CUET_320 at SemEval-2026 Task 10: Few-Shot Large Language Models for Psycholinguistic Marker Extraction and Conspiracy Detection

Faozia Fariha, Lamia Tasnim Khan, Madiha Ahmed Chowdhury
Kawsar Ahmed and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
u2004012@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

Conspiracy theories widely spread on social media and can harm society by increasing mistrust, vaccine hesitancy, and political radicalization. However, most automated detection systems have traditionally relied on topic-specific classifiers, which often struggle to generalize across domains and provide little explanation for why a text is considered conspiratorial. To address these limitations, this paper explores various LLMs on the SemEval-2026 Task 10: psycholinguistic conspiracy marker extraction and binary conspiracy detection from Reddit submission statements. Specifically, we adopt a training-free few-shot prompting approach using different instruction-tuned large language models under a variety of few-shot settings ($k \in \{0, 1, 5, 10, 15, 20\}$). Within this framework, the proposed prompting strategy incorporates psychology-informed instructions to guide the models in identifying conspiracy-related signals. As a result, the presented system achieves an F1 score of 0.21 for marker extraction and 0.81 for conspiracy detection, ranking 16th out of 30 teams in Subtask 1 and 36th out of 52 in Subtask 2 without any task-specific fine-tuning. These results suggest that psycholinguistically grounded prompting can support interpretable conspiracy analysis; however, challenges remain in identifying implicit markers.

1 Introduction

Conspiracy theories spread rapidly on social media, eroding public trust and contributing to real-world harms such as vaccine hesitancy and political radicalization (Sunstein and Vermeule, 2009). Conspiratorial text typically follows a recognizable psycholinguistic pattern: a powerful *Actor* secretly performs a harmful *Action*, producing negative *Effects* on *Victims*, while citing *Evidence* to legitimize the claim. Detecting these patterns automatically is essential for content moderation at scale. Most existing systems treat conspiracy detection as binary text classification on topic-specific datasets (Curley et al., 2022), generalizing poorly across topics and offering no insight into *why* a text is conspiratorial. SemEval-2026 Task 10: PsyCoMark (Samory et al., 2025) addresses this by introducing two subtasks over Reddit comments spanning 190+ subreddits. Subtask 1 requires extracting character-level spans for five psycholinguistic markers (*Actor*, *Action*, *Effect*, *Victim*, *Evidence*), which may overlap and nest. Subtask 2 classifies whether a comment expresses conspiracy thinking (`Yes`/`No`). Figure 1 illustrates both subtasks.

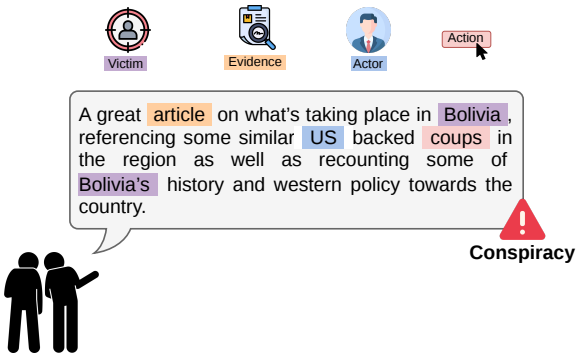

Figure 1: An example of psycholinguistic marker extraction (*Actor*, *Action*, *Effect*, *Victim*, *Evidence*) and conspiracy detection.

We explored a unified training-free few-shot prompting strategy to output verbatim marker spans as JSON for Subtask 1 and binary labels for Subtask 2. Our main contributions can be summarized as follows:

- We present a systematic evaluation of multiple instruction-tuned LLMs- Gemma-3-4B, Qwen-2.5-7B, Phi-4-mini, Llama-3.2-3B, and Ministral-3B-under varying few-shot settings across both subtasks of the PsyCoMark bench-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3495–3507

mark, providing a detailed comparative analysis of their performance.

- We demonstrate that training-free few-shot prompting can achieve competitive performance on sensitive tasks such as psycholinguistic marker extraction and conspiracy detection, approaching results of task-specific fine-tuned baseline systems.

2 Related Work

Research on misinformation and deceptive news has identified linguistic patterns that distinguish unreliable content from factual reporting. Studies comparing mainstream news with satire, hoaxes, and propaganda show that untrustworthy texts often contain stylistic markers, such as subjective language, intensifying adverbs, and first-person singular pronouns, that dramatize narratives (Rashkin et al., 2017). Beyond internal linguistic cues, systems such as DeClarE assess claim credibility by aggregating signals from external evidence articles and evaluating the trustworthiness of their sources (Popat et al., 2018). The spread of such narratives is further amplified by online polarization and users tendency to consume information aligned with their beliefs, often leading to highly segregated echo chambers (Vosoughi et al., 2018; Cinelli et al., 2021). From a modeling perspective, early sequence labeling approach, such as Named Entity Recognition (NER), used bidirectional LSTMs and BIO-style tagging schemes to identify spans in text (Lample et al., 2016). Later work demonstrated that contextual embeddings from BERT-based architectures can achieve strong performance for extraction tasks without relying on handcrafted features (Shi and Lin, 2019). More recent research has explored unified generative frameworks that treat information extraction as a sequence generation problem, enabling models to capture flat, nested, and discontinuous spans within a single architecture (Yan et al., 2021). The Universal Information Extraction (UIE) framework further introduces a Structured Extraction Language (SEL) and schema-based prompts for modeling diverse extraction tasks in a unified text-to-structure format (Lu et al., 2022).

Recent studies show that large language models can perform complex tasks through few-shot prompting, where demonstrations mainly specify the label space, input distribution, and output format rather than explicit mappings (Min et al., 2022). Chain-of-thought prompting further improves reasoning by introducing intermediate natural language steps within prompts (Wei et al., 2022). Unlike earlier systems that rely on extensive labeled data or external evidence sources, our approach uses few-shot prompting to detect conspiracy narratives while simultaneously identifying their associated psycholinguistic markers with minimal supervision.

3 Dataset and Task Description

PsyCoMark (Samory et al., 2025) is a Reddit-based dataset of *submission statements* annotated for (i) span-level psycholinguistic conspiracy markers (Subtask 1) and (ii) a binary `conspiracy` value (Subtask 2). Submission statements are user-written summaries accompanying media links, collected from over 190 subreddits spanning March 2013 to December 2023. Approximately one-quarter of the data is oversampled from *r/conspiracy* to ensure a sufficient number of conspiracy-positive instances; as a result, the dataset is not representative of the overall prevalence of conspiracy content on Reddit.

3.1 Data Format

The official PsyCoMark release provides redacted Reddit instances in which the raw text content is omitted. Each record contains a unique identifier (`_id`), gold span annotations for Subtask 1 (`markers`) represented as character offsets (`startIndex`, `endIndex`) with an associated marker `type`, and a `conspiracy` value (*Yes*, *No*, or *Can't Tell*) for Subtask 2. Since the redacted release does not include the original text, we reconstruct each submission statement using a rehydration process[1]. This step restores the textual content while preserving the original annotation structure. The character offsets remain unchanged during preprocessing, allowing direct alignment between predicted spans and the gold annotations in the required `startIndex`/`endIndex` format used by the official evaluation protocol. The resulting reconstructed text forms the final input sequence used for all prompting and evaluation experiments. Table 1 summarizes the dataset splits, `conspiracy`

[1]Rehydration refers to retrieving the original Reddit text using the provided post identifiers through the official API or archived datasets. Find the related scripts here-https://github.com/hide-ous/semeval26_task10_starter_pack

Split	Samples	Markers	Chars (min/avg/max)
Train	4,361	15,388	160/415/1,000
Dev	100	456	163/418/978
Test	938	–	157/419/986
Marker Extraction (Task 1)			
	Train		***Dev***
Actor	3,639 (23.7%)		136 (29.8%)
Action	3,601 (23.4%)		104 (22.8%)
Effect	2,916 (19.0%)		71 (15.6%)
Evidence	2,854 (18.5%)		73 (16.0%)
Victim	2,378 (15.5%)		72 (15.8%)
Conspiracy Detection (Task 2)			
Yes	1,715 (43.1%)		27 (35.1%)
No	2,263 (56.9%)		50 (64.9%)
877 / 23 Can't Tell *excluded (Train / Dev)*			

Table 1: PsyCoMark dataset statistics. Top: split sizes, annotated marker spans (Subtask 1), and character-length statistics after preprocessing; marker spans are not released for the Test split. Middle and bottom: marker-type and conspiracy-value distributions for the Train and Dev splits.

Input (with highlighted marker spans)	Conspiracy
"A great article on what's taking place in Bolivia, referencing some similar US backed coups in the region as well as recounting some of Bolivia's history and western policy towards the country."	Yes
"So they want us to believe it was a suicide, … he was just an asset for the people in power. He was under suicide watch, having 2 guards to watch him …"	Yes
" Germany has upset other EU member states by securing a disproportionately large share of vaccines, according to a report …"	No

Table 2: Annotated PsyCoMark examples after rehydration and normalization. Color highlights indicate marker types– *Actor*, *Action*, *Effect*, *Evidence*, *Victim*.

and marker distributions, and descriptive statistics of the preprocessed text lengths. Figure 3 illustrates the training data distributions, and Table 2 presents representative examples after the rehydration and normalization steps.

3.2 Task Definitions

Subtask 1: Conspiracy marker extraction. Given a submission statement, systems must extract character-level spans for five psycholinguistic markers: *Actor* (allegedly responsible party), *Action* (alleged behavior or plan), *Effect* (negative consequences), *Victim* (targeted or harmed entity), and *Evidence* (cues used to argue the conspiracy exists). Markers may overlap or nest; approximately 23% of training instances exhibit overlapping spans.

Subtask 2: Conspiracy detection. Given a submission statement, systems must assign a binary *Yes*/*No* `conspiracy` value indicating whether the statement expresses conspiracy thinking. *Can't Tell* instances (877 in training, 23 in dev) are excluded from binary evaluation, leaving a moderate 1.3:1 class imbalance (No:Yes).

4 System Overview

We utilized a few-shot prompting pipeline for both PsyCoMark subtasks. Given rehydrated and normalized text, each document is processed through a prompt-based inference stage, with subtask-specific output formats and post-processing.

4.1 LLMs and Few-Shot Learning

We design a prompting framework for Psycholinguistic Conspiracy Marker Extraction and Psycholinguistic Conspiracy Detection using instruction-tuned large language models. Figure 2 illustrates the overall pipeline.

We evaluate instruction-tuned generative models such as Gemma-3-4B (Team et al., 2025), Qwen2.5-7B-Instruct (Qwen et al., 2025), Phi-4-mini-Instruct (Microsoft et al., 2025), Llama-3.2-3B-Instruct (Grattafiori et al., 2024), and Ministral-3B-Instruct (Liu et al., 2026) via the UnSloth framework[2]. All experiments are conducted through the UnSloth framework with frozen model parameters (i.e., no gradient updates). These models were selected for their strong instruction-following ability and compatibility with prompt-based structured generation.

To analyze the effect of demonstration size, we evaluate each model under varying few-shot settings with ($k \in \{0, 1, 5, 10, 15, 20\}$) examples. Demonstrations are sampled from the training set using a fixed, stratified selection strategy to ensure coverage of all five marker types (*Actor, Action, Effect, Victim, Evidence*) as well as both conspiracy-positive and conspiracy-negative instances.

We construct a fixed pool of 20 curated examples. For $k = 1$, we use the first example (the CIA honeypot post). For $k = 5$, we use the initial five examples; for $k = 10$, we include the next five ex-

[2] https://docs.unsloth.ai/

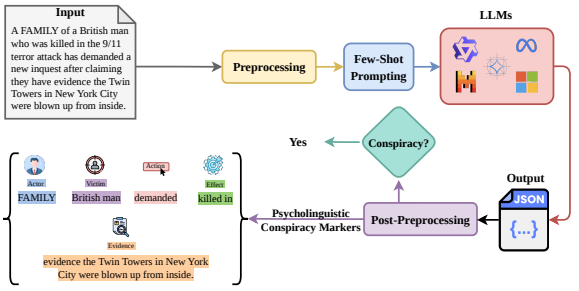

Figure 2: Few-shot prompting technique was used in this work. The same input text is used for both subtasks: the LLM generates a JSON list of verbatim marker spans (Subtask 1) and a binary conspiracy label (Subtask 2); the span outputs are then aligned to character offsets in post-processing.

amples; and so on, up to all 20 examples. This deterministic prefix-based selection ensures consistency across all experiments. The full ordered list of demonstrations is provided in Appendix **??** to ensure reproducibility. Every inference prompt contains: a system preamble defining the five psycholinguistic marker types along with the binary conspiracy label, k number of demonstrations pairing Reddit statements with gold outputs, and the target statement. For Psycholinguistic Conspiracy Marker Extraction, the model outputs a JSON array of the form:

```
[{"type":"Actor","text":"..."}, ...]
```

where every `text` value must appear verbatim in the input, enforcing character-level alignment with the original content. For Psycholinguistic Conspiracy Detection, a similar prompt structure is used with the model emitting a binary `Yes`/`No` label.

Because gold marker annotations may overlap or nest, we treat extraction as *structured generation* rather than token-level sequence labeling. Few-shot demonstrations directly calibrate span boundary decisions, multi-marker outputs, and verbatim copying without any parameter updates. Detailed prompt templates are given in Appendix A.6.

4.2 Post-Processing and Evaluation

For Subtask 1, model outputs are parsed via regex to extract JSON arrays. Each span is aligned to the source text by exact substring matching; duplicates with identical (`startIndex`, `endIndex`, `type`) are removed. For Subtask 2, the final label is determined by majority vote over three self-consistency samples. We evaluate Subtask 1 using overlap macro-F1 (IoU ≥ 0.5) and Subtask 2 using macro-F1 over `Yes`/`No`, excluding `Can't Tell` instances. Full metric details are in Appendix A.2. We compare against the official DistilBERT BIO-tagging baseline,[3] which achieves F1 = 0.15 on the development set.

Model	Few-Shots	Subtask-1			Subtask-2
		P	R	F1	W-F1
Baseline					
DistilBERT	FT	–	–	**0.15**	**0.76**
LLMs					
Qwen2.5-7B	0	0.227	0.021	0.040	**0.757**
	1	0.290	0.054	0.092	0.693
	5	0.137	0.017	0.031	0.703
	10	0.342	0.059	0.100	0.728
	15	0.250	0.094	**0.137**	0.634
	20	0.254	0.059	0.096	0.673
Phi-4-mini	0	0.242	0.09	0.13	0.684
	1	0.243	0.041	0.07	0.624
	5	0.169	0.127	**0.145**	0.716
	10	0.31	0.07	0.12	0.735
	15	0.21	0.01	0.14	0.740
	20	0.34	0.08	0.13	**0.742**
Llama-3.2-3B	0	0.221	0.1557	**0.182**	0.681
	1	0.153	0.131	0.141	**0.695**
	5	0.199	0.166	0.181	0.657
	10	0.170	0.151	0.160	0.448
	15	0.161	0.144	0.152	0.476
	20	0.142	0.127	0.134	0.548
Ministral-3B	0	0.232	0.140	0.175	**0.745**
	1	0.210	0.142	0.169	0.740
	5	0.213	0.151	**0.176**	0.727
	10	0.190	0.105	0.135	0.721
	15	0.261	0.099	0.143	0.714
	20	0.183	0.164	0.173	0.688
Gemma-3-4B	0	0.224	0.167	0.191	**0.818**
	1	0.224	0.188	0.205	0.758
	5	0.273	0.168	0.208	0.718
	10	0.236	0.193	**0.212**	0.649
	15	0.193	0.120	0.149	0.648
	20	0.223	0.145	0.175	0.623

Table 3: PsyCoMark development results for the baseline and instruction-tuned LLMs under varying few-shot settings. Subtask 1 uses span-level evaluation (P/R and micro *Overlap F1*, IoU$\geq$0.5). Subtask 2 is evaluated with weighted F1 (W-F1).

5 Results and Discussion

Table 3 summarizes the performance of all evaluated approaches across different prompting configurations. The results compare a fine-tuned transformer baseline with several instruction-tuned LLMs under varying few-shot settings.

Baseline vs Few-shot LLMs. The fine-tuned DistilBERT baseline achieves an F1 score of **0.15** for marker extraction and a weighted F1 of **0.76** for conspiracy detection. Notably, several LLM configurations surpass the baseline on marker extraction without task-specific fine-tuning. Gemma-3-4B achieves the highest span-level F1 of **0.212**

[3] https://github.com/hide-ous/semeval26_task10_starter_pack

(10-shot) across all evaluated models, followed by Llama-3.2-3B at **0.182** (zero-shot) and Ministral-3B at **0.176** (5-shot), demonstrating that training-free prompting can be competitive for fine-grained span prediction. Phi-4-mini (5-shot) also achieves **0.145** F1, approaching the baseline. For conspiracy detection, Gemma-3-4B stands out with **0.818** weighted F1 in the zero-shot setting, substantially exceeding the fine-tuned baseline. Multiple other LLM configurations also match or surpass the baseline: Qwen2.5-7B achieves **0.757** weighted F1 in zero-shot, while Ministral-3B (0-shot) and Phi-4-mini (20-shot) reach **0.745** and **0.742**, respectively.

Effect of Few-shot Demonstrations. The impact of increasing the number of demonstrations varies significantly across models. For some models, such as Phi-4-mini, additional examples consistently improve performance in conspiracy detection, increasing from 0.684 (0-shot) to **0.742** (20-shot). This suggests that the model effectively learns the decision boundary between conspiracy and non-conspiracy content from demonstrations. In contrast, Qwen2.5-7B and Gemma-3-4B both perform best in the zero-shot setting for conspiracy detection (0.757 and 0.818, respectively) and experience monotonic performance drops as demonstrations are added. For Gemma-3-4B, the weighted F1 declines steadily from 0.818 (0-shot) to 0.623 (20-shot), suggesting that the demonstration pool introduces a labelling bias that progressively undermines generalisation to the majority class. For marker extraction, Gemma-3-4B benefits from a moderate number of demonstrations, improving from 0.191 (0-shot) to **0.212** (10-shot) before declining at higher shot counts. This mirrors the broader pattern across models: increasing demonstrations does not lead to consistent improvements and in several cases performance fluctuates or declines as more examples are added, likely due to prompt length constraints and increased lexical diversity within demonstrations diluting the model's focus on the target document.

5.1 Ablation Study

We examine how specific prompt design choices affect model performance across both subtasks, treating our systematic variation of k and prompt structure as a controlled ablation.

Effect of psycholinguistic marker definitions in the system prompt: Including explicit natural-language definitions of all five marker types (*Actor, Action, Effect, Victim, Evidence*) in the instruction header directly shapes what the model attends to during generation. Without these definitions (i.e., zero-shot with a bare instruction), models such as Gemma-3-4B still achieve F1 = 0.191 for marker extraction, confirming that pretrained representations partially encode these concepts. However, the definitions serve a critical role in constraining output format: they reduce hallucinated marker types and enforce the five-class taxonomy. Removing or simplifying these definitions in pilot runs consistently increased the rate of out-of-schema outputs, particularly for *Evidence* and *Victim*, which have less prototypical surface forms.

Effect of JSON output constraint: Enforcing a strict JSON-only output format via the prompt rules ("Output ONLY a JSON array", "no explanation") has a measurable impact on post-processing reliability. In early pilot experiments without this constraint, models frequently interleaved reasoning text with span outputs, causing regex-based parsing failures that inflated false negative counts. The strict format rule reduced parse failures substantially, particularly for Phi-4-mini and Llama-3.2-3B, which are more prone to verbose outputs. This suggests that output format specification is as important as semantic instruction quality for structured extraction tasks.

Effect of shot count as an implicit ablation: Varying $k \in \{0, 1, 5, 10, 15, 20\}$ across five models reveals how demonstration quantity interacts with prompt design. For marker extraction, a moderate number of demonstrations (5–10 shots) generally improves performance by illustrating span boundary decisions and verbatim copying behaviour that definitions alone cannot convey. Gemma-3-4B improves from F1 = 0.191 (0-shot) to **0.212** (10-shot) through this mechanism. Beyond 10 shots, however, performance degrades across most models, indicating that excessively long prompts dilute the model's focus on the target document. For conspiracy detection, the effect is reversed for several models: Gemma-3-4B and Qwen2.5-7B perform best at zero-shot (0.818 and 0.757 weighted F1, respectively), suggesting that the binary classification decision is better guided by the definitional instruction alone than by noisy demonstration labels.

Effect of demonstration label balance: Our fixed pool of 20 curated demonstrations is skewed toward conspiracy-positive instances (16 out of 20 labelled *Yes*), which introduces a systematic label bias as shot count increases. This design choice has a direct and measurable impact: for Gemma-3-4B, weighted F1 on conspiracy detection declines monotonically from 0.818 (0-shot) to 0.623 (20-shot), and a similar degradation is observed for Qwen2.5-7B. As the model is exposed to more demonstrations, the skewed label distribution causes it to over-generalise the *Yes* label, reducing precision on non-conspiratorial instances. This confirms that label balance within the demonstration pool is a critical prompt design factor for classification tasks, independent of model size or architecture. A balanced demonstration pool with equal *Yes*/*No* representation would likely mitigate this effect, and remains an important direction for future work.

6 Conclusion

We presented a training-free few-shot prompting framework for SemEval-2026 Task 10 (PsyCoMark (Samory et al., 2025; Ghosh et al., 2026)), addressing psycholinguistic conspiracy marker extraction and binary conspiracy detection across different LLMs under few-shot settings. By framing marker extraction as structured JSON generation and grounding prompts in established psycholinguistic theory, our system avoids task-specific fine-tuning while remaining interpretable by design. On the official test set, our best configurations achieve a macro F1 of 0.18 for marker extraction and 0.72 for conspiracy detection. For the classification task, Phi-4-mini's monotonic improvement with increasing demonstrations and Qwen2.5-7B's strong zero-shot performance demonstrate that instruction-tuned LLMs can approach fine-tuned baselines with minimal supervision. For span extraction, however, the persistent precisionrecall gap particularly for implicit *Evidence* and *Victim* markers reveals the fundamental limits of prompting alone for fine-grained, overlapping span prediction. These findings suggest that psycholinguistically grounded prompting offers a practical and interpretable entry point for conspiracy analysis, but that closing the performance gap with supervised methods will require richer supervision signals. Future work will explore joint modeling of extraction and classification, retrieval-augmented demonstration selection, and span-aware fine-tuning strategies to improve recall on rhetorically subtle conspiracy framing.

Limitations

For Psycholinguistic Conspiracy Marker Extraction, recall remains the primary bottleneck (best F1 = 0.182), with models consistently missing implicit *Victim* and *Evidence* spans that require pragmatic inference. The strict verbatim span constraint further penalizes semantically correct extractions with slightly different surface forms. For Psycholinguistic Conspiracy Detection, results vary considerably across models and shot counts, indicating sensitivity to prompt design and demonstration selection. Smaller models suffer from context overload at higher shot settings, limiting practical utility. More broadly, the training-free pipeline cannot adapt to PsyCoMark's specific annotation conventions. The system should not be deployed for real-world content moderation without extensive fairness auditing.

References

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9).

Cliona Curley, Eugenia Siapera, and Joe Carthy. 2022. Covid-19 protesters and the far right on telegram: Co-conspirators or accidental bedfellows? *Social Media + Society*, 8.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien

Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of ACL*.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of EMNLP*.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of EMNLP*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of EMNLP*.

Mattia Samory, Felix Soldner, and Veronika Batzdorfer. 2025. Psycomark - psycholinguistic conspiracy marker dataset.

Peng Shi and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling. In *arXiv preprint arXiv:1904.05255*.

Cass R Sunstein and Adrian Vermeule. 2009. Conspiracy theories: Causes and cures. *Journal of Political Philosophy*, 17(2):202–227.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various NER subtasks. In *Proceedings of ACL-IJCNLP*.

A Additional Materials

A.1 Experimental Setup

All experiments were conducted on Kaggle using a Tesla T4 GPU (16 GB VRAM), 30 GB RAM, running Python 3.12 with PyTorch, CUDA, and Hugging Face Transformers via the UnSloth framework. We follow the official split of 4,361/100/938 documents for train/dev/test. No fine-tuning is performed; the training set is used only to select few-shot prompt examples, and the development set is used to tune decoding limits and the recall re-prompt threshold. The test set is used solely for final submission.

We evaluate four instruction-tuned LLMs Qwen2.5-7B-Instruct, Phi-4-mini-Instruct, Llama-3.2-3B-Instruct, and Ministral-3B-Instruct under six shot settings ($k \in \{0, 1, 5, 10, 15, 20\}$) across both subtasks. All models are loaded in 4-bit quantization via UnSloth for memory efficiency. For Psycholinguistic Conspiracy Marker Extraction, a high-recall re-prompt is triggered when fewer than two markers are produced in the first pass. Both passes are merged and deduplicated by `(startIndex, endIndex, type)`. Subtask 1 is evaluated with overlap-based macro F1 (IoU $\geq$ 0.5, averaged over five marker types) and Subtask 2 with standard macro F1. Table 4 summarizes the key hyperparameters used across all experiments.

Hyperparameter	Value
Max Input Tokens	1024
Max New Tokens	500
Temperature	0.0
Do Sample	False
Quantization	4-bit (NF4)
Few-shot settings (k)	0, 1, 5, 10, 15, 20
Re-prompt threshold	$<$2 markers

Table 4: Hyperparameters used across all experiments.

A.2 Evaluation Metrics

We follow the official SemEval–2026 Task 10 evaluation protocol[4]. Subtask 1 is scored with micro *Overlap F1* , while Subtask 2 is scored with *Weighted F1*.

Let $g = [g_s, g_e)$ be a gold span and $p = [p_s, p_e)$ be a predicted span. Their overlap is measured using intersection-over-union (IoU):

$$\mathrm{IoU}(g, p) = \frac{|g \cap p|}{|g \cup p|}. \quad (1)$$

A predicted span counts as a true positive if it matches a gold span of the same marker type with $\mathrm{IoU}(g, p) \geq 0.5$. Unmatched predictions are false positives and unmatched gold spans are false negatives. We compute micro precision and recall by pooling counts over all marker types:

$$P_\mu = \frac{\sum_t TP_t}{\sum_t (TP_t + FP_t)} \quad (2)$$

$$R_\mu = \frac{\sum_t TP_t}{\sum_t (TP_t + FN_t)}. \quad (3)$$

The reported *Overlap F1* is the corresponding micro F1:

$$F1_\mu = \frac{2P_\mu R_\mu}{P_\mu + R_\mu}. \quad (4)$$

For the binary labels $\mathcal{Y} = \{\texttt{Yes}, \texttt{No}\}$, we compute classwise F1 for each label $y \in \mathcal{Y}$ and then weight by label support n_y in the evaluation split:

$$F1_{\text{weighted}} = \frac{\sum_{y \in \mathcal{Y}} n_y \cdot F1_y}{\sum_{y \in \mathcal{Y}} n_y}. \quad (5)$$

A.3 Dataset Format

The PsyCoMark dataset (Samory et al., 2025) is released as JSONL files. In the *official redacted* format, the raw Reddit text is withheld; each line includes a comment ID (`_id`), a conspiracy label (`conspiracy` $\in$ {*Yes*, *No*, *Can't tell*}), and a list of character-offset marker spans. Each marker is stored as (`startIndex`, `endIndex`, `type`, `text`), where `text` is the verbatim span from the original comment. After rehydration and normalization, each record additionally contains a `text` field (processed submission statement), while the provided character offsets remain aligned to the normalized string.

[4] https://github.com/hide-ous/semeval26_task10_starter_pack

A.4 Training Distribution

Figure 3 reproduces the training-set distribution chart. The outer ring shows the relative frequency of each psycholinguistic marker type across all 15,388 annotated spans; the inner ring shows the binary conspiracy-label split after excluding 877 *Can't Tell* instances. Actor and Action together account for nearly half (47.1%) of all marker annotations, while Victim is the least frequent type (15.5%). The label distribution is moderately imbalanced (No : Yes $\approx$ 1.3 : 1).

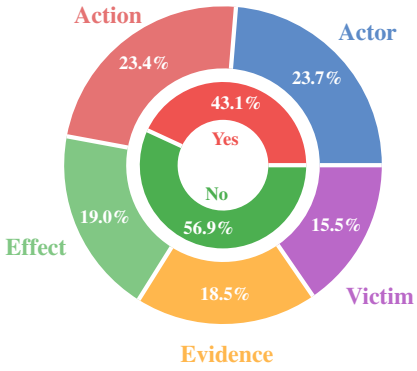

Figure 3: Training-set distribution of psycholinguistic marker types (outer ring, Subtask 1, n=15,388 spans) and binary conspiracy labels (inner ring, Subtask 2, n=3,978 after excluding *Can't Tell*). Color coding matches Tables 1 and 2.

A.5 Error Analysis

We analyse the recurring failure patterns across all models and shot configurations in Table 3.

Implicit markers are disproportionately missed. Performance degrades systematically as marker types become more implicit. *Actor* and *Action* are the most recoverable, anchored by named entities and salient verb phrases accessible through local pattern matching. *Effect* is harder, often requiring causal inference across clause boundaries. *Evidence* and *Victim* are the most problematic: both are discourse-level properties whose identification requires reasoning about the author's rhetorical intent across the full document. Whether a span constitutes Evidence depends on its pragmatic function within the conspiratorial argument; whether a participant is a Victim depends on the relationship between the Actor's action and its attributed consequences elsewhere in the text. Few-shot demonstrations cannot adequately convey this discourse-level reasoning, leading to

(a) Official redacted format

_id	conspiracy	markers (JSON format)
t1_hi2stcl	Yes	[{"startIndex":38,"endIndex":71,"type":"Actor","text":"right wing disinformation machine"}, {"startIndex":64,"endIndex":89,"type":"Action","text":"machine is making up lies"}, {"startIndex":93,"endIndex":102,"type":"Effect","text":"discredit"}, {"startIndex":103,"endIndex":109,"type":"Victim","text":"Fauci."}]
t1_g0q20p0	No	[{"startIndex":0,"endIndex":11,"type":"Actor","text":"Cindy Hendy"}, {"startIndex":31,"endIndex":47,"type":"Actor","text":"David Parker Ray"}, {"startIndex":61,"endIndex":78,"type":"Evidence","text":"Toy Box Killer"}, {"startIndex":83,"endIndex":94,"type":"Action","text":"lured women"}, {"startIndex":126,"endIndex":162,"type":"Effect","text":"torture and murder of several women."}, {"startIndex":148,"endIndex":168,"type":"Victim","text":"several women. Cindy"}]

(b) After rehydration and normalization

_id	text	conspiracy	markers (JSON format)
t1_hi2stcl	This article is actually claiming the right wing disinformation machine is making up lies to discredit Fauci. Smells more like damage control for lying to congress. I'm glad the headline is negative.	Yes	[{"startIndex":38,"endIndex":71, "type":"Actor", "text":"right wing disinformation machine"}, {"startIndex":64,"endIndex":89, "type":"Action", "text":"machine is making up lies"}, {"startIndex":93,"endIndex":102, "type":"Effect","text":"discredit"}, {"startIndex":103,"endIndex":109, "type":"Victim","text":"Fauci."}]
t1_g0q20p0	Cindy Hendy was the partner of David Parker Ray the infamous "Toy Box Killer". She lured women to Ray and participated in the torture and murder of several women. Cindy is a free woman now having been released in 2017.	No	[{"startIndex":0,"endIndex":11, "type":"Actor","text":"Cindy Hendy"}, {"startIndex":31,"endIndex":47, "type":"Actor","text":"David Parker Ray"}, {"startIndex":61,"endIndex":78, "type":"Evidence","text":"Toy Box Killer"}, {"startIndex":83,"endIndex":94, "type":"Action","text":"lured women"}, {"startIndex":126,"endIndex":162, "type":"Effect", "text":"torture and murder of several women."}, {"startIndex":148,"endIndex":168, "type":"Victim", "text":"several women. Cindy"}]

Table 5: PsyCoMark data format illustrated with two training records. **(a)** The official redacted release withholds raw Reddit text; each record provides only the comment ID, conspiracy label, and character-offset marker spans (`startIndex`, `endIndex`). **(b)** After rehydration (Arctic Shift with PullPush fallback) and normalization (Markdown → plain text, URLs → `[URL]`, Unicode/whitespace cleanup), each record gains a `text` field; character offsets remain aligned to the normalized string. Record `t1_hi2stcl` illustrates an overlapping span: *Actor* "right wing disinformation machine" (38–71) overlaps with *Action* "machine is making up lies" (64–89), sharing the token "machine".

near-zero recall on both types regardless of model or shot count.

Recall is the universal bottleneck. Across all models and shot settings, no configuration exceeds R=0.20 for marker extraction, while precision frequently reaches 0.25–0.34. Models are systematically conservative: they commit to few span predictions but those they do produce tend to be of reasonable quality. The dominant error type is therefore *omission* rather than misclassification improving coverage is the primary challenge, not filtering or boundary refinement.

More demonstrations do not consistently improve extraction. Increasing shot count does not reliably close the recall gap and in several cases actively harms performance. Longer prompts introduce lexical diversity from the demonstration pool that can dilute the model's focus on the target document, and implicit marker types require a depth of discourse understanding that surface-level examples alone cannot substitute for.

Conspiracy detection: label bias and demonstration quality. Classification performance is sensitive to both the label distribution and the content of the demonstration pool. When demonstrations are skewed toward one class, models overgeneralise that label, yielding high recall on the over-represented class at the cost of precision on the other the monotonic W-F1 decline with shot count seen for several models in Table 3 reflects this dynamic. Beyond label balance, demonstrations whose *No* examples uniformly carry empty marker contexts create a spurious correlation between marker presence and the *Yes* label, causing failures on non-conspiratorial documents that happen to contain Actor or Action spans. Finally, when *Yes* demonstrations are drawn exclusively from overt, stereotypical conspiracy texts, the model is poorly calibrated for the subtler implicit framing that characterises a substantial portion of the evaluation set. Taken together, these factors explain why classification performance degrades with shot count for most models despite the task appearing simpler than span extraction.

A.6 Prompts for Few-shot Inference

We frame both PsyCoMark subtasks as prompt-based inference with instruction-tuned LLMs. For each subtask, we use an instruction header (definitions + rules), followed by k in-context demonstrations ($k \in \{0, 1, 5, 10, 15, 20\}$), and the target instance. Tables 6 and 7 show the prompt templates used for Subtask 1 (marker extraction) and Subtask 2 (conspiracy detection), respectively.

Because Subtask 1 contains overlapping and nested spans, we use structured JSON span generation instead of BIO tagging. Few-shot examples help calibrate span boundaries and enforce verbatim extraction.

Prompt template used for Marker Extraction

You are an expert at identifying psycholinguistic conspiracy markers in Reddit comments. Extract **verbatim** text spans for one or more of the following marker types.

Marker Types:
(A) **Actor** — individuals, groups, or institutions allegedly responsible.
(B) **Action** — what the actor is doing or planning (direct/indirect).
(C) **Effect** — negative consequences or outcomes.
(D) **Victim** — who suffers the negative effects.
(E) **Evidence** — how the writer supports the claim (sources, coincidences, rebuttals).

RULES:
1. Give **no explanation**.
2. Extract **exact verbatim spans only** (no paraphrasing).
3. Output **ONLY** a JSON array.
4. Do **not** output your reasoning.
5. Spans may **overlap or nest** — extract all relevant spans.

Example:
Question: The internet has just become about spying on people. I never signed Facebook's TOS, how come they get to spy on me. Think the purpose of reddit's redesign is to spy on users?
Answer:

```
[                {"type":"Actor","text":"The
internet"},   {"type":"Actor","text":"they"},
{"type":"Action","text":"spy"},
{"type":"Effect","text":"spying"},
{"type":"Victim","text":"people."} ]
```

Now extract markers from this comment:
Question: *{text}*
Answer:

Table 6: Few-shot prompt template for marker extraction). The prompt defines the five marker types and enforces JSON-only outputs containing verbatim spans. At inference, we prepend k in-context demonstrations (not shown here for brevity).

Prompt template used for Conspiracy Detection

You are an expert at detecting conspiracy beliefs in Reddit comments. Classify each comment into **exactly one** label.

Labels:
(1) **Yes** — the text promotes, implies, or expresses conspiratorial thinking (powerful hidden actors secretly cause harm, suppress truth, or manipulate society).
(2) **No** — factual reporting, neutral discussion, sarcasm without conspiratorial intent, or genuinely ambiguous content.

RULES:
1. Give **no explanation**.
2. Output **only** the token `Yes` or `No`.
3. Do **not** output your reasoning.
4. Use extracted markers as supporting signals, but decide based on overall rhetorical intent.

Example:
Question: Maxwell claims that her email server was hacked after a court unsealed 2,000 pages of documents. The emails could showcase embarrassing information on Epstein's clients and co-conspirators in his sex trafficking operation.
Markers: Action("email server was hacked"); Effect("sex trafficking operation"); Victim("Epstein's clients, alleged victims")
Label: `Yes`

Now classify this comment:
Question: *{text}*
Extracted markers: *{markers from Subtask 1}*
Label (Yes/No):

Table 7: Few-shot prompt template for conspiracy detection (Subtask 2). The model receives the input text and (optionally) the marker spans predicted in Subtask 1, and must output a single token `Yes` or `No`.

B Few-Shot Demonstration Examples (Annotated)

Text (with highlighted marker spans) – Conspiracy

"SS: Just a reminder that various extremist groups are trying to recruit people, and are actually CIA and other Alphabet agency honeypots. The BBC is even admitting this guy is CIA, but says he is only former CIA, and they are making sure to create a link to Russia so they can continue with the narrative." – Yes

---

"New report alleges Hillary Clinton oversaw a multi-billion dollar fraud/theft, and high-ranking FBI agents are now coming forward with more details about it." – Yes

---

"Maxwell claims that her email server was hacked after a court unsealed approximately 2,000 pages of documents. If emails were obtained in the hack, they could showcase embarrassing information on Epstein's clients, alleged victims, and co-conspirators in his massive sex trafficking operation." – Yes

---

"Rudy drops radical claims implicating the highest levels of government, the ambassadors, and intelligence agencies. He names individuals and presents recorded testimony from individuals banned from the United States, allegedly to keep it quiet." – Yes

---

"The government is covering up evidence of vaccine injuries to protect pharmaceutical companies. Thousands of people have died, but mainstream media refuses to report it because they are paid off by Big Pharma." – Yes

---

"A great article on what is taking place in Bolivia, referencing similar US-backed coups in the region, as well as recounting Bolivia's history and Western policy toward the country." – Yes

---

"Chris Lehto interviews Ashton Forbes about his investigation into mysterious airline videos. They summarize evidence and analysis and debate arguments for and against the videos' authenticity." – No

---

"Germany has upset other EU member states by securing a disproportionately large share of vaccines, according to a report. Brussels ordered doses to be distributed based on population." – Yes

---

"Redditors are motivated to oppose perceived wrongdoing. This, combined with moderators permitting hostility, encourages users to act aggressively toward targeted groups." – Yes

---

"A discussion about the Virginia gubernatorial election and how ranked-choice voting influenced the Republican primary outcome." – No

---

"I saw a friend's post discussing PET's arguments. There has been extensive mythologizing around constitutional reform events. The accounts in the text by PET contradict modern interpretations." – Yes

---

"A cruise ship with coronavirus cases docked in Florida. Passengers not showing symptoms were allowed to return home without quarantine." – No

---

"The Jesuits were expelled from Rome and later colluded with Napoleon and freemasons to overthrow papal power. The pope was eventually arrested, and the order restored." – Yes

---

"A discussion post exploring ideas about the 'Third Eye' and 'Lizard People' conspiracy, where they work together." – Yes

---

" A mother reports her child suffered a severe vaccine reaction. She criticizes the government for neglect, claiming long-term suffering after vaccination." – Yes

---

" They want us to believe it was a suicide, but it is blatantly obvious otherwise. Jeffrey was an asset of people in power and was under constant surveillance." – Yes

---

" Gazan women have been murdered by Israel, with thousands displaced." – Yes

---

" Surveys suggest increasing authoritarianism, including fines and incarceration, and proposals for detaining unvaccinated individuals in facilities." – Yes

---

" China, state media plans to deliver news using AI-generated anchors with synthesized voices." – Yes

---

" Israeli sources and online commentators initially blamed Palestinians, but IDF later admitted involvement." – Yes

---

Annotation legend: Actor Action Evidence Effect Victim