CuriosAI at SemEval-2026 Task 10: Hybrid approaches to conspiracy span extraction and conspiracy detection

Hiroki Takushima, Daichi Yamaga, Fumika Beppu
Aiswariya Manoj Kumar, Yuki Shibata, Takayuki Hori

SoftBank Corp.
{hiroki.takushima,daichi.yamaga01,fumika.beppu,
aiswariya.manojkumar,yuki.shibata04,takayuki.hori}@g.softbank.co.jp

Abstract

We present CuriosAI's system for SemEval-2026 Task 10, addressing Conspiracy Marker Extraction and Conspiracy Detection. For marker extraction, we employ multi-label token classification with a bidirectional transformer (DeBERTa-v3-large) to predict overlapping spans. Alternative feature-based and LLM-based approaches do not surpass the encoder baseline. For Conspiracy Detection, we compare heterogeneous models, including transformer fine-tuning, lexical classifiers, embedding-based models, and LLM-based refinement. Development-optimal models do not always generalize best; logit-level ensembling achieves the strongest test performance ($F_1 = 0.7620$). These results highlight the importance of bidirectional token modeling for span extraction and calibration-aware ensembling for robust detection.

1 Introduction

Conspiracy discourse operates at both span and post levels. SemEval-2026 Task 10 (Samory et al., 2026) reflects this distinction through Conspiracy Marker Extraction and Conspiracy Detection.

For marker extraction, we formulate the task as multi-label token classification using a bidirectional transformer encoder. For detection, we evaluate heterogeneous models, including transformer fine-tuning, lexical classifiers, embedding-based models, and LLM-based refinement. We observe that development-optimal models do not necessarily yield the strongest test performance. Logit-level ensembling improves robustness under distribution shift.

Our contributions are: (1) an empirical comparison of span extraction strategies, (2) a heterogeneous modeling framework for detection, and (3) evidence that calibration-aware ensembling improves robustness in shared task settings.

2 Related Work

2.1 Conspiracy and Misinformation Detection

Conspiracy detection is closely related to misinformation and stance classification tasks. Pretrained transformer models such as BERT (Devlin et al., 2019) have become standard baselines for text classification. Beyond contextual encoders, lexical and feature-based approaches have been explored for interpretable and lightweight settings.

Our work differs in that we systematically compare heterogeneous modeling families within a shared task framework, including encoder-based, lexical, embedding-based, and LLM-based strategies.

2.2 Span-Level Detection

Span-level detection has been studied in tasks such as propaganda technique identification (Da San Martino et al., 2020) and toxic span detection (Pavlopoulos et al., 2021). These tasks typically employ sequence labeling or token classification architectures. Multi-label token classification is particularly suitable when overlapping spans are permitted.

In addition to encoder-based token modeling, we evaluate lexical feature augmentation using Empath (Fast et al., 2016) and parameter-efficient LLM fine-tuning via LoRA (Hu et al., 2022), analyzing their effectiveness under overlapping-label constraints.

2.3 Ensembling and Calibration

Model ensembling is widely used to improve robustness. Deep ensembles enhance uncertainty estimation and generalization (Lakshminarayanan et al., 2017), while modern neural networks are often poorly calibrated (Guo et al., 2017). We extend this line of work by applying logit-level ensembling to heterogeneous classifiers for Conspiracy Detection.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 497–502

3 Conspiracy Marker Extraction

3.1 Task Formulation

In the Conspiracy Marker Extraction subtask, the objective is to predict labeled text spans belonging to five marker types: Actor, Action, Effect, Victim, Evidence. Spans may overlap, and multiple labels may apply to the same token. Performance is evaluated using token-level Intersection-over-Union (IoU) with a threshold of 0.5.

Span-level detection has been studied in related domains such as propaganda technique identification (Da San Martino et al., 2020) and toxic span detection (Pavlopoulos et al., 2021). Following these paradigms, we formulate the task as multi-label token classification.

Given an input sequence of length n, a transformer encoder produces contextualized representations $H \in \mathbb{R}^{n \times d}$. Each token representation is projected to $K = 5$ independent logits:

$$Z = HW + b,$$

followed by a sigmoid activation. Training minimizes binary cross-entropy across token–label pairs, allowing multiple marker types per token.

3.2 Multi-Label Transformer Baseline

Our primary model is based on DeBERTa-v3-large (He et al., 2021), a bidirectional transformer architecture built upon the self-attention mechanism introduced by Vaswani et al. (2017). All encoder parameters are fine-tuned jointly.

At inference time, token-level probabilities are thresholded, and contiguous positive tokens are merged into span predictions. Character offsets are reconstructed from token boundaries. This bidirectional formulation naturally supports overlapping markers and aligns with the IoU-based evaluation protocol.

To assess the impact of encoder architecture, we additionally evaluate RoBERTa-base (Liu et al., 2019) and ELECTRA-base (Clark et al., 2020) under the same multi-label token classification setup. Both models consistently underperform DeBERTa, highlighting the benefit of disentangled attention for fine-grained span detection.

3.3 Feature-Augmented Modeling

We augment token representations with lexical category features derived from Empath (Fast et al., 2016). Category scores are selected based on correlation with the binary conspiracy label and concatenated with token embeddings before classification. However, this approach significantly underperforms the encoder-only baseline, suggesting that global lexical signals introduce noise at the token level.

3.4 Retrieval and Dictionary-Based Methods

We explore two non-parametric alternatives.

Retrieval-Based Span Transfer. Training posts are embedded into a semantic space, and nearest neighbors are retrieved for each input. Annotated spans from retrieved posts are transferred via string matching. While effective for recurring surface patterns, this approach is sensitive to paraphrasing and lacks contextual generalization.

Dictionary Matching. Frequent span expressions are compiled using TF-IDF filtering. Matches are directly applied to the input text. This method suffers from high false positives due to the absence of contextual disambiguation.

3.5 LLM-Based Marker Extraction

We additionally investigate parameter-efficient fine-tuning of a decoder-only large language model, Qwen3-14B (Yang et al., 2025), using Low-Rank Adaptation (LoRA) (Hu et al., 2022). The model jointly predicts sequence-level labels and token-level markers.

Despite reasonable sequence-level performance, token-level span extraction remains substantially weaker than encoder-based models. We attribute this gap to the causal attention constraint in decoder-only architectures, which restricts access to right-side context critical for precise span boundary detection.

4 Conspiracy Detection

4.1 Task Formulation

In the Conspiracy Detection subtask, each Reddit post x is classified as conspiratorial ($y = 1$) or non-conspiratorial ($y = 0$). Models output a probability $p(y = 1 \mid x)$, and predictions are obtained via thresholding:

$$\hat{y} = \mathbf{1}[p \geq t],$$

where t is selected based on development data. Training is performed using binary cross-entropy loss.

4.2 Transformer Fine-Tuning

We fine-tune DeBERTa-v3-large (He et al., 2021) with a linear classification head. As a bidirectional transformer architecture built upon self-attention (Vaswani et al., 2017), it captures contextual semantics effectively and provides a strong baseline.

However, development-optimal checkpoints do not always generalize best to the test set, indicating sensitivity to threshold selection and distributional variation.

4.3 Lexical, Embedding, kNN, and Additional Models

To introduce diverse inductive biases, we evaluate complementary model families.

TF-IDF + Logistic Regression. We construct word (1–2 grams) and character (3–5 grams) TF-IDF features and train a balanced logistic regression classifier. Lexical models are sensitive to explicit trigger terms, but less robust to paraphrasing.

Embedding + Linear Classifier. We extract fixed transformer embeddings, apply pooling and normalization, and train a calibrated linear SVM. Freezing the encoder reduces variance while retaining competitive development performance.

kNN-Derived Features. We compute local neighborhood statistics in embedding space, such as the positive neighbor ratio and class-conditional similarity averages. These features capture distributional structure that complements lexical signals. Among single models, DeBERTa (pretrained) + kNN achieves the highest development score ($F_1 = 0.8318$).

LLM-based classifier. We also evaluate a Qwen3-14B model fine-tuned with LoRA for sequence-level classification, using a task-specific fine-tuning setup for the detection task, and included as an additional neural baseline.

4.4 Calibration-Aware Logit Ensembling

Neural networks are often poorly calibrated (Guo et al., 2017), and heterogeneous classifiers exhibit distinct confidence scales. Ensemble methods are known to improve robustness (Lakshminarayanan et al., 2017).

Our final ensemble combines predictions from seven heterogeneous models introduced in Sections 4.2 and 4.3: (1) TF-IDF + Logistic Regression, (2) DeBERTa-v3-large fine-tuned classifier, (3) embedding-based linear classifier, (4) kNN-derived feature-based model, (5) RoBERTa-based classifier, (6) embedding-based model trained on GPT-refined data, and (7) Qwen-based model with LoRA fine-tuning.

Given model probabilities p_i, we compute the ensemble prediction as:

$$p_{\text{ens}} = \sigma\left(\sum_{i=1}^{M} w_i \log \frac{p_i}{1-p_i}\right)$$

where probabilities are clipped to avoid numerical instability.

We assign a smaller weight to the Qwen-based model ($w_{\text{Qwen}} = 0.05$), while distributing the remaining weight uniformly across the other six models ($w_i = 0.95/6$). This design reflects the comparatively lower standalone performance and higher variance observed for the Qwen-based model, while preserving its complementary contribution within the ensemble.

We do not apply explicit temperature scaling, as preliminary experiments indicated that logit-space aggregation already mitigates calibration differences across heterogeneous models.

The decision threshold t is selected based on development performance by evaluating a small neighborhood of candidate values $\{0.54, 0.55, 0.56, 0.57\}$. Within this range, performance differences are relatively small, and we adopt $t = 0.55$ as the final threshold, as it provides stable performance across models and configurations.

We also explored alternative ensemble strategies, including uniform weighting, majority voting, and probability-space averaging. However, logit-level aggregation with heuristic weighting consistently provided a better trade-off between robustness and performance.

5 Experiments

5.1 Experimental Setup

Data preprocessing. We use the official train and development splits provided by the shared task. For Conspiracy Detection, we exclude samples labeled as "Can't tell" and remove duplicate entries based on Reddit IDs, retaining only the first occurrence. All preprocessing steps, including deduplication and label filtering, are applied strictly to the training and development data and do not affect the official test set.

Conspiracy Marker Extraction. For Conspiracy Marker Extraction, we train multi-label token classification models using the same transformer architecture as the classification models. Each token is assigned binary labels for each marker type, and models are optimized using binary cross-entropy loss over token–label pairs. For the main DeBERTa-v3-large model, we use a learning rate of 2×10^{-5}, batch size of 24, and train for up to 10 epochs with a linear learning rate scheduler and a warmup ratio of 0.1. Token-level predictions are thresholded and merged into spans based on contiguous segments, following the IoU-based evaluation protocol.

Conspiracy Detection. For Conspiracy Detection, transformer-based models are fine-tuned using binary cross-entropy loss with independent sigmoid outputs. We use AdamW with a learning rate of 2×10^{-5}, batch size of 24, and train for 3 or 10 epochs depending on the configuration. Gradient accumulation is applied (steps = 1 or 3), resulting in an effective batch size of 24 or 72. We use a linear learning rate scheduler with a warmup ratio of 0.1, weight decay of 0.05, and a maximum input length of 512 tokens. All experiments are conducted with a fixed random seed.

For embedding-based classifiers, we extract sentence representations from fine-tuned transformer models and train downstream classifiers. We explore several pooling strategies (CLS, mean, last4, CLS+mean), and adopt last4 pooling with L2 normalization in the main setting. A LinearSVC classifier is used, and probabilities are obtained via CalibratedClassifierCV with sigmoid calibration.

For kNN-based models, we compute nearest neighbors in the embedding space using cosine similarity ($k = 20$). Features include the ratio of positive neighbors, average similarity to positive and negative neighbors, and their margin. These features are used as input to a logistic regression classifier.

We also construct a GPT-refined dataset by re-annotating training instances using GPT-5 via the Azure OpenAI API. Predictions with confidence scores above 0.95 are used to update labels. This high threshold is chosen to ensure that only highly confident predictions are applied, thereby minimizing the risk of introducing additional noise into the training data. As a result, only a small number of corrections are made (14 instances in the training set), while no changes are applied to the development set. Models trained on the refined dataset are treated as additional components in the ensemble.

Model	Dev F1	Test F1
DeBERTa (baseline)	0.2033	0.1400
DeBERTa + Multi-label	0.2297	0.1813
RoBERTa-base	0.1703	0.1111
ELECTRA-base	0.1571	0.1285

Table 1: Conspiracy Marker Extraction results.

Model	Dev F1	Test F1
Single Models (Original Data)		
TF-IDF + LR	0.7851	0.6877
DeBERTa-v3 Large	0.8100	0.7409
DeBERTa (pretrained) + SVM	0.8310	0.7495
DeBERTa (pretrained) + kNN	0.8318	0.7372
Qwen3-14B (LoRA)	0.7950	0.7326
GPT-Refined (0.95)		
DeBERTa (pretrained) + SVM	0.8073	0.7466
DeBERTa (pretrained) + kNN	0.8181	0.7458
Final Ensemble		
Logit Ensemble (t=0.55)	0.8182	**0.7620**

Table 2: Conspiracy Detection results (weighted F_1).

Evaluation. For Conspiracy Marker Extraction, we report macro F1 based on token-level IoU. For Conspiracy Detection, we report weighted F1 as the primary metric, along with precision and recall for the positive class.

5.2 Conspiracy Marker Extraction Results

Table 1 summarizes development results for the span extraction approaches.

The multi-label DeBERTa model achieves $F_1 = 0.2297$ (Macro) on the development set and $F_1 = 0.1813$ (Macro) on the test set, consistently outperforming the single-label baseline ($F_1 = 0.2033$ Dev, $F_1 = 0.1400$ Test).

5.3 Conspiracy Detection Results

Table 2 reports both single-model and ensemble performance for Conspiracy Detection.

Although DeBERTa (pretrained) + kNN achieves the highest development performance among single models ($F_1 = 0.8318$), it does not yield the strongest test performance. The logit-level ensemble, which attains a lower development score ($F_1 = 0.8182$), achieves the best test result ($F_1 = 0.7620$). This reversal indicates that selecting a single development-optimal model may lead to suboptimal generalization under distribution shift.

5.4 Analysis

For Conspiracy Marker Extraction, the multi-label DeBERTa model consistently outperforms the single-label baseline, indicating that multi-label token modeling is better suited for overlapping span prediction. However, overall performance remains low, reflecting the difficulty of precise boundary detection under the IoU-based evaluation.

For Conspiracy Detection, the development-optimal model does not necessarily achieve the best test performance. While the DeBERTa-based kNN model yields the highest development score, the logit-level ensemble achieves the best test performance, demonstrating improved robustness under distribution shift.

These results highlight the benefit of combining heterogeneous models. Lexical models capture explicit trigger patterns, whereas embedding-based and transformer models generalize better to implicit or paraphrased expressions. Logit-level aggregation further stabilizes predictions by reducing calibration differences across models.

6 Error Analysis

6.1 Conspiracy Marker Extraction

The primary source of error is boundary mismatch under IoU-based evaluation. Even semantically correct spans may be penalized due to slight token alignment differences.

We also observe type confusion, particularly between Action and Effect, where distinguishing between intended actions and resulting outcomes often requires broader discourse context.

We observe variation across marker types, with some markers (e.g., Actor and Victim) being easier to detect than others (e.g., Action, Effect, and Evidence), suggesting differences in difficulty across marker types.

Feature-augmented and retrieval-based methods introduce additional false positives due to surface-level matching without sufficient contextual grounding, which explains their lower performance compared to the multi-label DeBERTa model.

6.2 Conspiracy Detection

False positives are often triggered by the presence of conspiracy-related vocabulary without explicit endorsement. Lexical models are particularly sensitive to such surface-level triggers, leading to errors when posts mention controversial actors or institutions in a descriptive or critical manner.

False negatives typically occur in longer posts where conspiratorial intent is implied rather than explicitly stated. In such cases, the absence of clear trigger terms leads to lower confidence predictions. Sarcasm and narrative-style framing further increase ambiguity.

7 Conclusion

We presented CuriosAI's system for SemEval-2026 Task 10, addressing both Conspiracy Marker Extraction and Conspiracy Detection.

For Conspiracy Marker Extraction, multi-label token classification with a bidirectional transformer consistently outperformed simpler baselines, highlighting the importance of contextual token modeling for overlapping spans.

For Conspiracy Detection, we showed that development-optimal models do not necessarily generalize best, and that logit-level ensembling of heterogeneous models improves robustness under distribution shift.

Overall, our results emphasize the importance of combining contextual modeling with calibration-aware ensembling in shared task settings.

8 Limitations

For Conspiracy Marker Extraction, performance remains limited under IoU-based evaluation due to sensitivity to span boundary alignment and overlapping labels.

For Conspiracy Detection, the relatively small and imbalanced development set may increase sensitivity to model and threshold selection.

Synthetic data generated by large language models did not consistently improve performance.

Finally, our evaluation is limited to the shared task dataset, and generalization to other domains or platforms remains an open question.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Ivan Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of*

the 14th International Workshop on Semantic Evaluation (SemEval).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of CHI*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval)*.

Mattia Samory, Felix Soldner, and Veronika Batzdorfer. 2026. SemEval-2026 task 10: PsyCoMark – psycholinguistic conspiracy marker extraction and detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.