MindMiner at SemEval-2026 Task 10: Multi-Model Approaches to Conspiracy Detection and Psycholinguistic Marker Extraction

Pramod Kumar Ajmeera
University of Colorado Boulder
pramod.ajmeera@colorado.edu

Akshara Sri Lakshmipathy
University of Colorado Boulder
akshara.lakshmipathy@colorado.edu

Abstract

Conspiracy narratives on social media often hide in subtle word cues and quiet reasoning patterns, making their detection a challenging task for natural language processing systems. SemEval-2026 Task 10 PsyCoMark introduces a benchmark for studying these phenomena, pairing binary conspiracy detection with the extraction of five key psycholinguistic markers: Actor, Action, Effect, Victim, and Evidence. In this paper, we examine how modern transformer-based models can grasp both the conspiratorial intent and the deeper reasoning structures behind such narratives, using rehydrated Reddit comments annotated by experts in psychology and linguistics. We test five models across these subtasks, emphasizing the gap that exists between classification and deeper discourse-level interpretation. Our best system reaches 0.80 weighted F1 on conspiracy detection and 0.16 macro F1 on marker extraction, with per-marker F1 ranging from 0.36 (Actor) to 0.00 (Victim). This work also contributes to the growing call for explainable NLP methods that integrate psycholinguistic insights to better illuminate misinformation and conspiratorial thinking online.

1 Introduction

Conspiracy theories continue to shape public discourse, influencing how people reason about institutions and how information spreads across social platforms (Douglas et al., 2019; Zhou and Zafarani, 2020). Recent work in psychology shows that conspiracy thinking extends beyond topical concerns such as vaccines, geopolitics, and government institutions. It is driven by underlying psycholinguistic processes such as identifying actors, victims, proposed activities, and cited evidence. Identifying such indicators is an especially difficult task in the context of social media.

Psycholinguistic Conspiracy Marker Extraction and Detection (Samory et al., 2026) introduces a new benchmark aimed at addressing these challenges. The competition is divided into two subtasks:

1. Subtask 2 (Conspiracy Detection) is a binary classification task, where models are asked to classify whether a Reddit comment is showing conspiratorial thinking.

2. Subtask 1 (Marker Extraction) takes it further by identifying specific markers such as Actor, Action, Effect, Victim, and Evidence, which form the "structure" of the narrative. In contrast to prior work on topic-centric conspiracy datasets, this dataset is unique in being annotated through the lens of evolutionary psychology.

We participated in this shared task for two reasons. First, automated detection of conspiratorial content has become increasingly relevant for platform moderation, where scale makes manual review impractical. Second, psycholinguistic marker extraction directly supports explainable AI: a system that not only classifies a post as conspiratorial but also identifies the specific spans that triggered that decision is far more useful for downstream auditing and intervention than a black-box label alone. In our study, we develop and compare five different modeling approaches: a baseline DistilBERT (Sanh et al., 2019), DeBERTa-v3-base, DeBERTa-v3-large, RoBERTa-large, and LLaMA-3.1-8B as a representative open-weight LLM. In the case of Subtask 2, we fine-tune the models on binary classification, and for Subtask 1, we adopt a per-marker token classification approach with simplified BIO tagging, training a separate model for each of the five marker types. This allows us to explore the trade-offs between model scale, architectural design, and task-specific generalization.

The contributions of this work are threefold:

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 22–29

- We compare five Transformer-based designs on psycholinguistic marker extraction and conspiracy detection across various model scales (66M to 8B parameters).
- We compare parameter-efficient fine-tuning of large language models (LLaMA-3.1-8B with LoRA) with encoder-only models (DistilBERT, DeBERTa, RoBERTa).
- Our findings provide competitive performance on conspiracy detection across model scales (F1: 0.75-0.80). However, marker extraction is still difficult because of overlapping span bounds and subtle differences across marker types.

Overall, our study highlights the importance of modern transformers to go beyond topical categorization to a more interpretable, psychologically grounded understanding of conspiratorial discourse. Our analysis uses the dataset released as part of the SemEval-2026 shared tasks (Ghosh et al., 2026; Samory et al., 2026).

2 Task and Data

2.1 Task Definition

The goal of the PsyCoMark challenge (Samory et al., 2026) is not only to determine whether a model can recognize conspiratorial thinking, but also to assess whether it can detect the psycholinguistic patterns associated with such reasoning. The shared task is divided into two subtasks.

2.1.1 Conspiracy Detection

Subtask 2 is a binary text classification problem. The objective is to determine whether a Reddit comment expresses conspiratorial thinking. Evaluation is performed using the weighted F1 score across the two classes (conspiracy vs. non-conspiracy).

2.1.2 Marker Extraction

Subtask 1 focuses on identifying where conspiracy reasoning appears within a comment. Annotators labeled spans corresponding to five core psycholinguistic indicators:

- **Actor**: the individuals or groups alleged to be responsible,
- **Action**: what these actors are doing or planning,
- **Effect**: the negative consequences being described,
- **Victim**: the individuals or groups harmed by the alleged conspiracy,
- **Evidence**: textual support or "proof" used to justify the conspiracy claim.

Unlike traditional NER tasks, marker spans in PsyCoMark may overlap, be nested, or appear in disjoint segments. Evaluation for Subtask 1 uses an overlap-based macro F1 score.

2.2 PsyCoMark Dataset

2.2.1 Source and Annotation

The PsyCoMark dataset is composed of Reddit comments collected from a broad range of social engagement communities. The comments were annotated by experts in psychology and linguistics. Each instance contains a binary classification label (conspiracy vs. non-conspiracy), and zero or more marker spans corresponding to psycholinguistic indicators of conspiracy reasoning.

The annotation framework draws on evolutionary psychology principles such as agency detection, threat perception, and evidence construction. This approach yields a dataset that is more generalizable than prior topic-specific conspiracy corpora.

2.2.2 Data Splits

PsyCoMark follows the SemEval dataset structure and is distributed in JSONL files. The benchmark is provided in three splits:

- **Training set**: fully labeled with binary and span labels,
- **Development set**: labeled for model validation and tuning,
- **Test set**: labels withheld; evaluation conducted through Codabench.

The dataset covers a wide variety of topics, and comments range from brief statements to long multi-sentence paragraphs.

2.2.3 Data Preprocessing

Due to Reddit's data policy, the released files include only comment IDs and annotation metadata. Participants must "rehydrate" the raw text using external APIs. The starter pack provides:

- `train_redacted.jsonl` — comment IDs and annotated spans without raw text
- `rehydrate_data.py` — script for retrieving text via Reddit APIs
- `train_rehydrated.jsonl` — final dataset containing reconstructed comments

Label Filtering. To comply with the task criteria, we eliminated samples with the label "Can't tell" so that binary classification could focus on the clear distinction between conspiracies and non-conspiracies. This filtering was applied to the training data.

Data Splits. The filtered training set contained 4,316 samples after removing "Can't tell" instances. We trained on this full set and used the official development set for model selection and hyperparameter tuning. The organizers provided the official test set for Codabench's final assessment.

Tokenization Strategy. Different models used different sequence lengths based on their architectural constraints:

- **DistilBERT**: max_length=128 (Subtask 2), 256 (Subtask 1)
- **DeBERTa-v3-base**: max_length=256 (both subtasks) with gradient checkpointing
- **DeBERTa-v3-large**: max_length=256 (Subtask 2), 128 (Subtask 1) with gradient checkpointing
- **RoBERTa-large**: max_length=512 for full context
- **LLaMA-3.1-8B**: max_length=512 (Subtask 2), 1024 (Subtask 1)

All models used `padding="max_length"` and `truncation=True`.

Span Annotation Alignment (Subtask 1). We used offset mappings to align character-level annotations with subword tokens for marker retrieval. To prevent label conflicts from overlapping spans, we adopted a simple binary tagging scheme (O vs. marker type) and trained a separate model for each of the five marker types.

2.2.4 Challenges Raised by the Dataset

The PsyCoMark dataset introduces several challenges:

- Conspiratorial cues may be subtle or implicit, making detection difficult
- Marker categories are imbalanced, with some types appearing far more frequently.
- Rehydrated comments may be incomplete or misaligned, breaking offset consistency.
- Topic diversity introduces domain shift across comment themes.
- Reasoning spans often stretch across multiple sentences, requiring long-range understanding.

3 Evaluation Metrics

3.1 Subtask 2: Conspiracy Detection

Subtask 2 is evaluated using the weighted F1 score as the primary metric, with accuracy reported as a secondary metric. Weighted F1 accounts for class imbalance by weighting each class's F1 score proportionally to its frequency in the test set, which prevents the dominant non-conspiracy class from inflating the overall score.

3.2 Subtask 1: Marker Extraction

Subtask 1 is evaluated using an overlap-based token-level F1 score. A predicted span counts as a match if its Intersection-over-Union (IoU) with the ground truth span exceeds 0.5, where IoU is computed over token index sets rather than character offsets. Two variants are reported: macro F1, which averages scores equally across all five marker types regardless of frequency, and micro F1, which aggregates counts across all spans. Macro F1 is used for leaderboard ranking since it treats rare markers like Victim on equal footing with frequent ones like Actor.

3.3 Performance Baselines

The task organizers provided baseline systems:

- **Subtask 2 baseline**: DistilBERT achieved 0.75 weighted F1 on the development set.
- **Subtask 1 baseline**: DistilBERT achieved 0.15 overlap F1 on the development set.

These baselines define the targets we aim to surpass.

Partial Match Handling The overlap-based F1 rewards partial matches, in contrast to exact-match measures. Partial credit is given to a prediction that captures the majority of a marker span, making the evaluation more robust to small boundary errors. Unmatched ground truth spans are counted as false negatives, and predictions with IoU < 0.5 are counted as false positives.

Multi-Label Evaluation Challenges Overlapping spans of various marker types are permitted in the PsyCoMark dataset. For instance, "government officials" might be marked as both Evidence and Actor. Our per-marker evaluation approach trains and assesses five distinct models, treating each marker category separately. This avoids the complexity of multi-label sequence tagging, but it requires combining the predictions from all five models.

4 System Overview

4.1 Overview of Approach

To assess the relationship between model scale, architectural design, and performance on psycholinguistic conspiracy marker tasks, we benchmark five Transformer-based architectures. Our model selection encompasses three categories: (1) an efficient distilled baseline (DistilBERT-base, 66M), (2) mid-to-large encoder-only transformers (DeBERTa-v3-base 184M, DeBERTa-v3-large 435M, RoBERTa-large 355M), and (3) a large language model with parameter-efficient fine-tuning (LLaMA-3.1-8B with LoRA).

This design allows us to examine several questions: whether larger models consistently outperform smaller ones on psycholinguistic reasoning tasks, whether parameter-efficient fine-tuning via LoRA can match fully fine-tuned encoders, and whether DeBERTa's disentangled attention provides meaningful gains over RoBERTa at comparable scale.

The filtered training set of 4,316 samples for both Subtask 2 and Subtask 1 was used to train all models. We trained on all available training instances and used the official development set for model selection and hyperparameter tuning rather than creating an additional internal train–validation split. To reduce the complexity of multi-label sequence tagging with overlapping spans, we used a separate-model-per-marker approach for Subtask 1.

A natural question is why we did not train a single joint model for all five marker types simultaneously. We chose the separate-model approach for two reasons. First, the five marker categories have very different frequency distributions in the training data, and a joint model would likely be dominated by the more frequent categories like Actor and Action while underperforming on rarer ones like Victim. Training separately allows each model to specialize on its own marker's distribution rather than competing for shared capacity. Second, the task formulation allows spans to overlap across marker types, meaning a single token can simultaneously belong to an Actor span and an Evidence span. A shared classification head would need to handle this through multi-label tagging, which adds considerable complexity without a clear performance benefit at our scale. Joint modeling with explicit span interaction remains an interesting direction for future work.

4.2 Model Architectures

Tables 1 and 2 summarize the training configurations for all five models on Subtask 2 and Subtask 1 respectively. The remainder of this section describes each model's setup and observations.

4.2.1 DistilBERT

We replicated the organizer-provided DistilBERT baseline (Sanh et al., 2019) using `train_binary.py` and `train_one_span.py` from the starter pack with `distilbert-base-uncased` (66M parameters) and no configuration changes. For Subtask 1 we trained five separate `DistilBertForQuestionAnswering` models, one per marker type. The reference scores of 0.75 F1 (Subtask 2) and 0.15 overlap F1 (Subtask 1) establish the official shared task baselines.

4.2.2 DeBERTa-v3-base

DeBERTa-v3-base (He et al., 2021) is the mid-sized DeBERTa variant (184M parameters, 12 layers). We fine-tuned it for Subtask 2 with gradient checkpointing for memory, and trained five token-classification models with simplified BIO tagging for Subtask 1. The model matches DistilBERT on both subtasks (0.75 and 0.15) despite its larger capacity, suggesting that the disentangled attention mechanism alone does not yield gains at this scale.

4.2.3 DeBERTa-v3-large

DeBERTa-v3-large is our largest encoder-only model (435M parameters). It uses relative positional encoding and an enhanced mask decoder during pretraining, which we expected to help with

Model	BS	LR	Ep	Len	GC	F1
DistilBERT	64	1.5e-5	10	128	No	0.75
DeBERTa-v3-base	8	2e-5	10	256	Yes	0.75
DeBERTa-v3-large	8	1e-5	5	256	Yes	**0.80**
RoBERTa-large	16	1e-5	5	512	No	0.79
LLaMA-3.1-8B+LoRA	2	1e-4	5	512	–	**0.80**

Table 1: Subtask 2 (Conspiracy Detection) hyperparameters and test F1 across all five models. BS = batch size, LR = learning rate, Ep = epochs, Len = max sequence length, GC = gradient checkpointing.

Model	BS	LR	Ep	Len	F1
DistilBERT	8	3e-5	5	256	0.15
DeBERTa-v3-base	8	1e-5	5	256	0.15
DeBERTa-v3-large	8	1e-5	5	128	**0.16**
RoBERTa-large	8	1e-5	5	512	0.14
LLaMA-3.1-8B+LoRA	1	1e-4	5	1024	0.14

Table 2: Subtask 1 (Marker Extraction) hyperparameters and macro overlap F1 across all five models. Encoder models use one model per marker with simplified BIO tagging; LLaMA uses a single generative model for all five markers.

fine-grained span identification. For Subtask 2 we used a linear classification head with gradient checkpointing. The training loss decreased smoothly from 0.67 to 0.19 over five epochs, and validation F1 stabilized around 0.78. On the test set the model reached 0.80 F1, matching LLaMA-3.1-8B despite being 18× smaller. The slight gain on the test set relative to validation likely reflects differences in label distribution between the two splits.

For Subtask 1 we trained five DeBERTa-v3-large token classification models with simplified BIO tagging (O vs. marker), with final training losses ranging from 0.079 (Victim) to 0.236 (Actor). Despite low training losses, the model achieved only 0.16 macro F1 on the test set. Table 3 reports the aggregate breakdown.

Metric	Value
Macro overlap F1	**0.16**
Micro overlap F1	0.20
Precision	0.234
Recall	0.182
True Positives	83
False Positives	272
False Negatives	373

Table 3: Aggregate test results for DeBERTa-v3-large on Subtask 1. The 272 false positives versus 83 true positives reveals a substantial precision-recall imbalance.

DeBERTa-v3-large emerged as our best overall model, reaching 0.80 F1 on Subtask 2 and leading marker extraction by a small margin (0.16 F1, vs. 0.14–0.15 for the other models).

4.2.4 RoBERTa-large

RoBERTa-large (Liu et al., 2019) is an enhanced BERT variant (355M parameters), trained without Next Sentence Prediction, with dynamic masking, larger batches, and longer pretraining. We used a linear classification head with batch size 16 and a 512-token maximum sequence length for Subtask 2, and five per-marker token classifiers for Subtask 1. Final Subtask 1 training losses were low across all five marker models (0.060 to 0.191), but did not translate to strong test performance (0.14 overlap F1, tied with LLaMA). On Subtask 2 the model reached 0.79 F1, almost matching the best models while training more efficiently than DeBERTa-v3-large.

4.2.5 LLaMA-3.1-8B with LoRA

Our largest model, LLaMA-3.1-8B (Grattafiori et al., 2024), has 8 billion parameters across 32 transformer layers. We used LoRA (Hu et al., 2021) with rank $r = 16$ and alpha $\alpha = 32$, training only 0.5% of the model's parameters (41.9M out of 8B). 4-bit quantization via bitsandbytes kept the model in single-GPU memory.

For Subtask 2 we framed the task generatively with the prompt *Task: Determine whether this is content related to conspiracy theories. Only respond "Yes" or "No."* followed by the comment. Training loss reached 1.509 over five epochs, and the model matched DeBERTa-v3-large at 0.80 F1.

For Subtask 1 we used a single generative model rather than five separate classifiers, with the prompt *Task: Identify conspiracy theory indicators in the text. Identify spans for Action, Actor, Effect, Evi-*

dence, and Victim. A longer 1024-token sequence length accommodated the input plus generated marker annotations. Final loss was 1.096; test overlap F1 was 0.14, tied with RoBERTa-large. Although the generative approach simplified training, it required careful prompt engineering to maintain a consistent output format and did not improve over the per-marker token classification setup (0.16 for DeBERTa).

5 Experimental Results and Discussion

5.1 Subtask 2: Conspiracy Detection

Our top-performing models, DeBERTa-v3-large and LLaMA-3.1-8B with LoRA, both obtained 0.80 weighted F1. BERT-family models remain competitive, as RoBERTa-large's 0.79 F1 shows, only 0.01 below our best models. The two smaller models, DistilBERT and DeBERTa-v3-base, both matched the organizer's baseline score of 0.75 F1.

For this task, architectural design and training quality matter as much as model size, as shown by the narrow range of scores (0.75-0.80) among models spanning 66M to 8B parameters. Notably, fully fine-tuned DeBERTa-v3-large (435M parameters) was matched by LLaMA's parameter-efficient setup (training only 41.9M parameters), indicating the effectiveness of LoRA for adapting large language models to specific classification tasks.

5.2 Subtask 1: Marker Extraction

Marker extraction proved substantially more challenging than binary classification, with our best model (DeBERTa-v3-large) achieving only 0.16 macro F1 compared to 0.80 on Subtask 2, a fivefold ratio. This difficulty reflects the fundamental challenges of psycholinguistic span detection: overlapping marker boundaries, subtle distinctions between marker types, and the ambiguous nature of conspiracy discourse where markers are often implicit rather than explicit.

DeBERTa-v3-large again performed best, achieving 0.16 macro F1 (0.20 micro F1).

5.2.1 Per-Marker Performance Analysis

A more detailed analysis of DeBERTa-v3-large reveals stark differences in marker detection difficulty (Table 4).

The dramatic variation between Actor (0.364) and Victim (0.000) reveals that different markers require fundamentally different approaches:

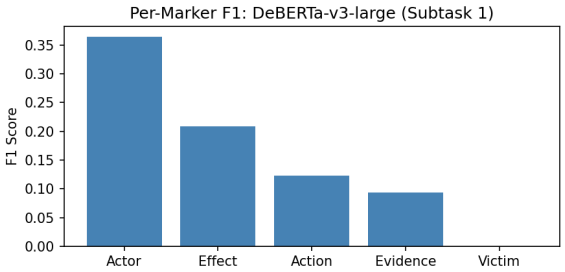

Figure 1: Per-marker F1 scores for DeBERTa-v3-large on Subtask 1, illustrating the dramatic performance gap between Actor (0.364) and Victim (0.000). The stark variation across marker types suggests that each category presents structurally different detection challenges rather than a uniform span extraction problem.

Marker Type	Precision	Recall	F1
Actor	0.368	0.360	**0.364**
Effect	0.185	0.239	0.209
Action	0.147	0.106	0.123
Evidence	0.109	0.082	0.094
Victim	0.000	0.000	**0.000**

Table 4: Per-marker F1 breakdown for DeBERTa-v3-large on Subtask 1.

- **Actor succeeds** (F1: 0.364): Actors are typically explicit noun phrases (e.g., "the government", "corporations") with clear syntactic boundaries.
- **Victim fails** (F1: 0.000): Victims are often implicit or pronominal references (e.g., "us", "the people") scattered across sentences, making span detection extremely difficult.
- **Evidence struggles** (F1: 0.094): Evidence spans frequently overlap with other markers and lack clear linguistic boundaries.

The low aggregate recall (0.182) and large number of false positives (272 FP vs. 83 TP) indicate that the model overpredicts boundaries while also missing many true markers, pointing to intrinsic uncertainty in span recognition rather than straightforward threshold problems.

5.2.2 Cross-Model Comparison

Beyond DeBERTa-v3-large, the remaining models clustered between 0.14 and 0.15 F1 with minimal differentiation. DeBERTa-v3-base and DistilBERT both achieved 0.15, while RoBERTa-large and LLaMA scored 0.14. Subtask 1 performance appears largely insensitive to model size, suggesting that the challenge lies in the fundamental task formulation rather than model capacity.

5.3 Cross-Task Analysis

Our results confirm that Subtask 2 (Conspiracy Detection) is considerably more tractable than Subtask 1 (Marker Extraction). Both DeBERTa-v3-large and LLaMA-3.1-8B reached 0.80 weighted F1 on Subtask 2, showing that encoder-based models and parameter-efficient LLMs can both pick up on conspiratorial framing effectively. Subtask 1, on the other hand, exposes the limits of current sequence labeling approaches: models handled more syntactically grounded markers like Actor and Action reasonably well but struggled with markers that require multi-sentence reasoning or implicit reference resolution.

Conspiracy theories share structural characteristics with political framing and narrative discourse, which may partly explain why span-level marker extraction demands more comprehensive discourse modeling than sentence-level classification. Closing this gap will likely require moving beyond token-level classification toward representations that capture argument structure and coreference across sentence boundaries.

6 Conclusion

This paper presented MindMiner, our system for SemEval-2026 Task 10 PsyCoMark, comparing five transformer-based models across conspiracy detection (Subtask 2) and psycholinguistic marker extraction (Subtask 1). Our results show that these two subtasks sit at fundamentally different levels of difficulty, and that this gap does not close with model scale.

For Subtask 2, DeBERTa-v3-large and LLaMA-3.1-8B with LoRA both reached 0.80 weighted F1. The narrow spread across all five models (0.75 to 0.80) suggests that modern transformer architectures have largely converged on this binary classification problem, and that further gains will likely require better data rather than larger models.

Subtask 1 told a different story. Even our best model achieved only 0.16 macro F1, and the per-marker breakdown revealed a dramatic performance gap between Actor (0.364 F1) and Victim (0.000 F1). Actors tend to be explicit noun phrases with clear boundaries, while victims are often implicit references scattered across a sentence. This tells us that the challenge is not model capacity but task structure: current span detection approaches are not well suited for psycholinguistic markers that are implicit, overlapping, and contextually grounded.

Going forward, the most productive directions are richer span representations that capture discourse-level context, marker-specific architectures that account for the structural differences between categories, and potentially reformulating Victim and Evidence detection as inference tasks rather than span extraction. The 0.000 F1 on Victim is less a failure of the model and more a signal that the task formulation itself needs rethinking for that category.

Limitations

- The marker distribution is skewed across categories, and our per-marker results show large variation in detection difficulty (F1 ranging from 0.000 for Victim to 0.364 for Actor). Models therefore handle markers with clear syntactic boundaries reasonably well while struggling on more implicit categories.

- The Reddit domain contains heavy linguistic noise such as sarcasm, slang, broken syntax, and inline hyperlinks. Even large models have trouble untangling this noise when the underlying task is conspiratorial reasoning.

- Many markers, especially Evidence and Effect, span multiple sentences and rely on long-range context. Standard Transformer models with fixed context windows do not capture these cross-sentence dependencies well, which likely contributes to weak Subtask 1 performance.

- Beyond technical limitations, our models also raise concerns about the social impact of NLP systems used for content moderation and risk detection on user-generated text (Hovy and Spruit, 2016). Misclassifying conspiratorial content can either suppress legitimate discourse or leave harmful narratives unchecked, which underscores the need for careful evaluation and transparent model behavior.

References

Karen M Douglas, Joseph E Uscinski, Robbie M Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding conspiracy theories. *Political psychology*, 40:3–35.

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, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. *Preprint*, arXiv:2006.03654.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

Mattia Samory, Felix Soldner, and Veronika Batzdorfer. 2026. SemEval-2026 task 10: PsyCoMark – psycholinguistic conspiracy marker extraction and detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.