HU at SemEval-2026 Task 10: Psycholinguistic Conspiracy Marker Extraction and Detection

Muhammad Quddussi Kashaf[1] **Shahmir Chaudhry**[1] **Marium Zeeshan**
Nahyan Javed **Sandesh Kumar** **Abdul Samad**
Habib University, Karachi, Pakistan
{kashafmuhammad, shahmirmustafa02}@gmail.com

Abstract

Modern media poses a complex challenge to verifying the credibility of information and public discourse due to the advent of conspiracy theory content. This paper presents our methodology in "SemEval-2026 Task 10: Psycholinguistic Conspiracy Marker Extraction and Detection". It consists of two subtasks: extracting psycholinguistic markers from text using Named Entity Recognition (NER) techniques, and classifying Reddit comments as conspiratorial or non-conspiratorial. Our approach involved: (1) diverse extraction methodologies, including traditional bio tagging schemes, the GlobalPointer framework, and the GLiNER2 architecture, (2) data augmentation and synthetic data generation via Large Language Models (LLMs), and (3) evaluating various transformer-based models, such as DistilBERT and Covid Twitter-BERT. Our final system achieves a macro F1 score of 0.26 on Subtask 1 and 0.76 on Subtask 2.

1 Introduction

The digital age propelled the rapid growth of social media, facilitating the spread of conspiracy theory (CT) content. These narratives typically involve suspecting individuals of malicious intent by hidden actors and/or provide alternative explanations that oppose mainstream beliefs. Detecting such content requires the use of sophisticated natural language processing techniques and models, capable of identifying subtle linguistic nuances and complex psychological patterns that characterize conspiratorial thinking.

This paper provides a structured framework for analyzing conspiracy-related rhetoric in Reddit statements, following the SemEval task description (Samory et al., 2026; Ghosh et al., 2026). The task is divided into two main parts:

Extracting Conspiracy Markers (NER Task): This task focuses on identifying five psycholinguistic markers: *Actor, Action, Effect, Evidence*, and *Victim*. The challenge is characterized by span-level extraction, in which a single document may contain multiple, potentially overlapping markers. System performance is evaluated using an overlap-based macro F1 score.

Conspiracy Detection (Classification Task): This subtask focuses on document-level classification, determining whether a Reddit post expresses a belief in a conspiracy. Performance is assessed using a weighted F1 score across two classes: "Yes" and "No", excluding the "Can't Tell" class according to task guidelines.

2 Related Work

2.1 Named Entity Recognition (NER)

NER has evolved massively with the advent of deep learning architectures. Previously, traditional approaches relied on feature engineering and statistical models such as Hidden Markov Models and Conditional Random Fields, such as the works of (Zhou and Su, 2002; Khan et al., 2022). At present, Modern NER systems leverage contextualized representations from pre-trained language models. Factors such as the choice of annotation scheme significantly impact NER system performance, as shown by (Alshammari and Alanazi, 2021)

Recent advances have focused on span-based approaches that directly model entity boundaries, bypassing the limitations of sequential BIO-tagging. The Global Pointer framework (Su et al., 2022; Zhang et al., 2023) is an efficient approach that treats NER as a multi-label classification problem over token pairs, naturally handling nested and overlapping entities. Furthermore, the emergence of the GLiNER model (Zaratiana et al., 2025) has introduced a bi-encoder architecture that utilizes semantic label descriptions. It enables zero-shot generalization across diverse and niche domains, a capability relevant to identifying nuanced psy-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1672–1677

Table 1: Comparison of NER Approaches Across Different Domains

Study	Domain	Dataset	Method	F1-Score	Additional Comments
(Meenachisundaram et al., 2023)	Biomedical	GENIA	TCN-CRF	91.54%	Temporal convolution
		CoNLL-2003	TCN-CRF	85.78%	Bio-tagging scheme
(Su et al., 2022)	General	People's Daily	Global Pointer	95.51%	Span-based, RoPE
		CLUENER	Global Pointer	79.44%	Multi-label loss
		CMeEE	Global Pointer	65.98%	Nested entity support
(Zhang et al., 2023)	Legal (Chinese)	LegalCorpus	RoBERTa-GP	90.54%	Word+Char embeddings
			RoBERTa-Biaffine	89.04%	Skip-Gram, Dictionary
(Jain and Sharma, 2024)	Legal (Indian)	Indian Court Judgments (46,545 examples)	RoBERTa-GCN-CRF	87.84%	14 entity types
			RoBERTa-GCN	86.40%	Graph convolution
			Text-GCN	83.50%	Fine-grained entities
			BiLSTM-CRF	85.34%	Sequential dependencies
(Tavan and Najafi, 2022)	Multilingual	MultiCoNER-2022	T5+Transformer+Subtoken	71.45%	Subtoken check, Byte Pair embeddings, multilingual evaluation

cholinguistic markers.

Success in specialized fields, such as legal and biomedical NER, underscores the need to capture context-dependent patterns. As shown in Table 1, architectures such as RoBERTa, combined with Global Pointer or Graph Convolutional Networks, have set benchmarks for processing complex structures in these domains (Zhang et al., 2023; Jain and Sharma, 2024), further highlighting the potential of these approaches for our task.

2.2 Conspiracy Theory Detection

CT detection, a subsect of sentiment analysis, presents its own challenges, as the content may be factually accurate in parts while weaving alternative explanatory narratives that invoke hidden actors and malicious intent. Psycholinguistic approaches to CT analysis have identified characteristic patterns, including high levels of negative emotion, distrustful expressions, and alternative explanations (Giachanou et al., 2023), which LLMs can extract.

Transformer-based models have become the standard for this task. (Moosleitner and Murauer, 2021) demonstrated the efficacy of BERT-base and RoBERTa in classifying conspiracy-related tweets, while (Huang et al., 2024) explored binary classification of public health messages into critical versus conspiracy narratives. These studies highlighted the importance of domain-adapted variants such as Covid-Twitter-BERT (CT-BERT), which was pre-trained on a massive corpus of COVID-19 discourse to capture the rhetorical styles and specialized vocabulary characteristics of conspiracy theories. Recent iterations, such as CT-BERT-PRCT (Marino et al., 2025), further specialize in detecting specific conspiracy tropes across social platforms, suggesting that domain-specific pretraining significantly improves classification robustness across social platforms.

3 Dataset

The PsyCoMark dataset provided for SemEval 2026 Task 10 comprises Reddit submission statements—user-written summaries accompanying media uploads—annotated with psycholinguistic conspiracy markers and belief labels. The dataset comprises 4,855 annotated entries, with 1715 labeled "Yes", 2263 labeled "No", and 877 labeled "Can't tell".

3.1 Data Collection and Annotation

The dataset spans a decade of Reddit discourse (March 2013 to December 2023), with a specific sampling strategy to ensure a high density of conspiratorial rhetoric. Approximately 25% of the comments were sourced from the *r/conspiracy* subreddit, while the remainder were gathered from over 190 diverse subreddits.

For Subtask 1, annotators labeled span-level psycholinguistic markers representing five key mechanisms of conspiratorial thinking:
Actor: Entities or groups attributed with agency in the conspiracy narrative.
Action: Activities or behaviors attributed to actors.
Effect: Consequences or outcomes of the conspiratorial actions.
Evidence: Information presented to support conspiracy claims.
Victim: Entities or groups portrayed as targets or harmed parties.

3.2 Preprocessing and Statistics

To ensure content quality and model compatibility, the raw text underwent several preprocessing steps: comments were filtered to a length between 160 and 1,000 characters to ensure sufficient context without exceeding transformer token limits, markdown was converted to plain text, URLs were masked with `[URL]` tokens, and quoted text was re-

moved to isolate the original author's language. As shown in Figure 1, *Action* and *Actor* are the most frequent markers in ST1, while *Victim* is the least frequent. For ST2, the dataset is moderately imbalanced, with *No* being the majority class compared to *Yes* and *Can't Tell*.

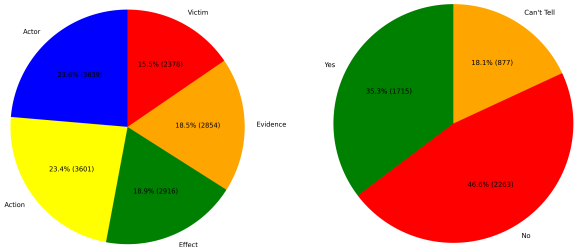

Figure 1: Distribution of Conspiracy Marker Instances and Classes for ST1 and ST2.

4 System Overview

In this section, we describe our methodology for marker extraction and conspiracy classification. Our methodology addresses two primary challenges: the inherent semantic complexity of conspiratorial rhetoric and the scarcity of domain-specific annotated data, which limits the model's ability to generalize well across this domain.

The complexity of implicit conspiracy theories introduces an additional level of difficulty to the entity extraction. Thus, we adopt a framework centered on the GLiNER2 architecture (Zaratiana et al., 2025), which generalizes across domains and is thus suitable for our task. Consequently, drawing on the literature, we employ domain-specific models, such as CT-BERT, for conspiracy theory detection.

The system architecture is built upon a two-stage pipeline. First, we enhance the models' representational power through diverse data augmentation strategies, including LLM-based synthetic generation. Second, we perform task-specific fine-tuning on the augmented and original datasets, ensuring that the models adapt to both artificial examples and the nuances of real-world social media discourse.

4.1 Data Augmentation Strategies

Data augmentation in the context of CT analysis is particularly challenging due to the high risk of semantic drift; simple modifications can strip a sentence of its conspiratorial intent. While we initially evaluated standard techniques such as Contextual Word Embedding and Synonym Replacement (Huang et al., 2025), given their widespread use, these methods failed to produce high-quality, label-preserving samples. Consequently, we utilized two more robust approaches:

Synthetic Generation via LLMs: We employed **Llama-3-70B** to generate a synthetic corpus that mimics the rhetorical structure of the original dataset. As illustrated in Figure 2, the prompt provides the original text, five entity definitions with task-specific descriptions, and requires JSON-structured output, as in (Dao et al., 2025), ensuring that the LLM captured the psycholinguistic essence of the categories rather than just literal entity types. To mitigate annotation errors, we randomly sampled and manually verified a subset of examples, verifying that entity spans appeared as exact substrings in the generated text.

Prompt

```
You are an expert data generator for Named Entity Recognition (NER).
ORIGINAL TEXT:
    "{entry['text']}"
TASK:
1. Write a NEW synthetic paragraph based on the same conspiracy
   theme/topic as the original.
2. The new text must be distinct.
3. Extract entities fitting these categories: {ENTITY_TYPES}.
4. Identify NESTED entities if applicable.
ENTITY DEFINITIONS:
- Actor: The person, group, or org responsible for a harmful agenda.
- Action: The behavior attributed to the actor that causes harm.
- Effect: The negative consequence resulting from the actor's action.
- Victim: The person, group, or population harmed or targeted.
- Evidence: Any claim, quote, or reasoning supporting the conspiracy.
OUTPUT FORMAT (JSON ONLY):
{
  "text": "Your new synthetic text here",
  "entities": [
    {"type": "Actor",  "text": "exact substring"},
    {"type": "Effect", "text": "exact substring"}
  ]
}
```

Figure 2: Structure of the prompt provided to Llama-3-70B for synthetic data generation.

Cross-Domain Dataset Mapping: A second augmentation stream was derived from the PAN 2024 Oppositional Thinking dataset (Korenčić et al., 2024), which shares significant overlap with our task. We developed a mapping schema to align the PAN labels with our labels, as shown in Table 2. This allowed us to incorporate both English and Spanish examples, thereby providing the models with broader exposure to linguistic markers.

4.2 Subtask 1

For psycholinguistic marker extraction, we explored four distinct methodologies: traditional BIO-tagging, the Global Pointer framework, a T5-based generative approach, and GLiNER2.

BIO-Tagging Baseline: Although traditional BIO-tagging cannot handle nested or overlapping enti-

PsyCoMark Label	PAN 2024 Term
Actor	Agent
Action	Objective
Effect	Negative Effect
Actor	Facilitator
Victim	Victim
Conspiracy (Yes)	Conspiracy
No Conspiracy (No)	Critical

Table 2: Label mapping scheme for data integration from the PAN 2024 dataset.

ties, we use it as our primary baseline. The scheme comprised 11 distinct classes (B and I tags for the five markers each, plus the O tag). To mitigate the severe class imbalance caused by the dominant **Outside (O)** class, we applied class weighting during training. Furthermore, we integrated a **subtoken check** mechanism (Tavan and Najafi, 2022) to ensure boundary consistency when tokens are split into sub-words by the tokenizer. We evaluated several backbones for this setup, including RoBERTa-base, DeBERTa-base, and the T5-large encoder. For the T5 variant, we appended an additional encoder layer to form a multi-encoder architecture, following (Tavan and Najafi, 2022).

Global Pointer: To address the limitations of sequential tagging regarding overlapping spans, we implemented the Global Pointer framework (Su et al., 2022), which is highly effective for extracting niche and nested entities (Zhang et al., 2023). For this method, we used transformer-based models, namely *BERT, T5,* and *DeBERTa* as contextual embedding generators. Rotary Position Embedding (RoPE) was incorporated, enabling better modeling of long-range dependencies in entity span extraction (Su et al., 2021) We optimized the models using the multilabel categorical cross-entropy loss since Global Pointer predicts entity spans across multiple types simultaneously. Unlike standard categorical cross-entropy, the multilabel variant does not force predictions to be in a single class, making it ideal for NER tasks where multiple labels can apply simultaneously (GeeksforGeeks, 2024).

Generative Extraction via T5: Third, we use **T5-base** to cast NER as a unified **text-generation** task, inspired by (Wang et al., 2023). It achieves overlap handling inherently through the structured output format. To frame extraction as a sequence generation task, we use T5's unique sentinel tokens ($\langle extra\_id\_n \rangle$). Ground-truth markers were first sorted by their `startIndex`. Each entity is then encapsulated within a tag-pair schema, where an even-indexed sentinel (e.g., $\langle extra\_id_{2i} \rangle$) denotes the beginning of the text span, and an odd-indexed sentinel (e.g., $\langle extra\_id_{2i+1} \rangle$) serves as a separator for its semantic category. For sentences devoid of conspiratorial elements, the model is trained to generate the explicit null token string `"none"`.

GLiNER2 with Descriptive Prompting: Finally, our best-performing approach involved fine-tuning the GLiNER2-large model. GLiNER2 leverages a bi-encoder architecture that jointly embeds the input text and the target entity labels, which we fine-tune for a few epochs, using the final dataset. During inference, we conducted an ablation study comparing two prompting strategies: (1) providing the model with standard entity names (e.g., *Actor*, *Victim*), and (2) providing the model with entity descriptions (e.g., *"Actor: Entities or groups attributed with agency in the conspiracy narrative"*).

4.3 Subtask 2

Sub-task 2 is a classic binary classification problem: identifying whether a given Reddit comment is a CT or not. We experimented with several transformer-based architectures to establish a baseline and explored the impact of domain-specific pre-training, including BERT Base, DistilBERT, XLM-ROBERTa, COVID-Twitter-BERT (CT-BERT), and Conspiracy Theory BERT (CT-BERT-PRCT).

Our primary approach was a staged pipeline: training the model sequentially on augmented data (AD), PAN English data, PAN Spanish Data (optionally), and finally on the original data. The strategy was to enable the model to learn an overall trend across the various ADs, and then fine-tune on the original content.

A secondary approach employed a mixed dataset of augmented + original samples. After training for several epochs, we employed a cosine learning rate scheduler to fine-tune the model; however, this approach yielded suboptimal results.

5 Results

In this section, we will evaluate our approaches for both subtasks on the SemEval *Codabench* test and development datasets.

5.1 Subtask 1

The BIO-tagging method performed well during training; however, across all models, its validation loss continued to increase. This indicated severe overfitting, as the model collapsed into predicting the dominant class *(Outside Entity)* to minimize loss. To address this, the T5-Large model, augmented with an additional encoder layer and a subtoken check mechanism, yielded relatively better performance. Table 3 presents the overlap-based macro F1 scores of both baseline approaches on the development set, alongside the official getting-started baseline provided by the task organizers, which serves as a reference point for comparison across all task participants.

Model	Overlap Macro F1
Baseline	0.15
RoBERTa-Large	0.14
T5-Large + Encoder	0.21

Table 3: Performance on the Dev dataset using BIO-Tagging. *Baseline* refers to the official getting-started script provided by the task organizers.

The Global Pointer (GP) method showed a significant increase in macro-F1 score on the training data. Table 4 details the development set scores for the various backbones tested with the GP framework, with DeBERTa-Base achieving the highest performance.

Model Backbone	Overlap Macro F1
BERT-Base + GP	0.17
T5-Large + GP	0.19
DeBERTa-Base + GP	0.22

Table 4: Performance of Global Pointer models on the Dev dataset.

Concurrently, evaluating the generative T5-Base model on a unified text-to-text extraction task yielded a macro F1 score of 0.19 on the development dataset.

Ultimately, finetuning the GLiNER2 model yielded the best overall results. As stated before, during inference, the model can be prompted with either the literal entity names or detailed semantic descriptions of those entities. Descriptive prompting proved highly beneficial for this task, as the psycholinguistic markers (e.g., Actor) deviate from their conventional semantic definitions in our case. Table 5 compares GLiNER2's performance on the development and test datasets.

GLiNER2 Prompting Setting	Dev	Test
Without Descriptions	0.24	0.22
With Descriptions	0.27	0.26

Table 5: GLiNER2 performance comparison (Overlap Macro F1) with and without semantic entity descriptions.

5.2 Subtask 2

In line with the literature, models with prior success in classifying CT content were deployed. As expected, CT-BERT-PRCT offered the highest F1 score. The results for all models are outlined in Table 6:

Model	Dev	Test	AD Test
Baseline	0.75	0.71	—
BERT-base	0.72	0.72	0.70
DistilBERT	0.75	0.73	0.71
XLM-RoBERTa (Base)	0.73	0.72	0.69
CT-BERT	0.79	0.76	0.74
CT-BERT-PRCT	0.80	0.76	0.75

Table 6: F1 Performance Comparison of Models on Dev, Test, and Augmented Approach on Test Datasets.

After training on the AD, we observed several trends. Firstly, the models overfit the training data, with a drastic reduction in training loss and an increase in validation loss. This may be due to imbalances unintentionally introduced by the data augmentation. For instance, merging the PAN 2024 data leads to the dataset having more COVID-19-based conspiracy theories and the dataset learning these behaviors rather than CT as a whole. Secondly, using multiple augmented samples per Reddit comment of the original data led to similar content across the augmented dataset, further causing overfitting. Lastly, using the Spanish data samples improved the initial validation loss but later overfitted, yielding similar results. As a global trend, the augmentation reduced model scores as outlined in Table 6 AD Test.

6 Future Work

We can explore the ROBERTa Model with Global Pointer and CRF layers methodology for marker

extraction, leveraging its strong performance in specialized domains. Additionally, the T5 model with a transformer architecture can be integrated, as suggested in the literature, to further enhance our results on the development dataset.

For both subtasks, better augmentation techniques need to be explored to prevent overfitting. For conspiracy detection, larger models fine-tuned for this task can also be explored by quantizing to a lower memory footprint; however, the performance trade-off needs to be monitored. These efforts aim to develop effective methods to counter online CT content.

References

Nasser Alshammari and Saad Alanazi. 2021. The impact of using different annotation schemes on named entity recognition. *Egyptian Informatics Journal*, 22:295–302.

An Dao, Hiroki Teranishi, Yuji Matsumoto, Florian Boudin, and Akiko Aizawa. 2025. Overcoming data scarcity in named entity recognition: Synthetic data generation with large language models. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 328–340, Viena, Austria. Association for Computational Linguistics.

GeeksforGeeks. 2024. Categorical crossentropy in multiclass classification.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. 2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2023. Detection of conspiracy propagators using psycho-linguistic characteristics. *Journal of Information Science*, 49(1):3–17.

Jiahong Huang, Zhongyuan Han, Ruihao Zhu, Mingcan Guo, and Kaiyin Sun. 2024. Conspiracy theory text classification based on ct-bert and beto models. *Working Notes of CLEF*.

Yi Huang, Yuhan Gao, and Chengjuan Ren. 2025. A survey of data augmentation in named entity recognition. *Neurocomputing*, 651:130856.

Arihant Jain and Raksha Sharma. 2024. Enhancing legal named entity recognition using roberta-gcn with crf: A nuanced approach for fine-grained entity recognition. *Lecture notes in computer science*, pages 261–267.

Wahab Khan, Ali Daud, Khurram Shahzad, Tehmina Amjad, Ameen Banjar, and Heba Fasihuddin. 2022. Named entity recognition using conditional random fields. *Applied Sciences*, 12(13).

D. Korenčić, B. Chulvi, X. Bonet Casals, M. Taulé, and P. Rosso. 2024. Pan24 oppositional thinking analysis (1.0.0) [data set].

Erik Bran Marino, Renata Vieira, and Davide Bassi. 2025. One model to detect them all? comparing llms, bert and traditional ml in cross-platform conspiracy detection. *arXiv preprint*.

Thiyagu Thavittupalayam Meenachisundaram, Sangeetha Ramachandran, Sudhakaran Gajendran, Om Kumar Chandra Umakantham, and Sathish Kuppani. 2023. Biomedical named entity recognition using tcn approaches and bio tagging. *Journal of Autonomous Intelligence*, 6.

Manfred Moosleitner and Benjamin Murauer. 2021. On the performance of different text classification strategies on conspiracy classification in social media. In *MediaEval*.

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)*.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition.

Ehsan Tavan and Maryam Najafi. 2022. Marsan at semeval-2022 task 11: Multilingual complex named entity recognition using t5 and transformer encoder. *ACL Anthology*, pages 1639–1647.

Shuhe Wang, Xiaoya Li, Rongbin Ouyang, Fei, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. Gliner2: An efficient multi-task information extraction system with schema-driven interface. *Preprint*, arXiv:2507.18546.

Xinrui Zhang, Xudong Luo, and Jiaye Wu. 2023. A roberta-globalpointer-based method for named entity recognition of legal documents. *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 473–480, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.