

CCNU at SemEval-2026 Task 10: Conspiracy Marker Extraction and Detection via Multi-task Learning and LLM-based Data Augmentation

Zijun Wang¹ and Guanyi Chen^{2*}

¹Wollongong Joint Institute, Faculty of Artificial Intelligence in Education,
Central China Normal University

²Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning,
National Language Resources Monitoring and Research Center for Network Media,
School of Computer Science, Central China Normal University
zjwang@mails.ccnu.edu.cn, g.chen@ccnu.edu.cn

Abstract

This paper presents the system of CCNU for SemEval-2026 Task 10: Psycholinguistic Conspiracy Marker Extraction and Detection. The task requires identifying fine-grained conspiracy markers that characterize conspiracy thinking, as well as determining whether a Reddit comment constitutes conspiratorial discourse. For Conspiracy Marker Extraction (Subtask 1), we adopt a Unified Multi-Task Sequence Labeling Framework that jointly models multiple conspiracy markers within a single labeling space. This formulation enables collaborative learning across marker types while maintaining a compact architecture. For Conspiracy Detection (Subtask 2), we formulate the problem as sentence-level classification. Across both subtasks, we apply data augmentation powered by large language models and ensemble inference to improve robustness and generalization. Our system achieves strong performance on Subtask 1, ranking 3rd on the official test set, and delivers competitive results on Subtask 2.

1 Introduction

Conspiracy thinking refers to a style of reasoning in which events are interpreted as the result of coordinated and hidden actions by powerful actors (Douglas et al., 2017; Douglas and Sutton, 2023). Rather than relying on verifiable evidence, such narratives often attribute secret intentions to governments, institutions, or social groups. For example, a comment may claim that “a foreign government secretly orchestrated political unrest in another country, destabilizing the elected leadership while suppressing the truth through mainstream media”. This example illustrates a core characteristic of conspiratorial discourse: the attribution of hidden coordination and malicious intent to identifiable actors, constructing a narrative of secret manipulation that is resistant to verifiable counter-evidence.

*Corresponding Author

SemEval-2026 Task 10, namely the PsyCoMark task, addresses the analysis of conspiratorial discourse through two complementary subtasks: *Conspiracy Marker Extraction* and *Conspiracy Detection*. Conspiracy Marker Extraction focuses on identifying fine-grained conspiracy markers that characterize conspiratorial narratives. These markers correspond to narrative roles, including *Actor*, *Action*, *Victim*, *Effect*, and *Evidence*. In the example above, “a foreign government” functions as the *Actor*; “secretly orchestrated political unrest” represents the *Action*; “another country” serves as the *Victim*; “destabilizing the elected leadership” reflects the resulting *Effect*; and “suppressing the truth through mainstream media” provides contextual *Evidence*. Extracting such markers requires detecting short and often implicit spans embedded in informal text (Fong et al., 2021). Moreover, markers frequently appear in incomplete or implicit forms, further complicating precise boundary detection and role assignment.

Conspiracy Detection, in contrast, operates at the sentence-level and aims to determine whether a comment constitutes conspiratorial discourse. In the example above, the coordinated attribution of hidden intent to a powerful actor and the framing of deliberate suppression together signal conspiratorial discourse. This task requires modeling global narrative coherence, pragmatic framing, and communicative intent. Conspiratorial discourse is often subtle, indirect, or deliberately ambiguous, relying on insinuation rather than explicit accusations. Such ambiguity complicates decisive categorization, as conspiratorial meaning may emerge through tone, implication, stance, or rhetorical questioning rather than clearly articulated narrative roles. Consequently, Conspiracy Detection demands discourse-aware semantic modeling that captures implicit framing beyond surface marker identification.

These challenges are further amplified by limited

annotated data and noisy supervision, as conspiratorial markers are relatively rare and unevenly distributed across comments. The sparsity of certain role types and the subtlety of conspiratorial framing restrict the coverage of diverse narrative patterns, increasing the risk of overfitting and unstable predictions. Motivated by both the interdependence of narrative roles and the scarcity of reliable training signals, we adopt a Unified Multi-Task Sequence Labeling Framework to jointly model multiple marker types within a unified representation space. To address the limited coverage caused by data sparsity and role imbalance, we further incorporate large language model-based data augmentation to enrich marker distributions and expose the model to more varied conspiratorial constructions. In addition, we apply ensemble inference strategies to enhance robustness and stabilize predictions across both subtasks. Experimental results demonstrate that these components consistently improve performance across both Conspiracy Marker Extraction and Conspiracy Detection.

2 Task Description

The PsyCoMark task focuses on the analysis of psycholinguistic patterns associated with conspiracy thinking in everyday conversations (Samory et al., 2026). The dataset consists of Reddit comments annotated according to psychological theories of conspiracy expression. The task is divided into two subtasks.

Subtask 1: Conspiracy Marker Extraction aims to extract fine-grained conspiracy markers from text. Each comment may contain zero or more markers belonging to five predefined roles: *Actor*, *Action*, *Effect*, *Evidence*, and *Victim*. The task is formulated as a span extraction problem, where systems are required to identify both the boundaries and the corresponding roles of each marker. Performance is evaluated using a token-based overlap F1 score, where predicted and gold spans of the same type are matched based on token-level Intersection-over-Union ($\text{IoU} \geq 0.5$).

Subtask 2: Conspiracy Detection focuses on determining whether a given Reddit comment constitutes conspiratorial discourse. This subtask is formulated as a binary classification problem at the sentence-level. Systems are evaluated using Macro F1-score, which accounts for class imbalance and measures overall classification performance across

both classes.

3 Methodology

We address the two subtasks using task-specific modeling strategies tailored to their supervision granularity and prediction objectives. Conspiracy Marker Extraction is formulated as a token-level sequence labeling problem, emphasizing structured role modeling and boundary precision. Conspiracy Detection is treated as a sentence-level classification task, where data-centric strategies play a central role in performance improvement.

3.1 Subtask 1: Conspiracy Marker Extraction

Conspiracy Marker Extraction aims to identify fine-grained conspiracy markers at the token-level, where each marker corresponds to a specific role.

3.1.1 Unified Multi-Task Sequence Labeling Framework

We formulate Conspiracy Marker Extraction as a unified multi-task sequence labeling problem using a shared BIO tagging scheme. Specifically, all marker roles are mapped into a single label space consisting of one outside label (O) and role-specific begin and inside labels for each marker type. A single transformer encoder is shared across all roles, followed by a token-level classification head that predicts labels from this unified space.

This formulation enables collaborative learning across psycholinguistic roles through shared contextual representations. By jointly optimizing predictions for all marker types, the model can capture cross-role dependencies, such as the frequent co-occurrence and proximity of Actors and Actions or Actions and Effects, which are difficult to model when each role is learned in isolation. This collaborative formulation encourages the encoder to learn role-aware representations that generalize across marker types.

In practice, the dataset does not contain overlapping marker spans, allowing the unified BIO tagging scheme to be applied without ambiguity. During training, standard cross-entropy loss is computed over the unified label space, and the model is optimized end-to-end. This design significantly reduces model complexity and training cost compared to role-specific models, while maintaining competitive or superior extraction performance.

Overall, the proposed multi-task sequence labeling framework provides an efficient and effective

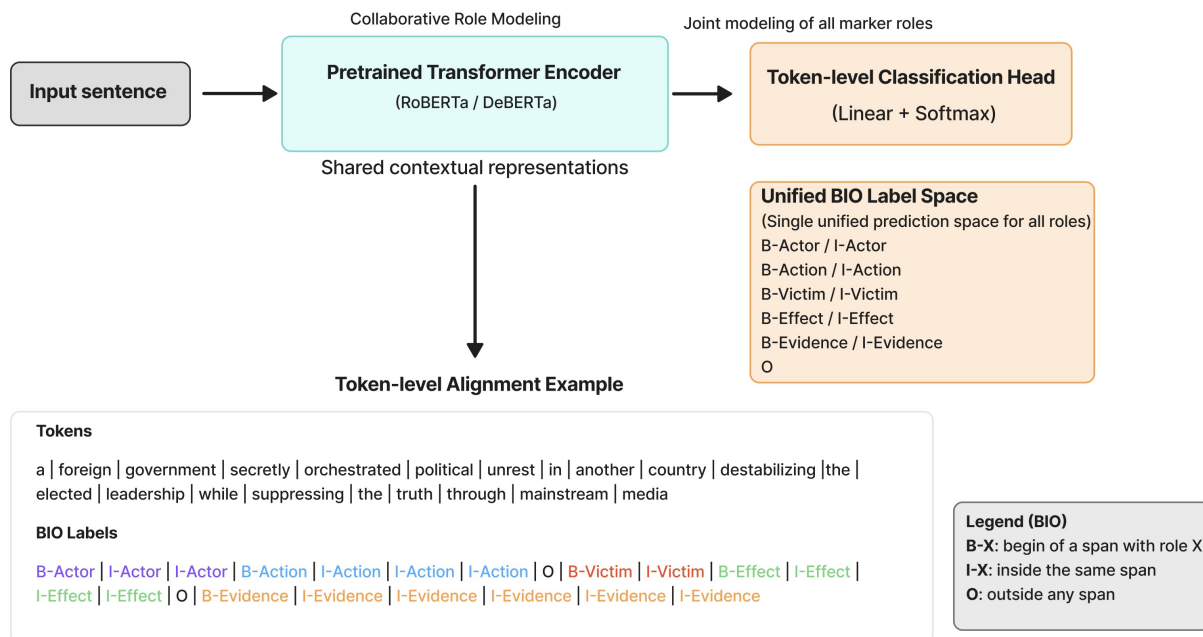


Figure 1: Unified Multi-Task Sequence Labeling Framework for Conspiracy Marker Extraction. A single shared encoder jointly predicts all roles within a unified BIO label space.

solution for fine-grained conspiracy marker extraction, balancing representational sharing with precise role differentiation.

3.1.2 Data Augmentation Strategy

To mitigate data sparsity while preserving span alignment, we adopt LLM-based data augmentation using Qwen. The model is prompted to generate semantically equivalent paraphrases that preserve the original conspiratorial structure while introducing lexical and syntactic variation.

We construct an augmented dataset approximately twice the size of the original training set by generating one paraphrased instance for each original sample. Because token-level supervision is highly sensitive to local lexical changes, this moderate augmentation setting increases training diversity without disrupting span boundaries or role alignment. The detailed prompt and augmentation examples are provided in Appendix A.

3.1.3 Ensemble Inference

We adopt pretrained transformer encoders, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021), as the backbone for sequence labeling.

Our ensemble strategy prioritizes diversity across model instances rather than architectural complexity. Specifically, we construct a five-model ensemble that integrates heterogeneous encoders

together with models trained both with and without data augmentation. This heterogeneous setup introduces complementary error patterns and improves boundary stability and cross-role consistency. Empirically, the proposed ensemble consistently outperforms individual models, indicating that training diversity is particularly beneficial for fine-grained marker extraction.

3.2 Subtask 2: Conspiracy Detection

Conspiracy Detection focuses on determining whether a comment constitutes conspiratorial discourse. Unlike token-level extraction, this task requires modeling global semantic coherence, pragmatic framing, and communicative intent. We therefore emphasize data-centric optimization strategies over architectural complexity.

3.2.1 Classification Architecture

For Conspiracy Detection, we adopt a standard transformer-based sentence classification framework built on pretrained encoders such as BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2020), followed by a lightweight classification head. As the task involves sentence-level binary prediction, strong pretrained language representations are sufficient to capture global semantic coherence and pragmatic framing.

3.2.2 Data Augmentation Strategy

Data augmentation constitutes the primary source of performance improvement for Conspiracy Detection. Because sentence-level classification is less sensitive to local lexical perturbations, we generate structurally diverse paraphrases that reformulate rhetorical framing, sentence organization, and discourse presentation while preserving the underlying conspiratorial intent. Using generation powered by the Qwen large language model, we construct an augmented dataset approximately three times the size of the original training set. The resulting diversity in expression, tone, and framing improves the classifier’s robustness to surface variation and implicit rhetorical cues. The corresponding prompt and examples are detailed in Appendix A.

3.2.3 Ensemble Inference

We further examine ensemble inference by aggregating predictions from multiple independently trained classifiers. However, empirical results indicate that ensembling primarily stabilizes model outputs rather than yielding performance improvements beyond the strongest individual model. While aggregation slightly reduces variance across runs, it does not surpass the best single-model performance.

This observation suggests that individual classifiers already capture highly similar decision boundaries after data augmentation, limiting the potential gains from prediction aggregation. Consequently, ensemble inference is not adopted as the main optimization strategy for Conspiracy Detection.

4 Experiments and Results

4.1 Experimental Setup

We evaluate our system on the official development and test sets provided by SemEval-2026 Task 10. All experiments are conducted using pretrained transformer encoders fine-tuned for the respective subtasks. Performance is assessed using the official evaluation metrics of the shared task: macro-averaged F1 score for Conspiracy Detection and span-level F1 score for Conspiracy Marker Extraction.

4.2 Results on Subtask 1

Table 1 presents the overall results for Conspiracy Marker Extraction on the development and test sets. Under the proposed multi-task sequence labeling

Single-Task Learning (ST)		
Model	Dev F1	Test F1
BERT	0.18	–
DeBERTa	0.21	–
Multi-Task Learning (MTL)		
Model	Dev F1	Test F1
RoBERTa	0.20	0.18
RoBERTa (DA)	0.21	0.19
BERT (DA)	0.20	0.18
DeBERTa	0.22	0.20
DeBERTa (DA)	0.23	0.20
Ensemble (5 models)	0.27	0.25

Table 1: Results on Subtask 1 (Conspiracy Marker Extraction).

Model	Data Augmentation	Dev F1
DeBERTa	×	0.22
DeBERTa	✓	0.23
BERT	×	0.18
BERT	✓	0.20
RoBERTa	×	0.20
RoBERTa	✓	0.21

Table 2: Effect of data augmentation on single-model performance for Subtask 1 (development set).

framework, all pretrained encoders achieve competitive performance, confirming the effectiveness of joint role modeling within a shared BIO tagging space.

For comparison, we report Single-Task (ST) results, where models are trained solely on the marker extraction objective without joint role modeling. ST models show slightly lower performance than their MTL counterparts (e.g., BERT: 0.18 vs 0.20 F1; DeBERTa: 0.21 vs 0.23 F1), confirming the benefit of multi-task learning.

Among individual models, DeBERTa consistently outperforms BERT and RoBERTa, indicating that stronger contextual encoders are advantageous for fine-grained span prediction, establishing DeBERTa as the strongest single-model baseline.

Data augmentation further improves performance across all model families. As shown in Table 2, applying LLM-based paraphrasing yields consistent gains of 1–2 F1 points on the development set. Although modest in magnitude, these improvements are systematic and stable, suggesting that augmentation enhances lexical robustness and mitigates overfitting under sparse supervision.

Despite these gains, single-model performance remains bounded. On the test set (Table 1), indi-

vidual models achieve F1 scores between 0.18 and 0.20, reflecting the difficulty of generalizing structured span predictions from limited annotated data. Notably, even the strongest DeBERTa model does not surpass 0.20 F1.

Ensemble inference produces substantially larger improvements. Our final ensemble integrates five heterogeneous models, including both augmented and non-augmented variants of DeBERTa and RoBERTa, together with an augmented BERT model. This configuration explicitly leverages diversity introduced by both architectural variation and data augmentation. The resulting ensemble achieves 0.27 F1 on the development set and 0.25 F1 on the official test set, clearly outperforming all individual models.

These findings indicate that while data augmentation provides consistent single-model improvements, ensemble diversity constitutes the primary driver of performance gains. The Unified Multi-Task Sequence Labeling Framework offers a strong structural foundation, but it is the heterogeneous ensemble—rather than any single encoder—that delivers the most substantial improvements in marker extraction.

4.3 Results on Subtask 2

Table 3 presents the results for Conspiracy Detection. Compared to Conspiracy Marker Extraction, this task exhibits substantially stronger baseline performance, with individual pretrained classifiers already achieving high Macro F1 scores.

Among single models, BERT attains the strongest performance, while DistilBERT remains competitive. Applying LLM-based data augmentation consistently improves both models, indicating that lexical diversity enhances robustness in sentence-level conspiracy detection. Although the absolute gains are modest, the improvements are stable across configurations.

In contrast to Subtask 1, ensemble inference does not yield additional gains. Combining DistilBERT and BERT leads to a decrease in performance (0.70), and aggregating DA and non-DA variants of both models achieves 0.72, which remains below the best single BERT model. This suggests that individual sentence representations are already sufficiently expressive, and additional model-level diversity contributes limited complementary information.

Overall, these findings suggest that Conspiracy Detection primarily benefits from strong pretrained

Model	Dev F1	Test F1
DistilBERT	0.76	0.70
DistilBERT(DA)	0.79	0.71
BERT	0.78	0.72
BERT(DA)	0.80	0.73
Ensemble (DistilBERT + BERT)	0.78	0.70
Ensemble (DA + non-DA)	0.80	0.72

Table 3: Results on Subtask 2 (Conspiracy Detection).

Model	Dev F1	Δ
BERT	0.78	–
BERT (+ Marker Features)	0.72	-0.06
RoBERTa	0.74	–
RoBERTa (+ Marker Features)	0.66	-0.08

Table 4: Cross-task transfer experiment (marker-informed detection)

representations and lexical diversity introduced through data augmentation. Increasing architectural complexity or ensemble diversity yields limited returns once a high-performance baseline is established.

4.4 Cross-Task Transfer Experiment

To explore whether fine-grained conspiracy markers can provide additional signals for sentence-level conspiracy detection, we conducted a cross-task transfer experiment. Specifically, outputs from the Conspiracy Marker Extraction model were transformed into structured marker indicators and appended to the input representation of sentence-level classifiers. However, results show that this strategy does not improve performance. As summarized in Table 4, incorporating marker-derived features consistently leads to performance degradation across models. BERT drops from 0.78 to 0.72 Macro F1, while RoBERTa decreases from 0.74 to 0.66. These findings suggest that fine-grained span predictions do not translate into reliable sentence-level signals under limited supervision. Error propagation from marker extraction and boundary ambiguity may introduce noise that outweighs potential structural benefits.

4.5 Analysis and Discussion

Beyond quantitative comparisons, our experiments reveal distinct optimization characteristics across the two subtasks.

For Conspiracy Marker Extraction, performance is primarily driven by model diversity. Conspiracy Marker Extraction under sparse supervision makes individual models sensitive to boundary am-

biguity. Heterogeneous encoders and augmentation regimes provide complementary predictions, making ensemble inference the most effective strategy for improving robustness.

In contrast, Conspiracy Detection is largely governed by representation strength rather than structural diversity. Even larger encoders such as DeBERTa and RoBERTa did not consistently outperform BERT or DistilBERT, suggesting that model capacity alone is not the limiting factor. Once strong sentence-level representations are established, additional architectural complexity yields diminishing returns.

Taken together, these findings suggest that optimization strategies should be tailored to task granularity. Structured prediction under sparse supervision benefits from ensemble diversity, whereas sentence-level classification is primarily influenced by representation quality and training data diversity.

5 Conclusion

This paper presents our system for SemEval-2026 Task 10, addressing both Conspiracy Marker Extraction and Conspiracy Detection. For marker extraction, we adopt a Unified Multi-Task Sequence Labeling Framework with a shared encoder and a single BIO tagging space. For conspiracy detection, we employ pretrained sentence classifiers enhanced by LLM-based data augmentation.

Our experiments demonstrate that the two sub-tasks exhibit fundamentally different optimization behaviors. For structured span prediction under sparse supervision, ensemble diversity plays a critical role in achieving robust performance. In contrast, sentence-level classification primarily benefits from strong pretrained representations and lexical diversity introduced through data augmentation, while additional architectural complexity and cross-task transfer from marker extraction provides limited gains.

These findings highlight the importance of tailoring optimization strategies to task granularity. Rather than uniformly increasing model complexity, performance improvements depend on whether diversity or representation strength constitutes the primary bottleneck. Future work may further explore adaptive strategies that dynamically balance structural modeling and data-centric optimization across heterogeneous tasks.

Acknowledgment

This work was supported by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No.25YJC740005), the National Language and Character Research Base (Project No.ZDI145-168), and Fundamental Research Funds for the Central Universities, Academy of Frontier Interdisciplinary Research, Central China Normal University (Project No.JC2026PT-004).

References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Karen Douglas, Robbie Sutton, and Aleksandra Cichocka. 2017. [The psychology of conspiracy theories](#). *Current Directions in Psychological Science*, 26.
- Karen M. Douglas and Robbie M. Sutton. 2023. [What are conspiracy theories? a definitional approach to their correlates, consequences, and communication](#). *Annual Review of Psychology*, 74:271–298.
- Alex Fong, Jon Roozenbeek, Daniel Goldwert, Steve Rathje, and Sander van der Linden. 2021. [The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on twitter](#). *Group Processes & Intergroup Relations*, 24(4):606–623.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [Deberta: Decoding-enhanced bert with disentangled attention](#). *Preprint*, arXiv:2006.03654.
- 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.
- 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)*.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. [Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter](#). *Preprint*, arXiv:1910.01108.

A Data Augmentation Examples

Subtask 1: Conspiracy Marker Extraction

[instruction]

You are an expert linguist. Rewrite the sentence to increase lexical diversity while strictly preserving the semantic structure and the specific conspiratorial claims. Do not change the core phrases or the order of arguments. Use synonyms and minor syntactic variations only. Keep the sentence length similar.

[Input]

Russia has better nuclear capabilities yet criminal politicians Trump and Netanyahu risk bringing a global nuclear catastrophe. Maybe we need to get real about exposing corruption BEFORE WE ALL DIE.

[Output]

Russia possesses stronger nuclear capabilities, yet corrupt politicians Trump and Netanyahu risk triggering a worldwide nuclear disaster. Perhaps we must seriously confront corruption before we all perish.

Role Preservation Analysis

[Actor]

Original: Trump and Netanyahu

Augmented: Trump and Netanyahu

[Action]

Original: exposing corruption

Augmented: confront corruption

[Victim]

Original: we / WE

Augmented: we

[Effect]

Original: global nuclear catastrophe

Augmented: worldwide nuclear disaster

[Evidence]

Original: None

Augmented: None

Subtask 2: Conspiracy Detection

[instruction]

You are a linguistic expert tasked with generating structurally diverse training data for conspiracy detection. Rewrite the text by significantly altering its tone, sentence organization, and discourse presentation while ensuring the underlying conspiratorial intent and belief remain identical. You should reformulate the message into a different framing, such as a rhetorical question or a skeptical inquiry. Ensure the paraphrased version is syntactically distinct from the original but semantically equivalent for sentence-level classification.

Conspiracy: Yes

[Input]

Russia has better nuclear capabilities yet criminal politicians Trump and Netanyahu risk bringing a global nuclear catastrophe. Maybe we need to get real about exposing corruption BEFORE WE ALL DIE.

[Output]

If Russia truly holds superior nuclear capabilities, why are criminal politicians like Trump and Netanyahu still pushing the world toward a global nuclear catastrophe? Perhaps the real urgency lies in exposing corruption before we all die.

Conspiracy: No

[Input]

A Colombian Instagram model moonlighted as an international underage sex trafficking madame who recruited vulnerable girls to service celebrity clients, according to authorities.

[Output]

Authorities report a case in which a Colombian Instagram model allegedly operated as an international underage sex trafficking madame, recruiting vulnerable girls for celebrity clients.