# JIA at SemEval-2026 Task 10: A Dual-Track System with BERT-based Encoders and LLMs for Conspiracy Analysis

**Jiayue Zhu**
University of Tübingen
Department of Linguistics
jiayue.zhu@uni-tuebingen.de

## Abstract

The proliferation of conspiracy theories on social media platforms like Reddit poses a severe challenge to public discourse, and the accurate identification of these causal narratives relies on the fine-grained capture of their intricate narrative structures. This study proposes a comprehensive monitoring framework that integrates BERT-based discriminative models with Large Language Models (LLMs) to jointly optimize conspiracy detection and psycholinguistic marker extraction. We systematically evaluated the performance of several architectures, including DistilBERT, BERT-Base, DeBERTa-V3, RoBERTa, and the instruction-tuned Qwen2.5 series. Experimental results demonstrate that Qwen2.5-14B (Full-shot) achieves a peak Weighted F1-score of 0.80 in the detection task. Furthermore, marker extraction remains a significant challenge; although the fine-tuned LLM leads in identifying "Actors" (F1 of 0.144), its performance in semantically ambiguous categories such as "Evidence" and "Effect" highlights a pronounced generalization gap.

## 1 Introduction

Reddit stands as one of the world's most active communities with 116 million daily active users (Reddit, Inc., 2025). Due to its unique 'subreddit' architecture, it has become a vital arena for public discourse. However, this interest-based division, combined with users' homophily preferences, easily leads to closed "echo chamber" effects. This, in turn, provides a hotbed for conspiracy theories to thrive and spread rapidly (Papacharissi, 2016). Conspiracy theories are defined as causal narratives that attribute events to covert plans orchestrated by a secret cabal (Banas and Miller, 2013). Such narratives not only distort facts but also pose substantial threats to public health and democratic institutions (Diab et al., 2024). Consequently, accurately identifying conspiracy theories within vast volumes of daily discourse and disrupting their propagation is critical.

Conspiracy theories are not mere rumors but intricate narrative structures comprising six key elements: events, actors, goals, actions, consequences, and targets (Introne et al., 2020). These implicit and complex semantic characteristics make it difficult for traditional automated methods (e.g., proxy-based methods (Bessi et al., 2015), keyword matching (Kim and Kim, 2023), and pattern matching (Kou et al., 2017)) to capture the contextual logic. The development of deep learning methods, particularly the application of Pre-trained Language Models (PLMs), has significantly enhanced the generalized understanding of textual content (Patwardhan et al., 2023). Large Language Models (LLMs) have provided a new paradigm for decoding the narrative logic of conspiracy theories. Beyond merely identifying latent biases, these models leverage their reasoning capabilities to conduct targeted analyses of complex conspiracy discourses (Costello et al., 2024).

While past research has achieved notable results in applying pre-trained models to identify various conspiracy theory texts, deconstructing their internal formation mechanisms remains crucial for the interpretation of conspiracy theories. In light of this, SemEval-2026 Task 10: Psycholinguistic Conspiracy Marker Extraction and Detection aims to integrate psychological theory with NLP techniques (Samory et al., 2026). By extracting conspiracy markers (Subtask 1) and detecting conspiracy comments (Subtask 2), the task seeks to reveal the expressive mechanisms of conspiracy theories.

To address these challenges, this study proposes a general conspiracy theory monitoring framework for Reddit comments. The framework incorporates both BERT-based PLMs and LLMs to analyze the semantic characteristics and narrative logic of comments, enabling the accurate detection of conspiracy content and the automated extraction of crucial



*Proceedings of the The 20th International Workshop on Semantic Evaluation (2026)*, pages 1332–1339

psycholinguistic markers. Furthermore, we explore how model architecture, parameter scale, and pre-training data volume affect the overall performance of conspiracy analysis.

## 2 Related Work

Conspiracy theories are causal narratives that interpret major social or political events as the outcomes of secret plots orchestrated by powerful actors for specific ends (Douglas and Sutton, 2018; Banas and Miller, 2013). This distinct causal logic makes conspiracy theories predominantly certainty-oriented, prioritizing closed-loop causal explanations aimed at cognitive closure over open-ended factual inquiry (Fong et al., 2021).

In the early evolution of detection technologies, researchers primarily relied on external features or predefined rules as criteria. On one hand, proxy-based methods identified conspiracy content indirectly by analyzing information sources, user behaviors, or dissemination pathways, such as specific media domains or URLs (Bessi et al., 2015; Starbird et al., 2019). On the other hand, keyword-matching and syntactic-pattern methods located target texts by querying predefined lexicons (Kim and Kim, 2023) or specific narrative triplets, such as "Subject-Action-Object" (Kou et al., 2017; Samory and Mitra, 2018). While these approaches offer high interpretability, they struggle to capture the underlying narrative logic within the semantically dense contexts of social media.

To overcome the limitations of traditional paradigms regarding generalization and semantic comprehension, subsequent research turned to deep learning to automatically extract high-dimensional features from the text itself. Early efforts—such as the bidirectional GRU networks within the BORJIS framework—surpassed traditional methods in detection accuracy (Galende et al., 2022). Transformer-based architectures like BERT and RoBERTa utilize dynamic embedding generation to significantly bolster the model's ability to capture deep semantic and logical structures (Patwardhan et al., 2023). Building upon these advancements, the emergence of LLMs has introduced a new paradigm for parsing the complex narrative logic of conspiracy theories. Fine-tuning LLMs enables the identification and analysis of misinformation, propaganda narratives, and manipulative tactics within news media (Pavlyshenko, 2023). Furthermore, to address classification challenges in complex environments, LLMs have demonstrated remarkable potential in zero-shot scenarios for identifying fine-grained (Peskine et al., 2023).

## 3 Data

### 3.1 Dataset

This study employs the SemEval PsyCoMark dataset, which is sourced from more than 190 subreddits on the Reddit platform (Samory et al., 2025). The dataset consists of 4,316 labeled training instances, along with 100 development instances and 337 test instances, both of which are unlabeled. Regarding the data structure, each sample includes not only the core text, the conspiracy-theory classification label, and five categories of psycholinguistic markers (Actor, Action, Effect, Evidence, and Victim), but also retains key metadata including the sample ID, the originating subreddit, and the annotator information.

### 3.2 Data Preprocessing and Distribution Analysis

To ensure data quality, this study first cleaned the raw data: (1) **Deduplication:** Removed redundant samples with identical attributes , including 'ID', 'text', 'markers', 'subreddit', and 'annotator'. (645 entries, 14.94%); (2) **Conflict Resolution:** Removed noisy samples with identical text but inconsistent conspiracy labels or markers (434 entries, 10.06%).

The distribution of the cleaned data is shown in Table 1. The 'Non-conspiracy' category is dominant (1,541 samples), followed by 'Conspiracy' (1,108) and 'Unknown' (588). Additionally, text-length analysis (see Appendix A) reveals a distribution clustering between 30–60 words.

Further analysis highlights significant psycholinguistic distinctiveness: conspiracy texts show a 92% presence of 'Actor' and 'Action' markers, with substantially higher proportions of 'Effect', 'Evidence', and 'Victim' markers compared to non-conspiracy texts (e.g., 'Victim' is only 38% in the latter). This indicates that conspiracy theories rely on constructing complex causal chains with clear subjects and consequences.

## 4 Methodology

This study proposes a conspiracy analysis framework, as illustrated in Figure 1, using BERT-based PLMs and LLMs, designed to identify conspiracy

Table 1: Statistics of conspiracy labels and markers.

| Conspiracy Label | Total Texts | with Actor | with Action | with Effect | with Evidence | with Victim |
|---|---|---|---|---|---|---|
| Can't tell | 588 | 424 (72%) | 407 (69%) | 288 (49%) | 331 (56%) | 252 (43%) |
| No | 1,541 | 963 (62%) | 955 (62%) | 791 (51%) | 726 (47%) | 582 (38%) |
| Yes | 1,108 | 1,015 (92%) | 1,019 (92%) | 857 (77%) | 824 (74%) | 755 (68%) |
| Total | 3,237 | 2,402 (74%) | 2,381 (74%) | 1,936 (60%) | 1,881 (58%) | 1,589 (49%) |

content in unstructured social media text and further extract critical psychological markers.

### 4.1 BERT-based Discriminative Architecture

This methodology integrates conspiracy detection and marker extraction tasks, extracting key features from social media contexts through a unified encoder framework. Conspiracy detection is modeled as a binary classification task $f(T, M) \rightarrow y$. Simultaneously, the marker extraction task performs token-level labeling using the BIO scheme for 'Actor', 'Action', 'Effect', 'Evidence', and 'Victim'.

This study selected four representative BERT-based models to evaluate their performance from multiple dimensions (Qiu et al., 2020). BERT-Base (Devlin et al., 2019) provides a performance benchmark, DistilBERT (Sanh et al., 2019) validates lightweight efficiency, RoBERTa-Large (Liu et al., 2019) enhances semantic capture by increasing model capacity, and DeBERTa-V3 (He et al., 2021) optimizes feature extraction using a disentangled attention mechanism. These models share the encoding layer, providing the global marker vector and token-level contextual vectors for downstream tasks.

In feature processing for the detection task, discrete metadata are mapped to low-dimensional dense vectors to enhance representation and prevent overfitting. Specifically, subreddit information is mapped to a 32-dimensional embedding vector, while source ID and annotator are each mapped to 8-dimensional embedding vectors. These metadata features are concatenated with the text representation after being processed by embedding layers. The fused vector is normalized via a LayerNorm layer and fed into a two-layer MLP to output classification probabilities. Concurrently, each token vector undergoes entity boundary determination through a linear projection layer and a softmax function.

### 4.2 LLM-based Generative Architecture

This methodology employs Qwen2.5 (7B/14B)-Instruct as the decoder, utilizing structured prompt engineering to transform the analysis into a controlled text-generation process. As detailed in Appendix B, the prompt design consists of three core components: task description, instruction, and output format. By integrating an expert persona, core definitions of conspiracy theories, and granular criteria for five psycholinguistic markers, the framework explicitly instructs the model to exclude ambiguous options in binary classification. This forces the model to generate deterministic judgments based on rigorous logical deduction.

Our training and inference framework encompasses a three-tier strategy: Zero-shot prompts to assess inherent capabilities, Few-shot learning with five-example guidance for pattern recognition, and Full-set fine-tuning to maximize the model's performance. To achieve deeper knowledge alignment, we utilize Parameter-Efficient Fine-Tuning (PEFT) for instruction tuning. During the training phase, a minimal set of trainable parameters is injected specifically into the attention projection layers (Ding et al., 2023).

To ensure the rigor of the generated outputs, the system mandates a JSON-structured response containing the reasoning logic, the conspiracy label, and the list of identified markers. The generated marker text is then precisely back-mapped to the original text indices. This process guarantees token-level index consistency for the psycholinguistic extraction task while maintaining strict adherence to the required submission format.

### 4.3 Experimental Setup

This research employs a dual-track parallel strategy: Track A constructs discriminative models based on the BERT family for multi-task learning, while Track B uses Qwen-based LLMs for generative inference. For the BERT track, the training data was partitioned using a 9:1 hold-out split, with validation performance monitored in real-time to identify

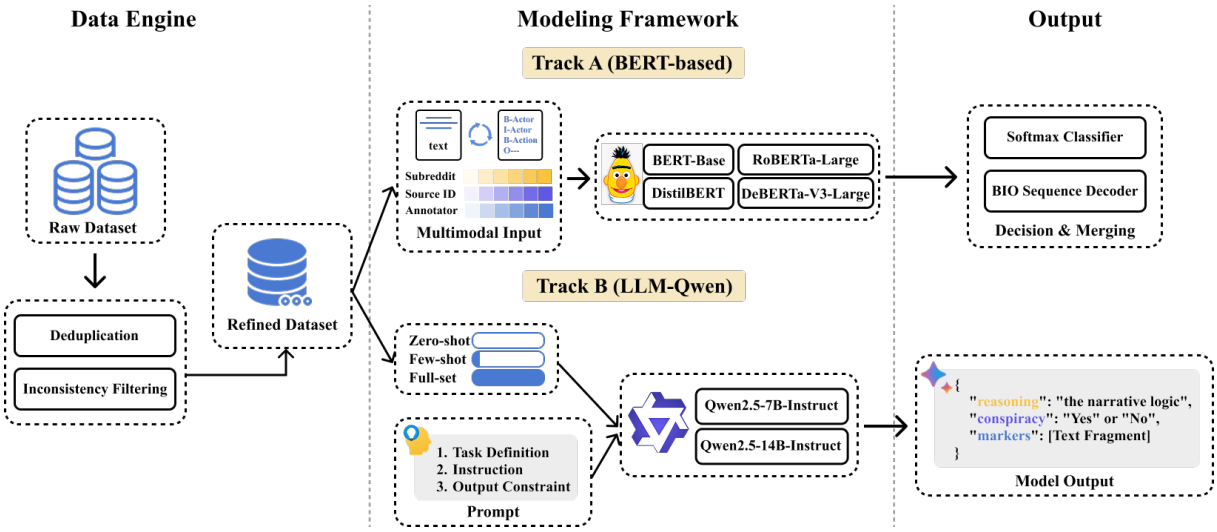


Figure 1: Overview of Methodology Framework.

the optimal epoch and maximize discriminative accuracy. In the LLM track, we implemented instruction fine-tuning via the LoRA algorithm to strengthen the model's grasp of complex conspiracy narrative logic and its structured representation capabilities. Detailed hyperparameter configurations are provided in Appendix C.

### 4.4 Evaluation Metrics

The evaluation framework follows the official competition guidelines (Samory et al., 2025). For Conspiracy Detection, which is framed as a binary classification task, performance is measured using Accuracy and Weighted F1-score. For Conspiracy Marker Extraction, the system adopts a token-based Intersection over Union (IoU) strategy to handle variations in span boundaries. A predicted segment is classified as a True Positive (TP) only if its entity type matches the ground truth and its token-level IoU score reaches a threshold of 0.5. Based on this criterion, the model is evaluated using both Micro F1 and Macro F1-scores. Detailed mathematical definitions for all metrics are provided in Appendix D.

## 5 Results and Analysis

### 5.1 Conspiracy Detection

As illustrated in Table 2, the experimental results reveal distinct performance between discriminative and generative architectures. In the discriminative track, DeBERTa-V3-Large distinguishes itself through exceptional stability, achieving both accuracy and Weighted F1-scores of 0.76. Compared to its performance on the development set (detailed in Appendix E), DeBERTa displays high consistency in Weighted F1-scores across both test and validation sets (both approximately 0.76), effectively proving the superior robustness of the architecture. Conversely, while DistilBERT achieved a high score of 0.80 on the development set, its Weighted F1-score dropped significantly to 0.72 on the test set, exposing a clear bottleneck in generalization performance.

Table 2: Results for Conspiracy Detection.

| Model | Acc. | W-F1 | F1 (No) | F1 (Yes) |
|---|---|---|---|---|
| DistilBERT | 0.72 | 0.72 | 0.75 | 0.67 |
| BERT-Base | 0.70 | 0.71 | 0.75 | 0.66 |
| RoBERTa-Large | 0.74 | 0.74 | 0.76 | 0.71 |
| **DeBERTa-V3-Large** | **0.76** | **0.76** | **0.80** | **0.71** |
| Qwen2.5-7B (Zero-shot) | 0.66 | 0.59 | 0.78 | 0.31 |
| Qwen2.5-7B (Few-shot) | 0.66 | 0.59 | 0.77 | 0.32 |
| Qwen2.5-7B (Full-shot) | 0.79 | 0.79 | 0.83 | 0.73 |
| Qwen2.5-14B (Zero-shot) | 0.73 | 0.72 | 0.79 | 0.62 |
| Qwen2.5-14B (Few-shot) | 0.77 | 0.76 | 0.82 | 0.68 |
| **Qwen2.5-14B (Full-shot)** | **0.80** | **0.80** | **0.84** | **0.76** |

In the generative track, the instruction-tuned Qwen2.5-14B (Full-shot) demonstrates overwhelming performance dominance, reaching a Weighted F1-score of 0.80. This result not only substantially outperforms the zero-shot and few-shot configurations of the same model but also exceeds the peak performance of discriminative models on the validation set. These findings provide strong evidence that applying domain-specific instruction fine-tuning to large-scale foundation models is a pivotal strategy for enhancing the detection of complex conspiracy narratives.

### 5.2 Conspiracy Marker Extraction

As illustrated in Figure 2, Qwen-7B (Full-set) delivers the most robust overall performance, securing the lead in both F1 Aggregate (0.094) and Actor (0.144). Among the baselines, RoBERTa-Large shows a targeted advantage in the Victim (0.085) metric. The results of large language models in zero-shot and few-shot scenarios—where most F1 scores remain below 0.041—further confirm that full-domain fine-tuning is indispensable for fine-grained semantic extraction in conspiracy detection. Furthermore, compared to the superior development set performance detailed in Appendix E (e.g., where the Actor score reached approximately 0.38), the drop in metrics on the test set across complex dimensions like Evidence and Effect highlights the generalization difficulties models encounter when dealing with long-tail semantics and ambiguous boundaries.

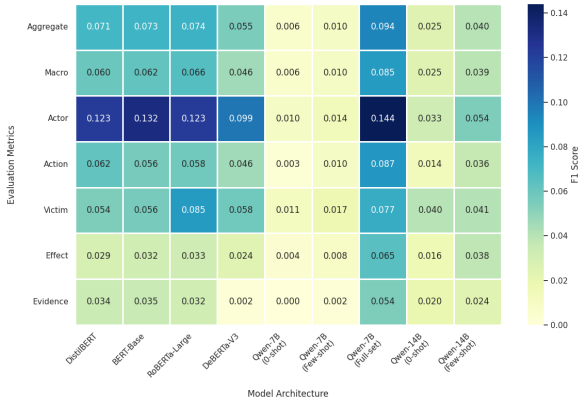


Figure 2: Performance Comparison of Marker Extraction (F1 Score).

## 6 Conclusion

Results demonstrate that Qwen2.5-14B, following full-domain instruction fine-tuning, achieves exceptional performance in conspiracy detection, accurately deconstructing the intricate causal logic and narrative patterns inherent in Reddit discourse. While discriminative BERT-based models, such as DeBERTa-V3-Large, exhibit robustness and consistency across datasets, they fall short of large-scale generative models in terms of deep reasoning capabilities. In marker extraction task, the models perform better in identifying "Actors" and "Victims," further validating that entities with well-defined semantic boundaries are more readily captured and represented by the models.

Despite these performance gains, several limitations remain. Foremost is a notable generalization gap between development and test environments, particularly within marker extraction, where scores for "Evidence" and "Effect" suffered a substantial decline. This suggests that current architectures struggle to navigate the long-tail semantics and ambiguous boundaries characteristic of abstract narrative elements. Additionally, the heavy reliance on full-domain fine-tuning exposes a deficiency in the zero-shot and few-shot capabilities of LLMs for this specific task. Performance gains remain difficult to achieve without extensive labeled data.

To address these constraints, future research will explore external knowledge and data augmentation strategies. We intend to incorporate external knowledge bases specifically related to conspiracy theories to enhance the models' ability to structurally represent causal chains and evidentiary logic. Concurrently, we plan to utilize generative techniques to construct diverse synthetic samples, expanding the training set to bolster model robustness in low-resource scenarios.

## References

John A Banas and Gregory Miller. 2013. Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Human communication research*, 39(2):184–207.

Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, 10(2):e0118093.

Thomas H Costello, Gordon Pennycook, and David G Rand. 2024. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814.

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.

Ahmad Diab, Rr Nefriana, and Yu-Ru Lin. 2024. Classifying conspiratorial narratives at scale: False alarms and erroneous connections. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 340–353.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, and 1 others. 2023.

Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5(3):220–235.

Karen M Douglas and Robbie M Sutton. 2018. Why conspiracy theories matter: A social psychological analysis. *European Review of Social Psychology*, 29(1):256–298.

Amos Fong, Jon Roozenbeek, Danielle Goldwert, Steven 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.

Borja Arroyo Galende, Gustavo Hernández-Peñaloza, Silvia Uribe, and Federico Álvarez García. 2022. Conspiracy or not? a deep learning approach to spot it on Twitter. *IEEE Access*, 10:38370–38378.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Joshua Introne, Ania Korsunska, Leni Krsova, and Zefeng Zhang. 2020. Mapping the narrative ecosystem of conspiracy theories in online anti-vaccination discussions. In *International Conference on social media and society*, pages 184–192.

Soojong Kim and Jisu Kim. 2023. The information ecosystem of conspiracy theory: Examining the QAnon narrative on Facebook. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–24.

Yubo Kou, Xinning Gui, Yunan Chen, and Kathleen Pine. 2017. Conspiracy talk on social media: collective sensemaking during a public health crisis. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–21.

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. *arXiv preprint arXiv:1907.11692*.

Zizi Papacharissi. 2016. Affective publics and structures of storytelling: Sentiment, events and mediality. *Information, communication & society*, 19(3):307–324.

Narendra Patwardhan, Stefano Marrone, and Carlo Sansone. 2023. Transformers in the real world: A survey on NLP applications. *Information*, 14(4):242.

Bohdan M Pavlyshenko. 2023. Analysis of disinformation and fake news detection using fine-tuned large language model. *arXiv preprint arXiv:2309.04704*.

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding gpt for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4054–4063.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China technological sciences*, 63(10):1872–1897.

Reddit, Inc. 2025. Reddit Announces Third Quarter 2025 Results. https://investor.redditinc.com/news-events/news-releases/news-details/2025/Reddit-Announces-Third-Quarter-2025-Results/default.aspx. Accessed: 2025-12-15.

M. Samory, F. Soldner, and V. Batzdorfer. 2025. PsyCoMark - Psycholinguistic Conspiracy Marker Dataset (0.0.2). Data set.

Mattia Samory and Tanushree Mitra. 2018. 'The Government Spies Using Our Webcams' the language of conspiracy theories in online discussions. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–24.

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. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–26.

## A Distribution of Text Lengths

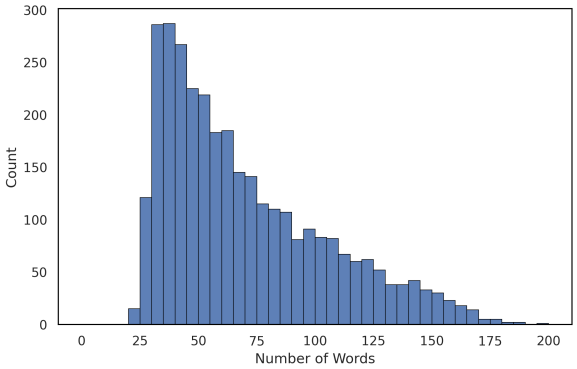


Figure 3: Distribution of text lengths (word count).

## B Prompt Engineering

In this section, we detail the structural design of our prompt engineering for Large Language Models (LLMs).

**Task Definition**

**You are an expert in psycholinguistics and conspiracy theory detection.**
**Your task is to analyze Reddit comments to:**
**1. Identify if they indicate a belief in a conspiracy theory (Binary Classification).**
**2. Extract psycholinguistic markers (Actor, Action, Victim, Effect, Evidence) supporting this (Span Extraction).**

**Instruction**

**1. Core Definition**
**A conspiracy theory is a causal narrative explaining a significant event as the secret result of powerful, deceptive actors working together for malevolent goals, rather than as a random occurrence.**
**2. Marker Definitions**
**You must extract the following narrative elements based on the logic above:**
**Actor: The deceptive, coordinated individuals or groups perceived as working together to initiate the plot (e.g., "The Deep State", "Global Elites").**
**Action: The secret, malevolent activities performed by the Actor to achieve their goal (e.g., "orchestrating the virus", "suppressing the truth").**
**Victim: The individual or population that is intentionally disenfranchised or harmed by the conspiracy.**
**Effect: The negative consequences or the specific goal the actors aim to achieve (e.g., "total control", "depopulation").**
**Evidence: Any arguments, documents, or data used to provide a closed-ended causal explanation or to express high certainty about the plot.**

**Output Format**

**You must respond with a strict JSON object. Identify all 5 marker types if present in the text:**

```
{
  "reasoning": "Brief analysis of the narrative logic. If ambiguous, explain why you chose the final label.",
  "conspiracy": "Yes" or "No",
  "markers": [
    {"text": "exact phrase from text", "type": "Actor"},
    {"text": "exact phrase from text", "type": "Action"},
    {"text": "exact phrase from text", "type": "Victim"},
    {"text": "exact phrase from text", "type": "Effect"},
    {"text": "exact phrase from text", "type": "Evidence"}
  ]
}
```

Figure 4: Overview of Prompt Engineering for Task Definition, Instruction, and Output Constraint.

## C Hyperparameter Configurations

Table 3: Hyperparameter Configurations for Track A (Bert-based).

| Hyperparameter | Conspiracy Detection | Conspiracy Marker Extraction |
|---|---|---|
| Optimizer | AdamW | AdamW |
| Loss Function | Cross-Entropy | Cross-Entropy |
| Dropout | 0.2 (Fusion), 0.1 (Hidden) | 0.1 |
| Learning Rate | $2 \times 10^{-5}$ | $2 \times 10^{-5}$ |
| Batch Size | 16 | 32 |
| Epochs | 5 | 10 |
| Max Sequence Length | 256 | 256 |

## D Mathematical Definitions of Evaluation Metrics

### 1. Weighted F1-score (Conspiracy Detection)

To address class imbalance, the Weighted F1-score sums the F1-scores of each category $c$ (Yes/No) weighted by their relative sample weights $w_c$:

$$\text{Weighted F1} = \sum_{c \in \{Yes, No\}} w_c \times \frac{2 \cdot P_c \cdot R_c}{P_c + R_c}$$

Table 4: Hyperparameter Configurations for Track B (LLMs-based).

| Hyperparameter | Value |
|---|---|
| LoRA Rank ($r$) | 64 |
| LoRA Alpha ($\alpha$) | 128 |
| LoRA Dropout | 0.1 |
| Optimizer | Paged AdamW (32-bit) |
| Learning Rate | $2 \times 10^{-4}$ |
| Batch Size | 1 |
| Gradient Accumulation | 16 |
| Max Sequence Length | 1024 |
| Training Epochs | 2 |
| Learning Rate Scheduler | Cosine |

### 2. IoU-based Matching (Conspiracy Marker Extraction)

The Intersection over Union (IoU) for a predicted span $S_{pred}$ and ground truth span $S_{gt}$ is calculated based on their token sets $T$:

$$\text{IoU}(S_{pred}, S_{gt}) = \frac{|T_{pred} \cap T_{gt}|}{|T_{pred} \cup T_{gt}|}$$

### 3. Micro and Macro F1-scores

For each entity type $i$, the F1-score is the harmonic mean of Precision ($P_i$) and Recall ($R_i$):

$$F1_i = \frac{2 \cdot P_i \cdot R_i}{P_i + R_i}$$

The Macro F1 is the unweighted arithmetic mean across all $n$ types:

$$\text{Macro F1} = \frac{1}{n} \sum_{i=1}^{n} F1_i$$

The Micro F1 is derived from the global sums of True Positives (TP), False Positives (FP), and False Negatives (FN):

$$\text{Micro F1} = \frac{2 \times \sum TP_i}{2 \times \sum TP_i + \sum FP_i + \sum FN_i}$$

## E Detailed Results on Development Set

The following table presents the detailed performance metrics for Bert-based models on the development set.

The following figure presents the performance (F1 score) of conspiracy marker extraction in development set.

Table 5: Performance comparison of BERT-based models on the conspiracy detection development set.

| Model | Accuracy | Weighted F1 | F1 (No) | F1 (Yes) |
|---|---|---|---|---|
| **DistilBERT** | **0.8052** | **0.8023** | **0.8544** | **0.7059** |
| BERT-Base | 0.7922 | 0.7879 | 0.8462 | 0.6800 |
| RoBERTa-Large | 0.7662 | 0.7680 | 0.8163 | 0.6786 |
| DeBERTa-V3 | 0.7662 | 0.7614 | 0.8269 | 0.6400 |

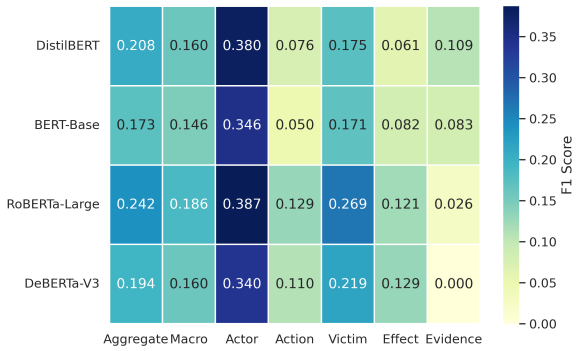


Figure 5: Performance Heatmap of Conspiracy Marker Extraction.