candidate sense B
Piano keys
Score: 4.8 / 5
Confirmed by the final sentence — the twist resolves the ambiguity.

Key insight: Sense A seems locally plausible until sentence 5. The model must read the *entire* story — not just the sentence containing the target word.

APPROACH OVERVIEW

We propose two complementary approaches to score word-sense plausibility in narrative text.

- **Approach 1:**
  - retrieves similar examples
  - reasons through the full narrative step-by-step
  - self-corrects before scoring.
- **Approach 2:**
  - aggregates multiple prompt strategies
  - fuses hem with transformer representations,
  - calibrates outputs to match human ratings.

APPROACH OVERVIEW

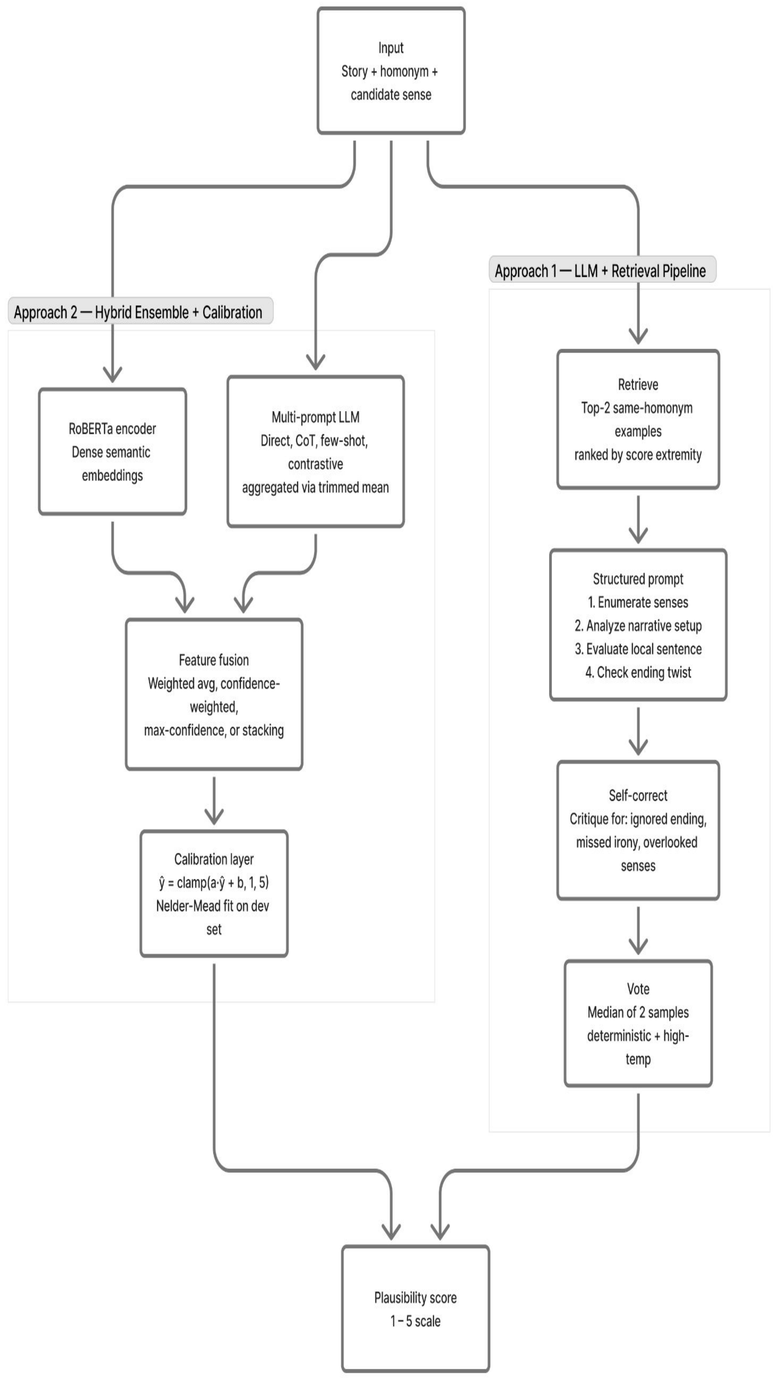

Figure: High-level overview of the two approaches

RESULTS

Both approaches were evaluated on the SemEval-2026 Task 5 development set using the official scoring script, measured by Spearman rank correlation (ρ) and accuracy.

System	N	Spearman ρ	Accuracy
RoBERTa-base	–	–	64.1%
Approach 1 (Llama RAG)	930	0.5187	60.3%
Approach 2 (Hybrid)	588	0.7393	80.1%

CONCLUSION

We presented two systems for rating word-sense plausibility in narrative text, a task where meaning depends on global story context, not just local word usage.

What we found:

- **Hybrid modeling wins:** Approach 2 achieves the strongest results (ρ = 0.7393, 80.1% accuracy), showing that combining LLM reasoning with transformer representations and calibration is more reliable than either alone.
- **Reasoning still matters:** Approach 1 demonstrates that structured narrative reasoning captures rank order well (ρ = 0.5187), even without calibration or transformer fusion.
- **Twist endings are the core challenge**: models that ignore the final sentence consistently misrate plausibility. Global coherence is not optional.

REFERENCES

Gehring, J., Roth, M., & Meyer, S. (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.*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Meta AI. (2024). The Llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

ACKNOWLEDGEMENTS

This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714. The work is solely the responsibility of the authors and does not necessarily represent the official view of the Office of Naval Research.