SemEval-2026 Task 10: PsyCoMark – Psycholinguistic Conspiracy Marker Extraction and Detection

Mattia Samory
Sapienza University of Rome
mattia.samory@uniroma1.it

Felix Soldner
Independent Researcher
felix.soldner@pm.me

Veronika Batzdorfer
Karlsruhe Institute of Technology
veronika.batzdorfer@kit.edu

Abstract

Despite the need to address the proliferation of conspiracy theories in online discussions, there is a lack of benchmarks for effectively detecting conspiracy-related content in everyday conversational settings. We introduce a novel dataset of comments from Reddit, ranging from politics to TV series, as well as two synergetic tasks: (1) extracting five psycholinguistic markers, grounded in evolutionary psychology, and (2) detecting conspiracy content. The data enable multi-task approaches, allowing testing of whether marker extraction improves detection performance.

1 Introduction

Most related NLP work aiming to detect conspiracy theories focuses on detecting specific claims, such as the belief that vaccines cause autism (Mitra et al., 2016). However, conspiracy theories may adapt to a variety of topical contexts, complicating their detection in realistic settings (Samory and Mitra, 2018). Yet, psychology research points to invariants in the structure of conspiracy theory narratives due to the functions that they serve in group life, which are independent of the specific claims of each theory (Prooijen and Vugt, 2018; Raab et al., 2013). We propose extracting the markers of conspiracy theory narratives and detecting conspiracy theories as two synergistic tasks. We introduce two novel aspects to the computational analysis of conspiracy theory language: psychologically grounded markers and topic-agnostic conversations.

Psychologically-Grounded Markers Evolutionary psychology literature posits that conspiracy beliefs evolved as adaptive mechanisms for detecting dangerous coalitions (Prooijen and Vugt, 2018). To shield oneself and one's in-group from threats, conspiracy thinking trains the individual to anticipate the secret scheming of malicious opponents. We introduce a novel dataset that maps the psychological mechanisms of conspiracy beliefs to linguistic markers in conspiracy theory discussions. Table 1 summarizes the relationship between psychological mechanisms and linguistic markers in our annotation schema. This grounding ensures that the annotations capture the underlying psychological drivers of conspiracy thinking, affording deeper insights and explainability.

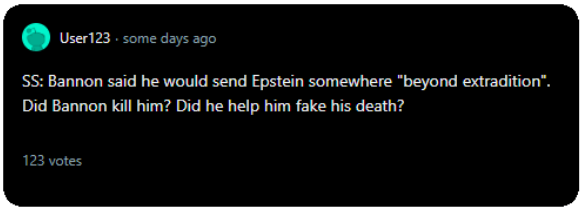

Figure 1: Example of a submission statement on Reddit (anonymized).

Topic-Agnostic Conversations Existing corpora focus on pre-determined conspiracy topics — in fact, there is criticism about the influence of topicality even in the psychological measurement of conspiracy thinking. Few resources, like the LOCO dataset (Miani et al., 2021), encompass a broad range of topics but do not include conspiracy theories as they are discussed on social media. Our dataset offers a contextually rich resource for analysis across a broad range of topics in everyday social media conversations—including subreddits both overtly about conspiracy theories (e.g., `conspiracy`, `ufobelievers`, `epstein`) and of general interest (e.g., `hongkong`, `truecrime`, `news`). The diversity ensures that models cannot rely on topic-specific shortcuts (e.g., associating "vaccines" with conspiracy theories) but need to learn markers that generalize across domains.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3628–3639

Psychological driver	Marker	Documents
Alliance detection	Actor	4012
Threat management	Effect	3240
	Victim	2603
Pattern perception	Evidence	3178
Agency detection	Action	3961

Table 1: Relationship between the psychological mechanisms underpinning conspiracy beliefs and the linguistic markers used in our annotation schema.

2 Task Structure

By combining detection and marker extraction, the PsyCoMark task enables the development of generalizable and interpretable models of conspiracy theory language.

Subtask 1: Conspiracy Marker Extraction Participants were tasked to develop models to extract the spans indicative of conspiracy theory markers. Each document may contain zero or multiple spans corresponding to each marker, which may overlap with each other.

Subtask 2: Conspiracy Detection Participants in this subtask have developed models to classify comments as either conspiracy-related or non-conspiracy-related.

Participants were encouraged but not required to contribute end-to-end systems that address both tasks jointly, and were allowed to leverage external data and pre-trained models to remove barriers to participation.

3 Data

Metric	Value
Documents	5354
Unique documents	4720
Median tokens per document	61
Annotators	138
Subreddits	226
Documents labeled "conspiracy"	1925
Documents labeled "not conspiracy"	2457
Documents labeled "can't tell"	972
Documents with any marker	4511
Total markers	27891

Table 2: Statistics of the annotated dataset.

Submission Statements To include texts that cover a wide variety of topics and include rich narrative information, we focused our data on submission statements on Reddit. Submission statements stem from the practice of accompanying submissions that contain only media—like a picture, video, or URL—with a comment describing why the media relates to a subreddit's topic. For example, the submission statement accompanying a newspaper article in the subreddit `conspiracy` may summarize how the article supports a conspiracy theory (see, e.g., Fig. 1), whereas in the subreddit `law` could connect news events to a broader legal discourse. Since submission statements encode users' summaries of the narrative in media, they represent an ideal format from which to extract linguistic markers of conspiracy theories. This may elicit higher-density marker usage than spontaneous comments, where conspiracy narratives may remain implicit. Submission statements are common practice in a wide range of subreddits, which encompass a broad topical range.

Listing 1: Example data point

```
{
    "_id": "t1_xxxyyyzzz",
    "conspiracy": "Yes",
    "markers": [{
            "type": "Actor",
            "text": "CNN",
            "startIndex": 11,
            "endIndex": 13
        }//, ...
    ],
    "subreddit": "conspiracy",
    "annotator": "annotator_1"
}
```

Sampling We selected first-level comments, authored by the discussion starter, that began with `ss|((submission )?statement)[^a-zA-Z\d]`, which is the standard format for submission statements. We manually excluded subreddits where submission statements were used for alternative purposes. We converted markdown to text and converted URLs to special tokens. We discarded comments containing quotes to limit the text by the comment's author. To balance content richness with the feasibility of annotation, we retained comments between 160 and 1000 characters. Since conspiracy theories are relatively rare across Reddit, we oversampled comments that likely contain them: we sampled 1500 comments uniformly at random from the subreddit `conspiracy` and the remaining comments from the remaining subreddits with at least 10 submission statements, sampling an equal number per subreddit iteratively.

Annotation We annotated the data through the crowdwork platform Prolific. Annotators (N=138)

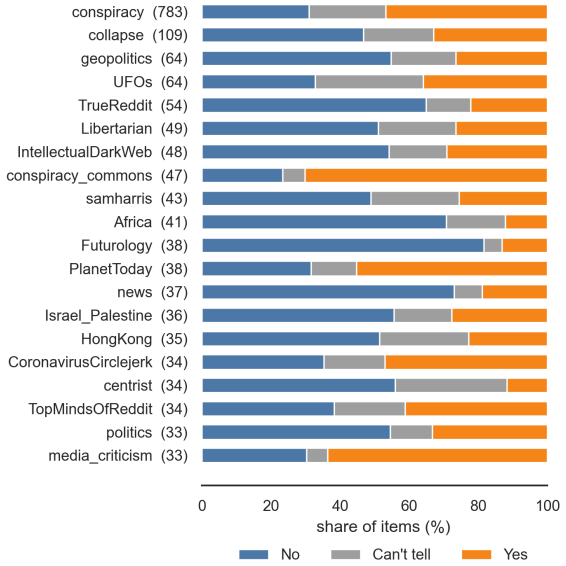

Figure 2: Conspiracy annotations for the top 20 subreddits in the dataset. Even self-professed conspiracy theory communities and those associated with them in the public discourse show a mixture of labels, e.g., the `conspiracy` and `UFO` subreddits show fewer than 50% "Yes" labels.

were native English speakers recruited from the U.S., U.K., Canada, and Australia. The annotation task comprised 1) a consent form, 2) a pre-survey about conspiracy thinking and related psychological predispositions and sociodemographics, 3) task instructions and codebook description, 4) a training task, 5) annotation of 10 documents for markers and three-way classification (conspiracy, not conspiracy, can't tell), and 6) a debriefing page. Annotators were compensated with £9/h on average. To ensure quality, whenever an annotator did not annotate more than two markers in at least one document of the batch, we discarded the entire batch from the final data.

Statistics The annotated dataset comprises 5,354 summary statements (see Tab. 2 for summary statistics). Annotators substantially agreed, at a Krippendorff's $\alpha = 0.58$ for the binary conspiracy labels, and unitizing $\alpha = 0.15$ (macro F1=0.19 at IoU=0.5) for marker spans. Data encompasses 226 subreddits, with 1256 comments coming from the `conspiracy` subreddit. The dataset is available [1] under a CC-BY license. A sample of 20% of the data was retained as a private test set. Listing 1 showcases an example data point.

[1] https://dx.doi.org/10.5281/zenodo.15114172

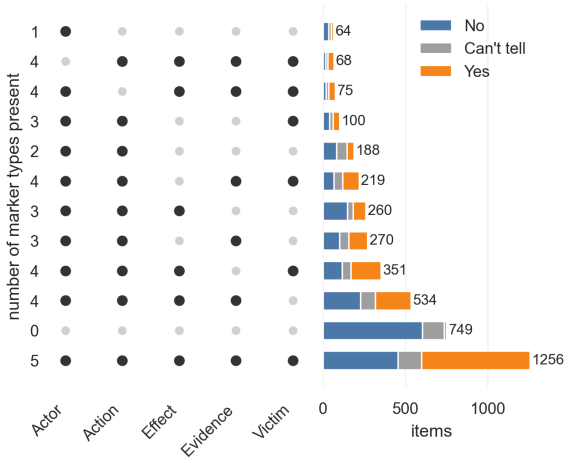

Figure 3: Disaggregation of conspiracy labels based on the observed combinations of marker types in the dataset. The more complete the set of markers a narrative contains, the more likely it relates a conspiracy theory, thus validating the hypothesized association between theoretical drivers of conspiracy thinking and their pragmatics in conspiracy narratives.

3.1 Data Exploration

Next, we describe the annotated dataset to surface opportunities for modeling and to provide due context for our evaluation methodology.

Label Composition Labels are far from homogeneous within communities, and the "hardest" ones are not always the most conspiratorial (see Fig. 2). Some cases follow expectations about their conspiracy theory content, such as `media_criticism` and `conspiracy_commons` standing out as high-Yes communities (>60% Yes), while `Futurology`, `Africa`, and `news` skew heavily No. Yet, most subreddits show a balanced mixture of labels: e.g., the majority of comments from the `conspiracy` community are labeled either No or Can't tell. Furthermore, the substantial Can't tell slices in subreddits like `centrist`, `CoronavirusCirclejerk`, `HongKong`, and `Israel_Palestine` signal that ambiguity is frequent, and may be structurally tied to political and epistemic gray zones — exactly where a binary conspiracy/non-conspiracy framing strains most visibly. Once closely inspected, conspiracy theories appear present in a broader range of topics than may be expected — and perhaps surprisingly, their frequency is not substantially different between fringe and mainstream conversational spaces. Therefore, modeling topical and community contexts may provide some, but limited information for the task.

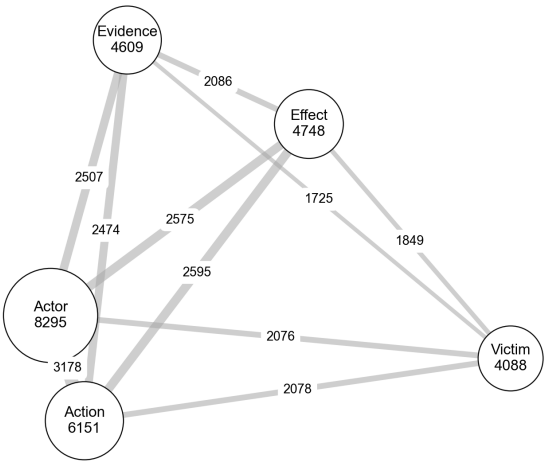

Figure 4: Network of annotated spans based on their marker type. Node size reflects the number of annotated spans; edge weight corresponds to the number of documents containing each pair of marker types.

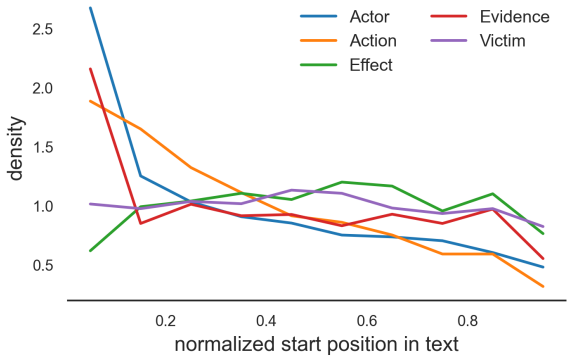

Figure 5: Distribution of marker positions in submission statements' text. Narratives typically start with Actor, Action, and Evidence.

Marker Distribution across Labels The presence of all five marker types simultaneously is strongly associated with the conspiracy label (see Fig. 3): items with all five markers (n=1,256) yield the highest Yes rate of any configuration, while items with zero markers (n=749) are predominantly labeled No, with varying proportions in between. Yet marker count alone is not a sufficient signal of the document's label: e.g., the Actor-only configuration (n=64) is overwhelmingly No, and several four-marker combinations show substantial No and Can't tell fractions, revealing that it is the specific combination of markers, not sheer quantity, that constitutes a conspiracy narrative. Yet, this association opens up opportunities for joint modeling of marker extraction and document classification.

Marker Co-Occurrence Figure 4 shows the co-presence of marker spans within documents. Actor (8,295 instances) and Action (6,151) are the backbone of the annotation space, and their co-occurrence edge (3,178) is the heaviest in the network, confirming that most annotated spans describe an agent doing something. Effect (4,748) and Evidence (4,609) cluster tightly with both Actor and Action, forming a near-complete subgraph that mirrors the rhetorical spine of a conspiracy claim: someone does something, with evidence, causing harm. Victim co-occurs least with Evidence (1,725) and most with Action (2,078), suggesting that victim framing in conspiracy text is structurally decoupled from evidentiary language — harm is asserted, not argued. Therefore, beyond the presence and number of markers, their relations appear to reflect the rhetorical structure of conspiracy narratives, information which may be leveraged in the marker extraction task.

Marker Position in Texts The rhetorical relations between markers are further elucidated by their position in the narrative's progression. Actors, Actions, and Evidence front-load their claims: the three marker types peak sharply in the first 20% of a document and decay monotonically toward the end, suggesting that conspiracy narratives tend to establish "who did what" and "why we know" early, before elaborating consequences. Effect markers, by contrast, show a mild right skew, accumulating through the middle and tail of texts, consistent with a rhetorical structure where harm or outcome is narrated after the conspiratorial premise has been established. Victim markers remain near-uniform throughout, indicating they are opportunistically placed rather than structurally motivated.

4 Evaluation Methodology

In light of our data analysis, our evaluation framework is designed to assess both the local precision of marker extraction (Subtask 1) and the global accuracy of conspiracy detection (Subtask 2). We supplement standard lexical metrics with a Bayesian hierarchical analysis to account for the inherent variance in subreddit difficulty and participant performance.

4.1 Subtask 1: Marker Extraction (Span Identification)

Evaluating the extraction of psycholinguistic markers presents a distinct challenge due to the "fuzzy"

boundaries of spans representing abstract concepts like *Threat*. Following recent span-labeling tasks (Vazquez et al., 2025), we adopt a **Token-based Intersection over Union (IoU)** matching strategy.

Span Matching We represent each predicted span S_p and ground-truth span S_g as sets of token indices. A prediction is classified as a True Positive (TP) if it satisfies an overlap threshold $\tau \geq 0.5$ with a ground-truth span of the same category:

$$\text{IoU}(S_p, S_g) = \frac{|S_p \cap S_g|}{|S_p \cup S_g|} \geq 0.5$$

Unmatched predictions are counted as False Positives (FP), and unmatched ground-truth spans as False Negatives (FN).

Aggregation We calculate Precision, Recall, and F_1-score per marker type. The primary metric for Subtask 1 is the **Macro-averaged** F_1, which ensures that models are evaluated on their ability to identify rare, reasoning-heavy markers (e.g., *Evidence*) as effectively as more frequent entity-based markers (e.g., *Actor*).

4.2 Subtask 2: Conspiracy Detection

Subtask 2 is evaluated as a binary classification problem (*Conspiracy* vs. *Non-Conspiracy*).

Primary Metric The official ranking is determined by weighted-F_1. Given that our dataset features a nearly balanced label distribution, weighted-F_1 serves as a robust indicator of overall system performance, effectively capturing the balance between precision and recall while respecting the specific class frequencies within the test set. This choice ensures that the final rankings reflect a model's true generalizability across the topic-agnostic corpus, providing a stable measure that accounts for every prediction made by the systems.

4.3 Baseline Model Setup

Both baselines leverage the `distilbert-base-uncased` model as their backbone. We utilize the hidden state of the special `[CLS]` token for sequence-level representation in Subtask 2, and the full sequence of hidden states for token-level boundary detection in Subtask 1. All models were implemented using the *Hugging Face Transformers* library and were optimized using the AdamW optimizer with a linear learning rate scheduler. We maintained a consistent set of hyperparameters across both subtasks to ensure a controlled comparison.

Marker Type	Precision	Recall	F1-Score
Action	0.1342	0.1031	0.1166
Actor	0.3772	0.2277	0.2840
Effect	0.1139	0.0797	0.0938
Evidence	0.1214	0.0884	0.1023
Victim	0.2541	0.1520	0.1902
Macro Average	—	—	0.1574
Micro Average	0.2110	0.1426	**0.1702**

Table 3: Baseline results for Subtask 1 (Token-level IoU with 0.5 threshold). The *Actor* and *Victim* categories appear more linguistically distinct, while reasoning-heavy markers like *Evidence* prove significantly more challenging for the baseline model.

Subtask 1: Sequence Labeling Baseline For the marker extraction task, given that markers may overlap or appear with varying frequencies, we simplify the problem into a "one-vs-rest" binary token classification for each psychological marker type. For a sentence with n tokens, the model produces a sequence of predictions $y = (y_1, y_2, \ldots, y_n)$, where $y_i \in \{0, 1\}$. The model is trained using a standard Cross-Entropy loss calculated over the valid sub-token offsets, effectively identifying spans that align with the evolutionary markers defined in Table 1.

Subtask 2: Sequence Classification Baseline The binary conspiracy detection baseline model treats the comment as a single unit of analysis, mapping the final layer representation of the `[CLS]` token to a bipartite output space: $\mathcal{Y} = \{\text{Conspiracy, Non-Conspiracy}\}$. To ensure the model focuses on the structural and psycholinguistic invariants rather than topic-specific keywords, the training data is filtered to remove "Can't Tell" or ambiguous labels, focusing the optimization on the most prototypical conspiratorial narratives.

4.4 Baseline Performance

Model 1 achieves 0.15 F1 in the Conspiracy Marker Extraction subtask, whereas model 2 achieves 0.70 F1 in the Conspiracy Detection subtask.

While the models effectively learn information about each task, they leave substantial room for model improvement. In particular, the marker extraction baseline model performs best on Actor ($F1 = 0.28$) and Victim ($F1 = 0.19$) (see Table 3). This suggests that the baseline is primarily picking up on named entities or referential pronouns that often characterize conspiratorial "us vs. them" narratives. The low performance on Effect (0.09)

Table 4: Subtask 1: System performance and approach summary. DA columns represent: (1) training-time augmentation and (2) use of external or additional data. ✓and – denote use and non-use, respectively.

Team	F1	DA		LLM	Conspiracy Specificities	Reference
		1	*2*			
HU	0.26	✓	–	–	Marker descriptions w/ GLiNER	(Kashaf et al., 2026)
CredenceAI	0.24	–	–	–	Cross-marker attn., joint pred.	(Karan, 2026)
CCNU	0.24	✓	–	Qwen-2.5	Joint marker prediction	(Wang and Chen, 2026)
dangphuduy	0.24	–	–	–	–	(Dang Phu, 2026)
UMUTeam	0.24	–	–	–	–	(Gomez Navalon, 2026)
UCSC NLP	0.22	–	–	–	Joint marker prediction	(Marhoefer et al., 2026)
NUST PsyAI	0.21	✓	–	–	–	(Akram and Fatima, 2026)
AILS-NTUA	0.21	–	–	GPT-5.2, Claude	DD-CoT w/ counter-arguments	(Spanakis et al., 2026)
YNU-HPCC	0.21	✓	–	–	–	(Chen et al., 2026)
Team A	0.19	–	–	–	–	(Pan, 2026)
CuriosAI	0.18	–	–	Qwen3-14B	Multi-task token classification	(Yamaga et al., 2026)
CUET_320	0.18	–	–	Gemma-3-4B	Marker descriptions	(Fariha et al., 2026)
wangkongqiang	0.15	–	–	–	–	(Wang and Tan, 2026)
zhangpeng	0.14	–	–	–	–	(Zhang and Lu, 2026)
Jia	0.08	–	–	Qwen2.5-7B,-14B	Joint generative & discriminative	(Zhu, 2026)
AGAI	0.07	–	–	–	–	(Ranxiandong, 2026)

and Evidence (0.10) suggests that these markers are not merely lexical but require understanding causal superfluity or rhetorical structure, which a vanilla DistilBERT baseline lacks. With a Micro F1 of 0.1702, the task is clearly "hard", which is an open invitation for participants to use multi-task learning (leveraging Subtask 2) or reasoning-aware LLMs.

4.5 Difficulty and Robustness Analysis

Latent Difficulty Estimation We model the probability of a system i correctly classifying an instance j from subreddit k as a function of system ability θ_i and subreddit difficulty β_k:

$$P(y_{ij} = 1) = \text{logit}^{-1}(\theta_i - \beta_k)$$

By estimating the posterior distribution of β_k, we identify "minefield" subreddits where high misclassification rates are statistically significant rather than being artifacts of sampling. This allows us to distinguish between systems that are globally robust and those that overfit to "easy" topics.

Significance Testing Following best practices (Dror et al., 2018), we apply Approximate Randomization testing with $R = 10,000$ iterations to verify that the performance deltas between the top-three teams and the DistilBERT baseline are statistically significant ($p < 0.05$).

5 Task Results

PsyCoMark attracted substantial interest from the NLP community, with participation numbers confirming the relevance and timeliness of the task. Across both subtasks, a combined total of 291 registered participants engaged with the competition platform, generating an aggregate of 3,370 submissions throughout the evaluation period (October 2025 – January 2026).

5.1 Subtask 1: Conspiracy Marker Extraction

Subtask 1 drew **134 registered participants**, who collectively produced **977 submissions** across the development and test phases. The final test leaderboard on CodaBench comprised 30 teams, of which 16 are included in the official leaderboard, shown in Tab. 4. The top-performing system, submitted by team **HU**, achieved an F1 of **0.26**, followed by **CredenceAI** and **CCNU**, both at **F1 = 0.24**. The score distribution across the leaderboard (0.07–0.26) reflects the inherent difficulty of fine-grained conspiracy marker extraction, a task that requires nuanced linguistic understanding well beyond standard sequence labeling benchmarks. Notably, though, the top-performing models surpassed individual human annotators (F1 = 0.19). Figure 6 shows that a challenge shared by all systems was maintaining consistent performance across marker types. While most systems cluster around the task mean, the highlighted systems demonstrate consistent performance across most narrative markers. Notably, the extraction of the NER-style markers "Actor" and "Victim" shows both the highest performance and the highest variance, suggesting these categories were the primary differentiators in system ranking, whereas "Evidence" and "Effect" remained a bottleneck for even the highest-ranked models.

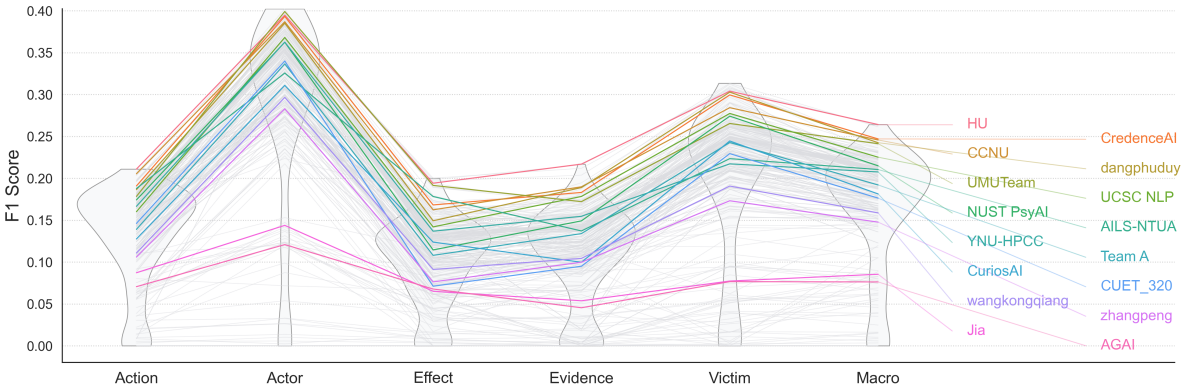

Figure 6: Per-marker system performance. Parallel coordinate trajectories for the top-performing systems are overlaid on the global distribution (represented by violin plots) of all submissions.

Most systems used pretrained transformer encoders fine-tuned for token-level sequence labeling, typically with BIO tagging to identify marker roles. Stronger submissions extended this approach with explicit span-aware modeling, including boundary-aware classifiers, span-consistency mechanisms, and structural constraints that encourage coherent contiguous spans and better alignment with the IoU-based evaluation metric. Several teams also explored joint modeling of multiple marker types, allowing the model to capture dependencies between roles within a unified representation. For example, team HU utilized GLiNER 2.0 to match text spans against semantically descriptive prompts for markers; CredenceAI incorporated a cross-marker attention module to model interdependencies between marker types; CCNU predicted markers on a joint BIO tagging space to capture cross-role contextual dependencies. Ensemble strategies and cross-validation averaging were common techniques for improving robustness. Overall, the most promising patterns combine transformer encoders with explicit span modeling, joint role prediction, and robust inference strategies, suggesting that architectures that capture both span structure and cross-role dependencies are particularly well suited for conspiracy marker extraction. This attention to the specifics of conspiracy narratives results in significantly higher performance than the baseline, and we will see next, will prove essential also in the detection subtask.

5.2 Subtask 2: Conspiracy Detection

Subtask 2 attracted **157 registered participants** and a substantially higher submission volume of **2,393 entries**, indicating sustained and active engagement throughout the competition. The final test leaderboard on CodaBench comprised 52 teams, of which 28 qualified for the official leaderboard. Performance on this binary classification subtask was considerably higher than on Subtask 1, with the leading team, **NJUST_KMG**, achieving an F1 of **0.89**. The second and third positions were held by **mdok-style** and **dangphuduy** (F1 = **0.78**), respectively. The top of the leaderboard was densely populated, with most teams scoring F1 ≥ 0.74, suggesting that the binary detection framing, while challenging, proved more tractable for participating systems than the extraction task.

Several trends emerge from comparing approaches: LLM usage, data augmentation, and explicit modeling of conspiracy narratives over general classification. First, transformer-based model backbones are almost ubiquitous, and transformer fine-tuning still outperforms pure prompting setups. Yet, most top performers integrate LLMs (often GPT/Qwen-style) in hybrid architectures. Furthermore, data augmentation correlates positively with higher performance, and is especially adopted by systems in the top quartile in the form of training-time augmentation. Pre-training or finetuning on external data, on the other hand, does not appear to provide clear performance benefits. Finally, explicit consideration of the peculiarities of conspiracy narratives, while pursued by many participants, is more common among top-quartile systems. The approaches, though, vary substantially: from identifying known conspiracy topics to separating them from stances toward conspiracy theories (e.g., to avoid the "reporter trap", in which the model incorrectly classifies objective reporting on a conspiracy theory as endorsing it), to incorporating emotional cues, and to account for ambiguity.

Table 5: Subtask 2: System performance and approach summary. DA columns represent: (1) training-time augmentation and (2) pre-trained/external data fine-tuning. ✓and – denote use and non-use, respectively.

Team	F1	DA		LLM	Conspiracy Specificities	Reference
		1	*2*			
NJUST_KMG	0.89	✓	✓	GPT-5.2	Keywords, subreddit	(Zheng and Yang, 2026)
mdok-style	0.78	✓	✓	Qwen3-32B	–	(Macko, 2026)
dangphuduy	0.78	✓	✓	GPT-OSS-120B	Sentiment analysis	(Dang Phu, 2026)
VARH-AI	0.78	✓	–	–	Task 1 markers as input	(Solanki et al., 2026)
UMUTeam	0.77	–	–	–	–	(Gomez Navalon, 2026)
HU	0.76	✓	–	–	Marker descriptions, PAN24	(Kashaf et al., 2026)
CuriosAI	0.76	✓	✓	GPT-OSS, Qwen3	Empath/VAD features	(Yamaga et al., 2026)
NUST PsyAI	0.76	✓	–	–	LIWC, NRC Lexicons	(Akram and Fatima, 2026)
CSECU-DSG	0.75	–	–	–	Sentiment analysis	(Chakraborty et al., 2026)
LATE-iimas	0.75	–	–	–	–	(Vázquez-Cerrillo et al., 2026)
davidinfotec	0.75	–	–	–	–	(Rodriguez Guierrez, 2026)
psy_detectives	0.75	–	✓	TinyLlama-1.1B	LIWC	(Carabas et al., 2026)
AILS-NTUA	0.75	–	✓	GPT-5, Claude	Topic/Stance separation	(Spanakis et al., 2026)
TruthGradient	0.75	–	–	–	–	(Goyal, 2026)
wangkongqiang	0.74	✓	✓	Gemma-3-27B	–	(Wang and Tan, 2026)
YNU-HPCC	0.74	✓	–	–	–	(Chen et al., 2026)
Unibuc-NLP	0.74	–	✓	Llama-3.1, Mistral	LIWC	(Marchitan, 2026)
Hidetsune	0.73	✓	✓	–	–	(Takahashi, 2026)
Team A	0.73	–	–	–	–	(Pan, 2026)
CCNU	0.73	✓	✓	Qwen-2.5	Multitask marker extraction	(Wang and Chen, 2026)
UCSC NLP	0.73	–	–	–	–	(Marhoefer et al., 2026)
zhangpeng	0.73	–	–	–	–	(Zhang and Lu, 2026)
Macaroni	0.73	–	–	–	Stylistic cues	(Rabehi et al., 2026)
TTLab	0.72	–	✓	Qwen2.5, Llama-3.2	–	(Richter and Marreddy, 2026)
AGAI	0.70	✓	✓	Qwen3-14B	–	(Ranxiandong, 2026)
CUET_320	0.59	–	–	Gemma-3-4B	Task 1 markers as input	(Fariha et al., 2026)
Jia	0.37	–	✓	Qwen2.5-Instruct	Multitask marker extraction	(Zhu, 2026)
GUNLP	0.34	–	✓	GPT-5, Claude	Sentiment analysis	(Ziaei et al., 2026)

A recipe for success, therefore, emerges: transformer fine-tuning + LLM component + data augmentation + marker-aware modeling. These trends are embodied by the top-performing system, which combines a transformer backbone, an LLM-based filter for comments not pertaining to conspiracy keywords, data augmentation, and explicit conspiratorial cue modeling. More specifically, (Zheng and Yang, 2026) use retrieval augmentation to update the prompt, which is then fed into the model for fine-tuning. First, samples are divided between those containing and not containing relevant keywords — unrelated topics (health, art, lifestyle) are directly classified as "No". A weighted classification is performed on the remaining conspiracy theory-related content, together with the subreddit tags obtained from the query.

Subreddit Difficulty Across all submissions, patterns emerge about which discussion communities are harder to predict correctly (Fig. 7; to ease interpretation, we divide subreddits based on whether they contained more comments labeled as "Yes" than "No", and refer to using the shorthands of "conspiracy" and "mainstream", respectively.

Mainstream subreddits are substantially easier to predict than conspiracy subreddits. Easy-to-predict mainstream subreddits have broader appeal than corresponding easy-to-predict conspiracy subreddits (`Health` vs. `AntiVaxxers`, `EverythingScience` vs. `ScienceUncensored`, `NEWPOLITIC` vs. `NewPatriotism`). Conversely, easy-to-predict conspiracy subreddits position themselves as alternatives to a mainstream (e.g., `outsideofthebox`, `awfuleverything`, `CoronavirusConspiracy`). This suggests that mainstream *positioning* may be a telltale sign of both conspiracy and non-conspiracy narratives.

Subreddits that explicitly elicit a diversity of viewpoints are among the hardest-to-predict, like `TheMotte`, `Foodforthought` (conspiracy), `TruthSeekers`, and `Israel_Palestine` (mainstream). The same is true for inherently contentious topics, such as political figures (`WayOfTheBern` and `samharris`), economics (`Wallstreetsilver` and `LateStageCapitalism`), and esoteric topics (`ufo` and `TheSaturnTimeCube`). It appears that, more than the topic, what drives difficulty is a polarizing interaction space, which may lead to the unsuspected presence or absence of conspiracy theories.

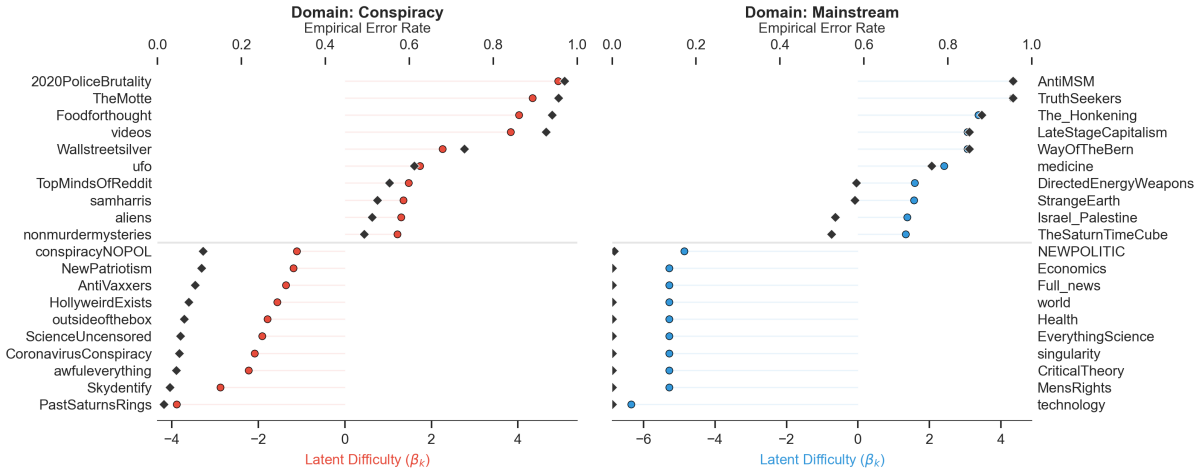

Figure 7: Top and bottom 10 most difficult-to-predict subreddits (positive $\beta_s \rightarrow$ difficult), disaggregated into "conspiracy" (left panel) and "mainstream" (right panel) based on whether they contain more "Yes" than "No" labels.

6 Related Work

Computational studies of online conspiracy theories often resort to distant-labeled data, such as, by reference to the platform or community of origin (e.g., Samory and Mitra 2018; Prooijen and Vugt 2018; Tangherlini et al. 2020; Zhang et al. 2021). Recent research contributed annotated datasets (e.g., Faddoul et al. 2020; Yi Liaw et al. 2023; Lei and Huang 2023), the majority focusing on content from Twitter and on conspiracy theories related to COVID-19 (e.g., Phillips et al. 2022; Moffitt et al. 2021; Gambini et al. 2024; Korenčić et al. 2024; Batzdorfer et al. 2022; Shahsavari et al. 2020; Langguth et al. 2023). Few datasets include conspiracy theories without topical restrictions. Most relevant to our work, Korenčić et al. present a span-based annotation schema to differentiate between conspiratorial and critical narratives (Korenčić et al., 2024). Our dataset takes stock of past work and extends it by 1) refining the annotation schema by grounding it in evolutionary psychology and 2) diversifying and expanding the topical focus of the data to a wide range of conversations.

7 Ethics

The raw data in this study were retrieved from publicly accessible archives. Although exempt, we obtained IRB approval from GESIS Cologne for the annotation task. Data is released without personal identifiers and raw comment text to preserve privacy, the right to erasure, and data minimization. We provide a script to rehydrate the comment text.[2]

[2] https://github.com/hide-ous/semeval26_starter_pack

8 Conclusions

The PsyCoMark task was particularly aimed at NLP researchers focused on explainability, misinformation detection, and fact-checking, as well as researchers on content moderators seeking granular interventions for conspiracy-related content. The high number of participants and submissions across both subtasks demonstrates strong community interest in computational approaches to conspiracy theory analysis and underscores the value of PsyCoMark as a shared evaluation framework for this emerging research area.

The contrast in difficulty between the two subtasks is clearly reflected in the performance distributions: Subtask 1 (extraction) presented a harder challenge with scores clustered in a lower range. Participants showed how accurately modeling the characteristics of conspiracy narratives yields significant gains over baselines and may achieve impressive performances. Much information, however, remains untapped. The participants explored a variety of approaches that operate at distinct lexical, semantic, local-distributional, and meta-textual levels: a principled integration is likely an avenue for further improvement. In particular, a minority of participants successfully leveraged the extracted conspiracy markers, despite their wealth of information and their strong association with conspiracy narratives found in our data exploration; further integration of the two tasks is expected to further boost systems' overall performance.

References

Mian M. Husnain Akram and Mehwish Fatima. 2026. NUST PsyAI at SemEval-2026 task 10: A comparative evaluation of psycholinguistic features, machine learning, and LLMs for conspiracy marker extraction and detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Veronika Batzdorfer, Holger Steinmetz, Marco Biella, and Meysam Alizadeh. 2022. Conspiracy theories on twitter: emerging motifs and temporal dynamics during the covid-19 pandemic. *International journal of data science and analytics*, 13(4):315–333.

Roxana Carabas, Anamaria Persida Nacu, Isac Lucian-Constantin, and Daniela Gîfu. 2026. psy detectives at SemEval-2026 task 10: PsyCoMark – psycholinguistic conspiracy marker extraction and detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Debashish Chakraborty, Sumaiya Tabassum, Sabrina Ibnath, and Abu Nowshed Chy. 2026. CSECU-DSG at SemEval-2026 task 10: Fine-tuning DeBERTa transformer model for conspiracy detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Junpei Chen, You Zhang, Jin Wang, Dan Xu, and Xuejie Zhang. 2026. YNU-HPCC at SemEval-2026 task 10: Pretrained distilbert models for conspiracy marker extraction and detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Duy Dang Phu. 2026. dangphuduy at SemEval-2026 task 10: Span-based conspiracy marker extraction and emotion-aware detection via gated fusion. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Marc Faddoul, Guillaume Chaslot, and Hany Farid. 2020. A longitudinal analysis of youtube's promotion of conspiracy videos. *Preprint*, arXiv:2003.03318.

Faozia Fariha, Lamia Tasnim Khan, Madiha Ahmed Chowdhury, Kawsar Ahmed, and Mohammed Moshiul Hoque. 2026. CUET_320 at SemEval-2026 task 10: Few-shot large language models for psycholinguistic marker extraction and conspiracy detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Margherita Gambini, Serena Tardelli, and Maurizio Tesconi. 2024. The anatomy of conspiracy theorists: Unveiling traits using a comprehensive twitter dataset. *Computer Communications*, 217:25–40.

Jorge Gomez Navalon. 2026. UMUTeam at SemEval-2026 task 10: Transformer ensembles for conspiratorial span extraction and detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Ekansh Goyal. 2026. TruthGradient at SemEval-2026 task 10: Dual-pooling DeBERTa for conspiracy theory detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Ishaan Karan. 2026. CredenceAI at SemEval-2026 task 10: A span-consistency network with cross-marker attention for conspiracy marker extraction. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Muhammad Quddussi Kashaf, Marium Zeeshan, and Shahmir Mustafa Chaudhry. 2026. HU at SemEval-2026 task 10: Leveraging GLiNER with descriptive prompts for conspiracy element extraction. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Damir Korenčić, Berta Chulvi, Xavier Bonet Casals, Alejandro Toselli, Mariona Taulé, and Paolo Rosso. 2024. What distinguishes conspiracy from critical narratives? a computational analysis of oppositional discourse. *Expert Systems*, 41(11):e13671.

Johannes Langguth, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, Jesper Phillips, and Konstantin Pogorelov. 2023. COCO: an annotated Twitter dataset of COVID-19 conspiracy theories. *Journal of Computational Social Science*, pages 1–42.

Yuanyuan Lei and Ruihong Huang. 2023. Identifying Conspiracy Theories News based on Event Relation Graph. *arXiv*.

Dominik Macko. 2026. mdok-style at SemEval-2026 task 10: Finetuning LLMs for conspiracy detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Teodor-George Marchitan. 2026. Unibuc-NLP at SemEval-2026 task 10: Unmasking conspiracies with pre-trained language models. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Dom Marhoefer, Glenn Grant-Richards, Aidan Pinero, Milos Suvakovic, and Ryan King. 2026. UCSC NLP at SemEval-2026 task 10: Boundary-aware span extraction and stratified classification for conspiracy detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Alessandro Miani, Thomas Hills, and Adrian Bangerter. 2021. Loco: The 88-million-word language of conspiracy corpus. *Behavior research methods*, pages 1–24.

Tanushree Mitra, Scott Counts, and James Pennebaker. 2016. Understanding anti-vaccination attitudes in social media. In *Proceedings of the International AAAI Conference on web and Social Media*, volume 10, pages 269–278.

J. D. Moffitt, Catherine King, and Kathleen M. Carley. 2021. Hunting conspiracy theories during the covid-19 pandemic. *Social Media + Society*, 7(3):20563051211043212.

Xintong Pan. 2026. Team a at SemEval-2026 task 10: Conspiracy extraction and classification. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Samantha C. Phillips, Lynnette Hui Xian Ng, and Kathleen M. Carley. 2022. Hoaxes and hidden agendas: A twitter conspiracy theory dataset: Data paper. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 876–880, New York, NY, USA. Association for Computing Machinery.

Jan Willem van Prooijen and Mark van Vugt. 2018. Conspiracy Theories: Evolved Functions and Psychological Mechanisms. *Perspectives on Psychological Science*, 13:770–788.

Marius Hans Raab, Stefan Andreas Ortlieb, Nikolas Auer, Klara Guthmann, and Claus-Christian Carbon. 2013. Thirty shades of truth: Conspiracy theories as stories of individuation, not of pathological delusion. *Frontiers in Psychology*, 4:1–9.

Rofaïda Rabehi, Nicolai Plenk, and Sung-Jin Miriam Han. 2026. Team Macaroni at SemEval-2026 task 10: PsyCoMark – psycholinguistic conspiracy marker extraction and detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Ranxiandong. 2026. AGAI at SemEval-2026 task 10: Enhancing conspiracy detection via instruction-tuned LLMs. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Samuel Richter and Mounika Marreddy. 2026. TT-Lab at SemEval-2026 task 10: Transformer-based approaches for psycholinguistic conspiracy detection in social media discourse. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

David Rodriguez Guierrez. 2026. davidinfotec at SemEval-2026 task 10: From lexical baselines to resource-efficient contextual transformers for conspiracy detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

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.

Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R. Tangherlini, and Vwani Roychowdhury. 2020. *Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news*, volume 3.

Hritav Singh Solanki, Shubham Sharma, and Manish Prasad. 2026. VARH-AI at SemEval-2026 task 10: Exploiting architectural diversity with transformer-ssm ensembles and confidence-based iterative refinement for conspiracy detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Panagiotis Spanakis, Maria Lymperaiou, Giorgos Filandrianos, Athanasios Voulodimos, and Giorgos Stamou. 2026. AILS-NTUA at SemEval-2026 task 10: Agentic LLMs for psycholinguistic marker extraction and conspiracy endorsement detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Hidetsune Takahashi. 2026. Hidetsune at SemEval-2026 task 10: A systematic exploration of training and inference strategies for detecting conspiracy beliefs. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Timothy R. Tangherlini, Shadi Shahsavari, Behnam Shahbazi, Ehsan Ebrahimzadeh, and Vwani Roychowdhury. 2020. An automated pipeline for the discovery of conspiracy and conspiracy theory narrative

frameworks: Bridgegate, Pizzagate and storytelling on the web. *PLoS ONE*, 15(6):e0233879.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.

José-Jorge Vázquez-Cerrillo, Helena Gómez-Adorno, and Gemma Bel Enguix. 2026. LATE-iimas at SemEval-2026 task 10: Conspiracy detection with ensemble models and loss optimization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Kongqiang Wang and Qingli Tan. 2026. wangkongqiang at SemEval-2026 task 10: PsyCoMark – psycholinguistic conspiracy marker extraction and detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Zijun Wang and Guanyi Chen. 2026. CCNU at SemEval-2026 task 10: Conspiracy marker extraction and detection via multi-task learning and LLM-based data augmentation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Daichi Yamaga, Fumika Beppu, Yuki Shibata, Aiswariya Manoj Kumar, and Takayuki Hori. 2026. CuriosAI at SemEval-2026 task 10: Hybrid approaches to conspiracy span extraction and binary detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Shao Yi Liaw, Fan Huang, Fabricio Benevenuto, Haewoon Kwak, and Jisun An. 2023. Younicon: Youtube's community of conspiracy videos. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):1102–1111.

Peng Zhang and Gehao Lu. 2026. zhangpeng at SemEval-2026 task 10: PsyCoMark – psycholinguistic conspiracy marker extraction and detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Yafei Zhang, Lin Wang, Jonathan J. H. Zhu, and Xiaofan Wang. 2021. Conspiracy vs science: A large-scale analysis of online discussion cascades. In *World Wide Web*, volume 24 of *World Wide Web*, pages 585–606.

Yuhan Zheng and Yang Yang. 2026. NJUST_KMG at SemEval-2026 task 10: PsyCoMark – subtask 2: Conspiracy detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Jiayue Zhu. 2026. Jia at SemEval-2026 task 10: A dual-track system with BERT-based encoders and LLMs for conspiracy analysis. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Rojin Ziaei, Mahsa Khoshnoodi, and Nazli Goharian. 2026. GUNLP at SemEval-2026 task 10: Improving conspiracy detection via emotion-aware augmentation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.