# Guys_LLM at SemEval-2026 Task 5: NLI-Informed Regression for Graded Word-Sense Plausibility in Narrative Contexts

**Niccoló Antonelli-Dziri*** **Sixtine Marcotte*** **Emanuele Rosapepe***
**Gabriele Santona*** **Omar Wafaay*** **Lorenzo Vaiani†**
**Riccardo Coppola†** **Flavio Giobergia†**
Politecnico di Torino
* {s350296,s350492,s346705,s343402,s355114}@studenti.polito.it
† {firstname.lastname}@polito.it

## Abstract

While large language models (LLMs) excel at semantic reasoning, their discrete token-based outputs introduce limitations for fine-grained regression tasks requiring continuous scoring. We address graded word-sense plausibility estimation by reformulating it as a Natural Language Inference (NLI) regression problem, adapting DeBERTa-v3-large with NLI pretraining and a regression head to predict continuous plausibility scores from story-sense pairs. We compare this model against BERT, vanilla DeBERTa, SmolLM variants and state-of-the art LLMs under various prompting strategies, and show that the NLI-finetuned model achieves superior rank correlation and alignment with human judgments. While several baselines collapse toward mean predictions and LLMs show unstable prompting sensitivity, our findings establish NLI-informed pretraining as highly effective for narrative plausibility regression, highlighting fundamental LLM limitations for word sense disambiguation.

**Code:** github.com/NiccoloAntonelliDziri/LLM-SemEval-T5

## 1 Introduction

Word Sense Disambiguation has long been a fundamental challenge in natural language processing, requiring systems to determine which meaning of a polysemous word is intended in a given context. Traditional approaches typically assume that only one sense is correct for any given usage, framing the task as a classification problem (e.g. (Navigli, 2009)). However, this binary view does not fully capture how humans perceive ambiguity in language, where multiple interpretations may be simultaneously plausible to varying degrees.

SemEval 2026 Task 5 (Gehring et al., 2026) addresses this challenge by focusing on rating the plausibility of different word senses within ambiguous sentences through narrative understanding. Rather than simply selecting the correct interpretation, this task requires systems to evaluate the plausibility of each potential sense in the context of short narrative stories. This approach acknowledges that ambiguity, underspecification, and individual linguistic experience all influence how humans interpret meaning, with narrative coherence playing a crucial role in these judgments.

In this paper, we compare specialized models, such as DeBERTa-v3 (He et al., 2021a), which are trained for specific tasks, with larger, state-of-the-art Large Language Models (LLMs) like Llama 3.1 and Deepseek-r1. Our focus is on two key aspects: (i) how similarly to humans do these models rank information (Spearman correlation) and (ii) how consistent their results are with human expectations (accuracy within the standard deviation).

We demonstrate that, while LLMs possess vast linguistic knowledge, they often fail to capture human intuition, a task for which specialized regression-tuned models remain superior.

## 2 Related Work

Word Sense Disambiguation is traditionally framed as a classification problem, selecting a single correct sense for a target word in context (Navigli, 2009). However, this formulation does not capture the graded nature of human interpretation. Recent work, including SemEval-2026 Task 5, instead models plausibility as a continuous variable, requiring systems to score candidate meanings rather than make discrete decisions.

Natural Language Inference (NLI) provides a



*Proceedings of the The 20th International Workshop on Semantic Evaluation (2026)*, pages 2487–2494

general framework for modeling semantic relationships between text pairs (Bowman et al., 2015; Williams et al., 2018). Models pretrained on NLI data have been shown to capture transferable reasoning abilities (Laurer et al., 2024a), and have been successfully applied to a variety of downstream tasks through premise–hypothesis formulations. Our approach follows this paradigm, extending it to regression for fine-grained plausibility estimation. This approach of leveraging the signal produced by Language Models trained on classification tasks (e.g., next token prediction, or NLI) to solve regression tasks has already been adopted in literature, e.g. for semantic similarity scoring (Reimers and Gurevych, 2019), or for annotator agreement estimation (Giobergia, 2026).

Transformer encoders such as BERT (Devlin et al., 2018) and DeBERTa (He et al., 2021b) remain strong baselines for structured prediction tasks when fine-tuned. In contrast, larger models are typically used via prompting, but can exhibit instability and limited calibration in numerical prediction settings. Similar observations have been reported in prior work on LLM-based classification (Giobergia et al., 2024). Our work compares these paradigms, showing that NLI-pretrained encoders are more reliable for continuous plausibility scoring.

## 3 Task description

The task is built upon the AmbiStory dataset (Gehring and Roth, 2025), an English language dataset, designed to probe how narrative progression resolves or maintains lexical ambiguity. Each instance consists of a five-sentence narrative featuring a target homonym.

Each record is structured into three parts:

- **Precontext**: three sentences that establish the setting and relevant events.
- **Ambiguous sentence**: the fourth sentence contains a homonym with two widely different candidate senses.
- **Ending**: an optional fifth sentence, where one of two possible endings (or no ending) may be provided and often biases the interpretation toward one sense.

At least five ratings were collected for each story.

Given the story context and a candidate sense definition, the system is expected to output a plausibility score for that sense, ranging from 1 to 5. Evaluation uses two complementary metrics:

- **Spearman correlation**: correlation between the model predictions and the mean human plausibility scores;
- **Accuracy within standard deviation**: proportion of predictions that fall within one standard deviation of the mean human rating, reflecting agreement under varying annotator consensus.

## 4 Methodology

### 4.1 The Gold Standard

The core idea of our approach was to adapt the task into a Natural Language Inference (NLI) problem, using a hypothesis-premise pair:

- **Premise**: The full story context, composed of precontext, target sentence, and ending (if available).
- **Hypothesis**: A structured sense definition: *The definition of '{homonym}' is: '{judged_meaning}' as in the following sentence: '{example_sentence}'.*

We use a pre-trained DeBERTa-v3-large model (He et al., 2021a), specifically the variant fine-tuned on multiple NLI datasets (Laurer et al., 2024b). We adapted this model for our regression task by replacing the original three-class classification head with a single-output regression head. The model is trained to minimize a Mean Squared Error (MSE) loss.

This formulation allows the model to leverage its pre-trained understanding of semantic entailment to assess the plausibility of each word sense within the given narrative context.

### 4.2 Alternative Baseline Architectures

To quantify the gain provided by DeBERTa-v3's architecture, we benchmarked several alternative families against our primary NLI-finetuned DeBERTa model:

**BERT-base (Devlin et al., 2018).** We utilized this state-of-the-art model as a baseline for the encoder family. This comparison measures the impact of parameter scale (110M for BERT base, 435M for DeBERTa v3 large) and the performance ceiling of early transformer architectures on narrative-dense tasks.

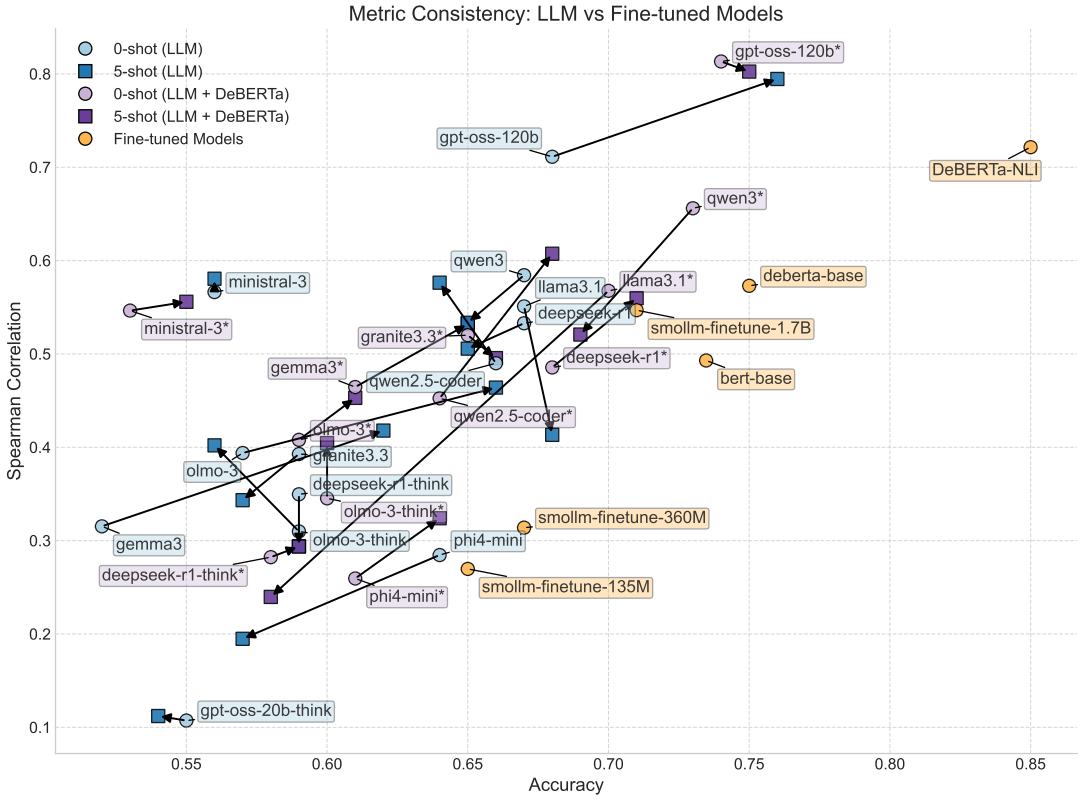



Figure 1: Metric consistency: accuracy and Spearman on a randomly selected 100 elements. The '*' next to model names indicates that they were run with DeBERTa injection (purple dots and squares).

**Vanilla DeBERTa-v3-large.** We fine-tuned a "vanilla" (i.e., non-NLI-tuned) version of the DeBERTa-v3-large to specifically quantify the gain provided by the NLI-specific pre-finetuning. This version uses the same attention mechanism as our primary model but lacks the internal representations of semantic entailment provided by the NLI datasets, enabling us to evaluate the gains from transfer learning.

**SmolLM (Allal et al., 2024).** Shifting from encoder-based masked models to decoder-only (causal) architectures, we fine-tuned the three variants of the SmolLM series. Specifically:

- **135M**: A minimal-parameter baseline for efficient edge deployment.
- **360M**: A mid-range model comparable in size to DeBERTa-v3-large but with a causal objective.
- **1.7B**: A model that explores the benefits of significantly larger parameter counts in a "small" language model format.

For fair comparison, all the above models were trained using the same training setup (A.1).

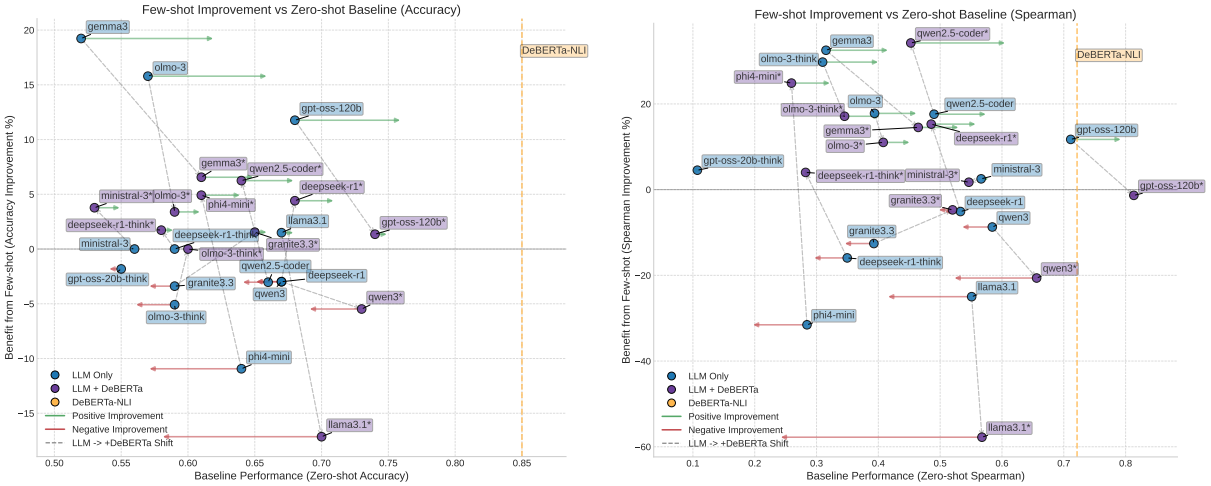
### 4.3 Large Language Model Benchmarking

Finally, we evaluated a broad range of state-of-the-art LLMs to assess their zero-shot and few-shot performance on the task, considering both instruction-tuned and reasoning ("thinking") variants.

We used four prompt configurations: zero-shot, few-shot, with and without DeBERTa-injection. Few-shot prompts were built from training data by selecting one representative example per score (1–5). In the hybrid setting, we appended an explicit "Additional Context" block with the DeBERTa prediction and its evaluation statistics, while keeping the core annotation task unchanged.

For reasoning models with significantly higher inference costs, we evaluated on a fixed set of 100 randomly selected instances to make the results comparable. This keeps the results comparable and avoids bias from the dataset ordering, where examples are grouped by homonym in consecutive blocks of six closely related items (same target sentence with two candidate senses and different ending variants).

To ensure comparable results across architectures, all LLMs were set to a temperature of 0 for deterministic (greedy) sampling.




(a) Accuracy within Std. Dev.

(b) Spearman correlation

Figure 2: Few-shot learning potential on a randomly selected subset of 100 examples. We report accuracy within standard deviation (left) and Spearman correlation (right) across the four prompting configurations. The arrows (red and green) point towards the values of the five-shot prompting. The grey dotted lines link the models with and without DeBERTa.

## 5 Experiments

To align with the first phase of the SemEval task and prevent data leakage, we split the training data into training and validation sets (90/10) and reserved the development data exclusively for final evaluation.

Model training and LLM evaluation were performed on an NVIDIA GeForce RTX 4060 laptop GPU, except for gpt-oss-120b, whose inference was run via Groq (free plan).

## 6 The Metric Consistency Landscape

The core of our evaluation lies in the relationship between Spearman correlation and accuracy within standard deviation. Figure 1 visualizes this landscape, mapping each model's ability to rank plausibility (y-axis) against its ability to remain within the boundaries of human consensus (x-axis).

### 6.1 Finetuned models

A first observation from Figure 1 is the grouping of fine-tuned models (BERT family and SmolLM series indicated as orange dots) along a line in the high-accuracy/low-correlation region. When these models are fine-tuned on the AmbiStory regression task with a Mean Squared Error objective, they can reduce their loss by gravitating toward a conservative, near-constant prediction near the global mean (around 3). This can yield relatively high accuracy within the standard deviation (many predictions fall within the typical rating range) while still failing to capture the monotonic ranking of senses, leading to a low Spearman correlation.

For the BERT family, gains appear to be driven by architecture choices (DeBERTa vs. BERT) and, crucially, by the intermediate fine-tuning regime (NLI-supervised weights vs. vanilla weights). DeBERTa-NLI (top-right) suggests that encoder-based models can achieve both high accuracy and strong rank correlation when they benefit from additional training on an entailment-style objective, which is closely related to our premise–hypothesis formulation. This gain is not architectural. DeBERTa and DeBERTa-NLI were trained with the same pipeline and hyperparameter regime, so the key difference is the NLI pretraining. The NLI weights already encodes cross-sentence entailment reasoning. That gives DeBERTa-NLI a head start on fine-grained ranking, instead of collapsing to a mean prediction.

The finetuned SmolLM variants reveal a consistent scaling trend within decoder-only architectures: the 1.7B variant substantially outperforms both 360M and 135M baselines. However, even the largest SmolLM variant falls short of DeBERTa-NLI, suggesting that causal language modeling, while capable of learning regression, has fundamental difficulties with fine-grained plausibility estimation. This architectural limitation fore-

shadows the broader struggle we observe across instruction-tuned LLMs: if even optimized decoder models trained explicitly on the task cannot match entailment-informed encoders, then general-purpose language models are unlikely to excel at this regression problem.

Our DeBERTa-NLI model achieved a Spearman correlation of 0.723 and an accuracy within standard deviation of 82.5%, ranking 32nd on the leaderboard with the new evaluation dataset.

### 6.2 Not all LLMs are equal

To investigate the LLMs, we benchmarked them in four distinct configurations to observe their movement across the landscape: zero-Shot, five-Shot, zero-shot + DeBERTa injection, and five-Shot + DeBERTa injection, to bridge the gap between the specialized precision of DeBERTa and the broad reasoning of LLMs.

Contrary to the expectation of a universal "horizontal shift" toward higher accuracy upon providing DeBERTa context, we observed a stratified response in Figure 2. Some models, such as Gemma3 and Granite3.3, showed significant alignment gains when injected with DeBERTa scores, effectively adopting the encoder's scale as a prior. However, other models, such as Llama3.1 or Olmo3, exhibited signal interference: injecting the external score occasionally caused the model to vacillate between its internal semantic logic and the teacher signal, leading to decreases in both accuracy and correlation. This suggests that, for some models, the DeBERTa prediction serves as adversarial noise rather than helpful support.

We observed a similar effect with few-shot prompting in some cases (Phi4-mini or Qwen3), where adding examples degraded performance instead of improving it.

While few-shot prompting is traditionally expected to improve task alignment, our experiments revealed a performance degradation in several LLMs when transitioning from zero-shot to five-shot configurations. We hypothesize that this degradation is rooted in an attenuation of lexical focus. In a zero-shot setting, the model's attention is strictly concentrated on the single target homonym and its relationship to the provided candidate sense within a single story. However, a five-shot prompt for AmbiStory contains approximately 25-30 sentences of extraneous narrative context.

Overall, DeBERTa score injection is most reliable in the zero-shot setting or when paired with explicit reasoning prompts. In contrast, for instruction-tuned LLMs, combining few-shot exemplars with a numeric teacher prior can introduce conflicting signals, often reducing both accuracy and rank correlation.

### 6.3 Gpt-oss-120b performance

Notably, the large open-source model 'openai-gpt-oss-120b' shows a pattern where its ability to rank candidate senses can be competitive with DeBERTa-NLI when given additional context, while its absolute accuracy within the human standard deviation range is slightly lower. For instance, in our 100-instance evaluation 'openai-gpt-oss-120b' (zero-shot) reports Spearman around 0.685 and accuracy around 0.707, whereas DeBERTa-NLI yields Spearman around 0.693 and Accuracy around 0.823. When the DeBERTa prediction is injected into the LLM prompt, Spearman increases to around 0.737 while accuracy remains comparatively unchanged, hinting that large models can improve ranking but not necessarily calibration.

These results suggest that very large LLMs can come close to ranking under carefully tuned prompting, but that their behaviour remains sensitive to prompting choices and auxiliary signals; in contrast, the small encoder-only DeBERTa-NLI remains the strongest and most stable model overall. Crucially, it achieves this with orders-of-magnitude fewer parameters, making it substantially more efficient in terms of performance per model size.

### 6.4 Impact of the ending sentence

To evaluate the effect of the ending sentence, we computed metrics separately for examples that include an ending and those that do not. The DeBERTa-NLI model shows nearly identical accuracy (82.65% without vs 82.14% with), but a clear improvement in correlation (Spearman 0.651 vs 0.707), indicating that the ending helps the model evaluate candidate choices more like human judgments, even when it rarely changes the single best prediction. This pattern suggests that the ending provides additional disambiguating cues that help the model better distinguish among the candidate senses.

## 7 Discussion and Comments

While the DeBERTa regression head provides a continuous score, LLMs are fundamentally optimized for token selection and discrete labeling.

When prompted to use the output of DeBERTa-NLI, the LLMs act as discretizers. This process destroys the regression's granularity.

We also report a random baseline that samples plausibility scores according to the empirical label distribution observed in the training set (i.e., not from a uniform distribution). Under this distribution-matched guessing strategy, performance behaves as expected, with a near-zero Spearman correlation and an accuracy of around 46%. We additionally consider a degenerate baseline that always predicts the central score 3; in this case, Spearman correlation is undefined (NaN) due to the lack of variance in the predictions, while accuracy rises to 53%. Taken together, these baselines confirm that the task is non-trivial (even a prior-matched or mean-only strategy remains weak) and that every model learns or understands meaningful patterns beyond label-frequency effects.

## 8 Conclusion

We reformulate graded sense-plausibility estimation as an NLI-style regression task and show that an encoder-only DeBERTa-v3 model with NLI pre-training achieves the best overall performance, outperforming both encoder baselines (BERT, vanilla DeBERTa) and decoder-only SmolLM variants. Across a broad LLM benchmark, results are comparatively unstable and sensitive to prompting and score-injection strategies, indicating that general-purpose LLMs remain less reliable for fine-grained continuous scoring. Overall, the strong performance of DeBERTa-NLI at a much smaller parameter budget highlights the effectiveness and efficiency of entailment-informed encoders for narrative plausibility regression. More broadly, the combination of narrative reasoning and continuous plausibility scoring makes this task a promising candidate for future LLM benchmarking.

## References


Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Leandro von Werra, and Thomas Wolf. 2024. Smollm - blazingly fast and remarkably powerful.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Flavio Giobergia. 2026. MINDS at GSI: Detect: From Logits to Degrees of Agreement in Gender Stereotype Detection with LLMs. In *Proceedings of the 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2026)*, volume 4195 of *CEUR Workshop Proceedings*, Bari, Italy. CEUR-WS.org.

Flavio Giobergia, Alkis Koudounas, and Elena Baralis. 2024. Large language models-aided literature reviews: A study on few-shot relevance classification. In *2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–5. IEEE.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024a. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024b. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

# A Appendix

This appendix provides additional details and qualitative examples omitted from the main paper for space.

## A.1 Training and hyperparameters

The following training arguments were use with Pytorch for all of the models that were fine-tuned. The "combined_score" metric is the average between Spearman correlation and accuracy within standard deviation. The decoder only models (Smollm familly) have a different number of num_train_epochs=6, a different per_device_train_batch_size=8 and gradient_accumulation_steps=1.

```
training_args = TrainingArguments(
    output_dir=f"{OUTPUT_DIR}/model/",
    num_train_epochs=4,
    per_device_train_batch_size=4,
    per_device_eval_batch_size=64,
    gradient_accumulation_steps=2,
    logging_dir=f"{OUTPUT_DIR}/logs/",
    logging_steps=10,
    eval_strategy="epoch",
    save_strategy="epoch",
    load_best_model_at_end=True,
    metric_for_best_model="combined_score",
    learning_rate=2e-5,
    save_total_limit=2,
    remove_unused_columns=True,
    warmup_ratio=0.06,
    weight_decay=0.01,per_device_train_batch_size
    )
```

## A.2 Zero-Shot Prompt

System prompt:

```
You are an expert NLU annotator. Your job is to
↪  rate how plausible a candidate meaning
↪  (sense)
is for the \textit{HOMONYM} used in the target
↪  sentence within the short story.

Return ONLY a single JSON object with one key:
↪  "score" and an integer value 1, 2, 3, 4 or 5.
Integer mapping:
  1 = Definitely not
  2 = Probably not
  3 = Ambiguous / Unsure
  4 = Probably yes
  5 = Definitely yes
```

User prompt:

```
[STORY]
{full_story_text}

[HOMONYM]
{homonym}

[CANDIDATE SENSE]
{sense_text}

[TASK]
Based on the STORY above, decide how plausible it
↪  is that the HOMONYM is used with the
CANDIDATE SENSE in the target sentence.

USER PROMPT:
[STORY]
{full_story_text}

[HOMONYM]
{homonym}

[CANDIDATE SENSE]
{sense_text}

[TASK]
Based on the STORY above, decide how plausible it
↪  is that the HOMONYM is used with the
CANDIDATE SENSE in the target sentence.

Return ONLY a single JSON object with one key
↪  "score" and an integer value (1-5)
as described by the system message. Example
↪  output: {{"score": 3}}
```

## A.3 Few-Shot Prompt

System prompt:

```
You are an expert NLU annotator. Your job is to
↪  rate how plausible a candidate meaning
↪  (sense)
is for the HOMONYM used in the target sentence
↪  within the short story.

Return ONLY a single JSON object with one key:
↪  "score" and an integer value 1, 2, 3, 4 or 5.
Integer mapping:
  1 = Definitely not
  2 = Probably not
  3 = Ambiguous / Unsure
  4 = Probably yes
  5 = Definitely yes

The response must be a JSON object and nothing
↪  else, for example: {"score": 4}

[EXAMPLES]
{few_shot_examples}
```

User prompt:

```
Now, label this new instance:

[STORY]
{full_story_text}

[HOMONYM]
{homonym}

[CANDIDATE SENSE]
{sense_text}
```

```
[TASK]
Based on the STORY above, decide how plausible it
↪  is that the HOMONYM is used with the
CANDIDATE SENSE in the target sentence.

Return ONLY a single JSON object with one key
↪  "score" and an integer value (1-5)
as described by the system message. Example
↪  output: {"score": 3}
```

### A.4 Prompts with DeBERTa Injection

In the DeBERTa-injection setting, the system prompts remain unchanged. We modify only the user prompt by adding the following text:

```
[ADDITIONAL CONTEXT]
A DeBERTa model (Accuracy:
↪  {deberta_accuracy:.2f}, Spearman Correlation:
↪  {deberta_spearman:.2f}) predicted a score of
↪  {deberta_prediction:.2f} for this example.
You can use this information to guide your
↪  decision, but rely on your own judgment if
↪  the context strongly suggests otherwise.
```