Team JCT 2026 at SemEval-2026 Task 5: AmbiStory Navigating Narrative Consistency through a Hybrid LLM-NLI Ensemble

Chava Laufer*, **Batel Sara Turjeman*** **and Chaya Liebeskind**
Jerusalem College of Technology
21 Havaad Haleumi St., 91160
Jerusalem, Israel
{chava.laufer, batelsara.turjeman, liebchaya}@gmail.com

Abstract

This paper details Team "JCT 2026" submission to SemEval-2026 Task 5: Rating Plausibility of Word Senses in Ambiguous Sentences. The objective is to predict a continuous plausibility score (1.0–5.0) for target homonyms embedded in narratives of five-sentences. To address the challenges of narrative coherence and semantic precision, we introduce a hybrid ensemble architecture. The system integrates a generative Large Language Model (Llama-3 8B, fine-tuned via LoRA) with a dual-expert bidirectional cross-encoder (DeBERTa-v3-large) optimized for both semantic similarity and Natural Language Inference (NLI) [1]. By aggregating these complementary models, the system effectively captures complex contextual dependencies. In the official test set, our architecture ranked 22nd out of 79 systems, achieving a Spearman Rank Correlation of 0.71 and an accuracy within the standard deviation of 82.04%.

1 Introduction

The evaluation of computational models is shifting from discrete Word Sense Disambiguation (WSD) to evaluating a continuous spectrum of plausibility. SemEval-2026 Task 5 (Gehring et al., 2026) addresses this evolution via the AmbiStory dataset, challenging models to rate homonym plausibility on a scale of 1.0–5.0 within five-sentence narratives. This task exposes a critical cognitive mismatch: while human annotators naturally exhibit a central tendency bias, Large Language Models (LLMs) inherently favor binary extremes and overconfidence. To mitigate this bias, we propose a hybrid ensemble architecture that fuses a LoRA-adapted Llama-3 8B model, acting as a global "Narrative Expert", with a dual-stream DeBERTa-v3-large module, serving as a local "Word Specialist" utilizing semantic similarity and Natural Language Inference (NLI).

This paper provides a comprehensive overview of our methodology and findings, beginning with an exploration of related work in graded semantics and the challenges of tracking narrative consistency across multi-sentence contexts. We detail our dual-pathway architecture, specifically focusing on the regression transformation of Llama-3 and the specialized input engineering required for the DeBERTa-based experts. Following an outline of our experimental setup and training configurations, we present our official competition results, in which our system ranked 22nd out of 79 participating teams. To conclude, we offer a qualitative error analysis examining the "Variance Trap" and the cognitive bias mismatches between artificial networks and human annotations, ultimately suggesting future directions for the development of distribution-matching paradigms in computational narrative understanding.

2 Related work

2.1 Word Sense Disambiguation and Graded Plausibility

Historically, Word Sense Disambiguation (WSD) relied on discrete knowledge bases like WordNet for rigid synset mapping. However, this discrete approach struggles with underspecified, overlapping meanings. Consequently, the field is shifting toward graded semantic plausibility, treating meaning as a continuous spectrum (Gehring and Roth, 2025). While LLMs effectively simulate coarse plausibility judgments (Amouyal et al., 2024), their internal probability distributions often misalign with human judgment variance (Basile et al., 2025). This discrepancy—driven by human "Central Tendency Bias" versus model extremity (Tjuatja et al., 2024; Zheng et al., 2023)—necessitates regression-based fine-tuning to bridge the gap between model confi-

*Equal contribution

[1]Our system is open-source and available for research purposes https://github.com/JCT-2026-SemEval-Task-5/AmbiStory-Hybrid-LLM-NLI-Ensemble

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 826–831

dence and human cognitive scales.

2.2 Narrative Understanding and Contextual Semantics

Evaluating narrative consistency requires models to track entities and logical constraints across multiple sentences. Decoder-only LLMs, such as the Llama-3 family (Grattafiori et al., 2024), have demonstrated profound capabilities in parsing complex, long-context dependencies. In the AmbiStory dataset (Gehring and Roth, 2025), lexical ambiguity typically emerges mid-narrative, with disambiguating clues embedded in surrounding sentences. Generative models leverage their extensive pre-training corpora to assess real-world logical coherence, proving to be effective surrogates for human evaluators when properly aligned via regression scoring (Akhauri et al., 2025).

2.3 Natural Language Inference for Semantic Entailment

Natural Language Inference (NLI) creates a framework for determining logical entailment, contradiction, or neutrality between text pairs. Models such as DeBERTa-v3 (He et al., 2023), which utilize ELECTRA-style Replaced Token Detection (RTD), currently dominate these benchmarks due to their refined contextual embeddings. For plausibility rating, NLI offers a mechanism to reformulate the regression problem: by treating the narrative as a premise and the target definition as a hypothesis, the model evaluates strict logical entailment. This approach provides a rule-bound counterweight to the associative, and occasionally hallucinated, tendencies of generative LLMs (Maynez et al., 2020).

3 System Overview

3.1 Architectural Philosophy

Our architectural philosophy posits that accurate semantic plausibility scoring requires two distinct modes of comprehension: narrative coherence and lexical precision. Relying on a single architecture often biases predictions toward either generalized world knowledge or strict syntactic matching. To mitigate these vulnerabilities, we employ a hybrid ensemble with two pathways: Model A (Narrative Expert), which uses Llama-3-8B to holistically evaluate the story's logical flow, and Model B (Word Specialist), which utilizes DeBERTa-v3-large to verify target word usage against its dictionary definition. The final score is generated through a weighted aggregation, relying solely on provided data without external resources like WordNet. The system architecture is illustrated in Figure 1.

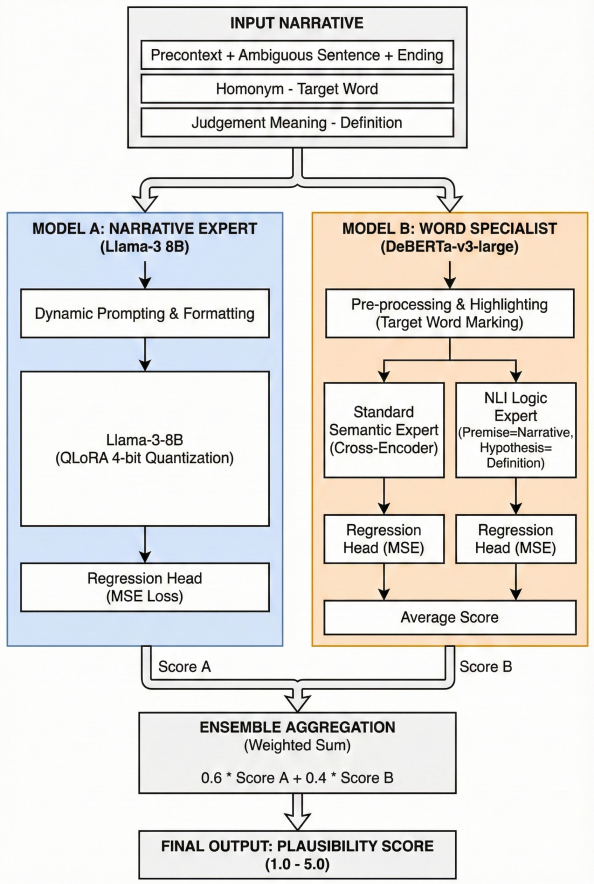

Figure 1: High-level architecture

3.2 Model A: The Narrative Expert

The Llama-3 8B model represents a highly optimized open-weight generative architecture, featuring a 128K-token vocabulary and Grouped-Query Attention (GQA) for efficient processing. To adapt this autoregressive generative model for the continuous scoring constraints of Task 5, we executed an extensive architectural transformation.

3.2.1 Generative to Regression Transformation

We replaced Llama-3's classification head with a dedicated Regression Neuron ($num\_labels = 1$) and utilized Mean Squared Error (MSE) loss. To ensure convergence on the original 1.0–5.0 scale without normalization, we isolated the regression head for full 32-bit precision training and applied output clipping during inference. This shift from discrete probabilities to continuous mathematical

distances allows the model to capture the nuanced magnitude of plausibility rather than binary fits.

3.2.2 Dynamic Prompting and Context-Aware Formatting

To accommodate the structural variability of the AmbiStory dataset, we implemented a context-aware formatting strategy using explicit string templates. When a story includes an ending, the input is constructed as: "*Instruction: Evaluate story consistency. Target: {homonym} ({definition}). Narrative: Context: {precontext} Sentence: {sentence} Ending: {ending}*".

Conversely, when no ending is provided, we shift focus to immediate lexical compatibility using the template: "*Instruction: Evaluate sentence fit. Target: {homonym} ({definition}). Narrative: Context: {precontext} Target Sentence: {sentence}*".

The rationale for this separation is to force the model to explicitly distinguish between early contradictions (within the target sentence) and late contradictions (appearing in the ending), ensuring the system correctly localizes the source of logical inconsistency.

3.2.3 Parameter-Efficient Fine-Tuning (LoRA)

Fine-tuning the entire 8-billion parameter weight matrix is computationally prohibitive and prone to catastrophic forgetting, which would degrade the model's foundational world knowledge. We circumvent this through Quantized Low-Rank Adaptation (QLoRA).

We load the base model in a 4-bit NormalFloat (nf4) format utilizing double quantization via BitsAndBytesConfig, drastically reducing the VRAM footprint while preserving high-fidelity float16 computation data types. Our LoRA configuration injects trainable low-rank matrices into multiple target modules across the attention and feed-forward layers (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj) with a rank (r) of 16, a scaling alpha of 32, and a dropout of 0.1.

A critical architectural safeguard we implement involves the modules_to_save parameter. We explicitly isolate the newly initialized score regression head, ensuring it remains completely unfrozen and is trained in full 32-bit floating-point precision. This guarantees that while the deep linguistic representations are preserved and subtly adapted via specialized adapters, our mathematical regression engine remains unconstrained and is trained entirely from scratch.

3.3 Model B: The Word Specialist

While Llama-3 excels at overarching logic, bidirectional encoders remain superior for token-level semantic disambiguation. We deploy two parallel `deberta-v3-large` instances (435M parameters), leveraging its Gradient-Disentangled Embedding Sharing to provide refined contextual representations. To stabilize gradients and improve convergence, we normalize human labels (1.0–5.0) to a [0,1] scale during training, rescaling to the original range for final inference.

To guide the self-attention mechanism, we apply a targeted input engineering step: the homonym within the ambiguous sentence is wrapped in focus markers (e.g., " homonym "). This programmatic injection serves as a hard attention anchor, explicitly signaling which token requires disambiguation against the provided definition.

3.3.1 Standard Semantic Expert

The first instance operates as a cross-encoder. We supply the dictionary definition as the primary sequence and the full narrative as the secondary sequence. Passing both simultaneously through deep attention layers enables the model to generate rich cross-attentional representations, measuring the direct semantic overlap between the narrative environment and the definition.

3.3.2 NLI Logic Expert

The second instance reformulates the task as Natural Language Inference (NLI). Initialized with a DeBERTa variant fine-tuned on multi-source inference datasets, this expert treats the full story as the *Premise*. We construct the *Hypothesis* using a rigid template: "The word 'homonym' implies: judged_meaning."

This reframing forces the network into strict logical reasoning. Rather than detecting mere topical relevance, which often triggers false positives due to deceptive vocabulary, the model calculates the probability that the definition is a logical, entailed consequence of the described events. Both experts utilize a linear regression head squashed by a Sigmoid activation function and optimized for MSE.

3.4 Ensemble Aggregation Strategy

Our final architecture employs a weighted ensemble to merge the divergent capabilities of Models A and B. Following hyperparameter tuning on

the validation set, we determined an optimal ratio of 60% Model A and 40% Model B (calculated as the mean of the two DeBERTa experts). This weighting reflects our empirical observation that high-level narrative coherence influences human plausibility ratings more significantly than granular lexical matching in the AmbiStory dataset.

4 Experimental Setup

4.1 Dataset Utilization

We trained and evaluated our system utilizing the official AmbiStory dataset provided for SemEval-2026 Task 5. The data consists of English short narratives, typically containing a precontext, an ambiguous target sentence, and a specific ending (if present). The dataset's structural design ensures that varying these endings across different instances manipulates the contextual plausibility of distinct word senses. For our training and evaluation, we utilized the provided ground truth labels, which represent the arithmetic average of five human annotations per instance, scaled continuously from 1.0 to 5.0.

4.2 Hyperparameter Configuration

4.2.1 Model A (Llama-3 8B)

We utilized the `paged_adamw_8bit` optimizer and gradient checkpointing to manage VRAM. Inputs were tokenized to a maximum length of 512, using the EOS token for padding. As a causal decoder, we employed Last Token Pooling, applying the regression head directly to the final non-padded token's hidden state to capture the full sequence context. We set the learning rate to 1×10^{-4} and trained over 4 epochs with an effective batch size of 16 (per-device 2, accumulation 8). Evaluation and checkpointing occurred every 50 steps, selecting the best model based on validation Spearman correlation.

4.2.2 Model B (DeBERTa-v3-large Hybrid)

We trained the two DeBERTa experts independently over 5 epochs with an effective batch size of 16. Inputs were padded and truncated to a maximum length of 512. For regression scoring, we utilized `[CLS]` token pooling, feeding the initial token's final hidden state into the linear head to capture the bidirectional representation. We applied the AdamW optimizer with a conservative learning rate of 1×10^{-5} and 0.01 weight decay to prevent catastrophic unlearning of the pre-trained cross-encoder weights.

4.3 Evaluation Metrics

We benchmarked our system performance strictly against the official SemEval metrics for Task 5:

- **Spearman Rank Correlation (ρ):** A non-parametric measure evaluating the monotonic relationship between the model's predicted scores and the human ground truth. We utilize this metric to assess whether our system correctly ranks the relative plausibility of stories, focusing on rank alignment rather than the absolute numerical values predicted.
- **Accuracy within Standard Deviation (Soft Accuracy):** Recognizing the inherent noise and variance in subjective human annotation, this metric forgives minor numerical deviations. We classify a prediction as "correct" if the absolute error between the prediction and the human average is less than or equal to the Standard Deviation of the human scores (utilizing a minimum allowed threshold buffer of 1.0).
- **Combined Average:** The definitive ranking metric utilized for the competition leaderboard. We calculate this as the arithmetic mean of the Spearman Correlation and the Soft Accuracy.

5 Results

5.1 Official Competition Performance

Notably, our system demonstrates strong competitive performance relative to the reported human upper bound of 0.834 for Spearman correlation and 0.892 for Accuracy within SD. While a gap of 12 points in correlation remains, our model significantly narrows the gap to 7 points in soft accuracy, demonstrating its efficacy in mirroring human perception of word sense plausibility. Table 1 summarizes our performance across official metrics.

5.2 Quantitative Analysis and Model Comparison

As shown in Table 2, the ablation study highlights the complementary strengths of our modules. Within Model B, the Semantic Expert achieves higher Accuracy (0.7500), while the NLI Expert yields a stronger Spearman correlation (0.6636).

Metric	Test Set Performance
Accuracy within SD	0.8204
Spearman Rank Correlation (ρ)	0.7118
Combined Average	**0.7661**

Table 1: Official evaluation results of our proposed system on the AmbiStory test set.

Their integration into a dual-stream sub-ensemble boosts accuracy to 0.7942, confirming that they capture distinct and synergistic semantic signals.

To find the optimal balance, we tested different weights for Model A (ranging from 0.0 to 1.0). Our results show a clear bell-curve trend: performance improves as the models are combined, reaching a peak exactly at the 60/40 ratio ($a = 0.6$). At this point, the ensemble achieves its maximum accuracy (0.8180) and Spearman correlation (0.7220). Moving away from this balance by relying too much on either the narrative logic of Model A or the lexical precision of Model B, leads to a steady drop in scores. This peak confirms that a 40% corrective signal from the Word Specialist is essential to reach the system's full potential.

Configuration	Spearman (ρ)	Accuracy within SD
Model A		
Narrative Expert	**0.6900**	**0.7960**
Model B - Semantic Expert	0.5873	0.7500
Model B - NLI Logic Expert	0.6636	0.704
Model B		
Word Specialist		
Ensemble	**0.6600**	**0.7942**
Full Hybrid Ensemble (A + B)	**0.7220**	**0.8180**

Table 2: Comparative performance of individual modules versus the weighted ensemble on the dev split.

6 Error Analysis

A qualitative and quantitative review of our predictive failures on the official test set reveals a distinct "Human-AI Gap". When evaluating the absolute error against the Soft Accuracy threshold ($\max(SD, 1.0)$), we identified 167 out of 930 instances (17.9%) where our ensemble failed to align with human consensus. We categorized these errors into three primary distribution buckets:

Error Category	Count	Percentage
Under-prediction	80	47.9%
Over-prediction	75	44.9%
The Variance Trap	12	7.2%
Total Failures	**167**	**100%**

Table 3: Distribution of system errors categorized by primary linguistic and statistical phenomena on the test set.

6.1 Under-prediction: Rigid Logic & Implicit Ellipsis (47.9%)

Accounting for nearly half of our system's failures (80 instances), the ensemble frequently underpredicted scores compared to human annotators. These errors typically occur in scenarios requiring "imaginative leeway" or pragmatic flexibility. For example, in a narrative involving the consumption of wine inside a movie "trailer," our system predicted a rigid 1.1 (Logical Mismatch). In contrast, human annotators assigned a 2.5 average, extending a degree of pragmatic allowance to the scene that our model's strict logic lacks.

6.2 Over-prediction: Deceptive Lures & WSD Blindness (44.9%)

In 75 instances, the ensemble fell victim to "WSD Blindness" caused by deceptive lexical lures. In these cases, Model B was often misled by strong semantic overlaps between local vocabulary and the dictionary definition. For instance, in a story containing "river" (context) and "deposit" (target), the model over-indexed on the immediate semantic signal. It failed to integrate the global narrative constraint that a muddy river inside a financial building is physically impossible, leading to an erroneously high plausibility score.

6.3 The Variance Trap: Central Tendency Mismatch (7.2%)

In 12 instances, errors were driven by the "Variance Trap", where highly polarized human annotations average to a neutral score (2.5–3.5). The model is mathematically penalized for making definitive logical predictions against this flattened artifact of human disagreement.

7 Conclusion

Our participation as Team JCT 2026 in SemEval-2026 Task 5 underscores the necessity of hetero-

geneous architectures for nuanced semantic evaluation. By adapting a generative LLM into a regression engine via LoRA and grounding it with an NLI-focused cross-encoder, our system successfully combines holistic narrative comprehension with precise lexical disambiguation. Achieving 82.04% Soft Accuracy and a 0.71 Spearman Correlation, our hybrid ensemble demonstrates high efficacy in mirroring human perception of word sense plausibility.

However, error analysis reveals persistent challenges, primarily the "Variance Trap" and the cognitive mismatch between human central tendency and LLM extremism. Advancing continuous WSD plausibility requires a fundamental shift in optimization mathematics. Future work must transition from absolute distance metrics (e.g., MSE) to distribution-matching paradigms (such as Jensen-Shannon divergence). Specifically, during the training phase, models should be optimized to predict the full probability distribution of human disagreement ratings, rather than regressing toward a compromised central mean. Additionally, future iterations of this architecture could benefit from implementing a dynamic weight allocation mechanism—adaptively adjusting the influence of the Narrative and Word experts based on input features—and exploring the integration of external knowledge bases, such as WordNet, to further enrich the disambiguation capabilities of the semantic experts.

References

Yash Akhauri, Bryan Lewandowski, Cheng-Hsi Lin, Adrian N. Reyes, Grant C. Forbes, Arissa Wongpanich, Bangding Yang, Mohamed S. Abdelfattah, Sagi Perel, and Xingyou Song. 2025. Performance prediction for large systems via text-to-text regression. *Preprint*, arXiv:2506.21718.

Samuel Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2024. Large language models for psycholinguistic plausibility pretesting. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 166–181, St. Julian's, Malta. 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.

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.

Janosch Gehring and Michael Roth. 2025. AmbiStory: A challenging dataset of lexically ambiguous short stories. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 152–171, Suzhou, China. 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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.