GUNLP at SemEval-2026 Task 10: Emotion-Aware Multi-Task Learning for Conspiracy Detection

Rojin Ziaei[1*] nz204@georgetown.edu
Mahsa Khoshnoodi[1*] mk2524@georgetown.edu
Nazli Goharian[1] nazli@ir.cs.georgetown.edu

[1]**Department of Computer Science, Georgetown University**

Abstract

This paper presents the Georgetown University NLP (GUNLP) system for SemEval 2026 Task 10: Psycholinguistic Conspiracy Marker Extraction and Detection. Specifically, our system addresses Subtask 2 by classifying conspiratorial beliefs in Reddit posts. Our approach uses COVID-Twitter-BERT v2 (CT-BERT-v2) (Müller et al., 2023), a BERT-large model pretrained on COVID-19 social media data. We integrate this model into a multi-task learning framework with a dual-head architecture on the [CLS] token to jointly optimize conspiracy classification and emotion prediction. To enrich the training set, we use GPT-5 to generate chain-of-thought emotion annotations and apply paraphrase-based data augmentation. This effectively doubles the training corpus to approximately 8,600 samples. We evaluate two input configurations: text only, and text with emotion annotations. The emotion-aware configuration achieves the strongest performance, reaching an F1 score of **0.54** (**Yes**: 0.655, **No**: 0.752, **Can't Tell**: 0.222) on the official development set, with an official test set score of **0.34**[1]

1 Introduction

Conspiracy theories and misinformation have existed far longer than the technologies now used to disseminate them (Van Prooijen and Douglas, 2017; Douglas and Sutton, 2023). However, the rate at which these narratives spread is unprecedented; spreading information is faster and easier than ever before (Del Vicario et al., 2016; Vosoughi et al., 2018). In the modern digital landscape, social media platforms and recommendation algorithms determine the news we consume and the information to which we are exposed, creating fertile ground for the propagation of unsubstantiated and often harmful beliefs.

Research suggests that conspiracy theories are not merely false beliefs, but structured narratives involving specific psycholinguistic components. These include perceived persecutors ("Actors"), distinct "Actions," and threatened "Victims" (Korenčić et al., 2024; Tangherlini et al., 2020; Samory et al., 2026). These narratives often spread through high-arousal emotional dynamics rather than factual debate. The resulting harm is tangible, ranging from the erosion of trust in public institutions to the incitement of real-world violence and social polarization (Ecker et al., 2022). Consequently, automatically detecting and analyzing these narratives is a critical challenge for the natural language processing community.

Current approaches to conspiracy detection often rely on standard text classification (Holur et al., 2022), which may fail to capture the nuanced psychological structure of conspiratorial thinking (Uscinski et al., 2022). While existing models can identify overt hate speech or misinformation (Malik et al., 2025), they often fall short in distinguishing between genuine inquiry and the specific rhetorical patterns that define conspiracy theories. To address this, the SemEval 2026 Task 10: Psycholinguistic Conspiracy Marker Extraction and Detection (PsyCoMark) challenges systems to not only classify Reddit posts as conspiratorial, but also to extract the supporting psycholinguistic markers (Samory et al., 2026).

In this work, we hypothesize that surface-level keyword detection is insufficient for robust conspiracy classification; rather, the interplay between linguistic structure, emotional context, and domain-specific semantic representations is critical for accurate detection. Our central research question is whether jointly modeling conspiracy classification and emotion prediction within a domain-

[*] Equal contribution.

[1] As noted by the shared task organizers, this score reflects evaluation on the full label set, including labels absent from our training data; performance on the matched label subset was substantially higher (0.54). The official score therefore represents a lower bound on our system's true capability under the revised evaluation protocol.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2715–2722

adapted transformer framework can meaningfully improve performance on this task. To this end, we fine-tune COVID-Twitter-BERT v2 (CT-BERT-v2) (Müller et al., 2023), a domain-adapted BERT-large model pre-trained on COVID-19-related social media data, which is particularly well-suited for the informal and emotionally charged language prevalent in conspiracy-oriented Reddit discourse.

We propose a multi-task learning architecture with a dual-head design that simultaneously optimizes conspiracy detection and emotion classification, leveraging shared contextual representations from the `[CLS]` token. By incorporating emotion as an auxiliary signal, we aim to capture the affective patterns that underlie conspiratorial language, improving both model interpretability and F1 performance. Our results demonstrate that this joint modeling approach achieves an F1 score of 0.87 on the official development set, validating our hypothesis that emotional context provides a meaningful inductive bias for conspiracy detection. This work contributes a lightweight yet effective framework for conspiracy detection, underscoring the value of domain-adapted pre-training and auxiliary affective supervision in low-resource classification settings.

2 Related Work

The paradigm of text classification has shifted substantially from manual feature engineering to the use of language models pre-trained on vast corpora. Early transformer-based approaches, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b), established a standard where models were adapted to downstream tasks via a discriminative classification head. These models demonstrated that deep bidirectional representations could effectively capture semantic nuances necessary for tasks ranging from sentiment analysis to misinformation detection. Domain-adaptive pre-training has further extended this paradigm, showing that continued pre-training on in-domain corpora yields substantial gains over general-purpose models on specialized downstream tasks (Gururangan et al., 2020).

A prominent example of domain-adaptive pre-training for social media is CT-BERT-v2 (Müller et al., 2023), which has demonstrated strong performance on health-related misinformation and conspiracy detection tasks. Given that Reddit posts share similar linguistic properties with Twitter discourse, including colloquial language, implicit reasoning, and affective framing, CT-BERT-v2 serves as a natural foundation for the present task.

Prior work on conspiracy detection has largely relied on domain-specific pre-training and keyword-based approaches, focusing on identifying topical clusters or semantic embeddings associated with known conspiracy theories (Müller et al., 2023). While effective in controlled settings, these approaches often struggle to distinguish between conspiratorial and non-conspiratorial content when vocabulary overlap is high, for instance when users discuss conspiracy theories critically rather than endorsing them. Recent work has addressed this limitation by moving beyond surface-level lexical cues toward modeling the underlying narrative structure and rhetorical patterns of conspiratorial discourse (Uscinski et al., 2022). Furthermore, the emergence of large annotated datasets for conspiracy detection on social media platforms, including Reddit and Twitter, has enabled more fine-grained classification schemes that distinguish between endorsement, denial, and ambiguous stances (Samory et al., 2026).

Multi-task learning has emerged as an effective strategy for improving generalization in NLP by jointly optimizing related objectives within a shared representation (Liu et al., 2019a). Rather than training separate models for each task, multi-task frameworks allow auxiliary tasks to regularize the primary objective and surface complementary signals. In the context of social media analysis, joint modeling of sentiment or emotion alongside the primary classification target has been shown to improve performance, particularly when training data is limited (Felbo et al., 2017). This is especially relevant for conspiracy detection, where emotional tone is not merely incidental but is widely understood to be a defining feature of conspiratorial rhetoric (Uscinski et al., 2022).

The relationship between affect and conspiratorial belief has been widely documented in social psychology. Conspiracy beliefs are associated with perceived threat and anxiety, and commonly evoke fear and anger-related moral emotions. Moreover, reliance on emotional processing increases susceptibility to misinformation, suggesting emotion provides a meaningful inductive signal for detection models (Pummerer et al., 2024). Prior work on misinformation detection has leveraged sentiment and emotion features to augment textual representations, yielding improvements over text-only baselines (Ajao et al., 2019). Our work builds directly

on these findings by incorporating emotion classification as an auxiliary task within a multi-task fine-tuning framework, enabling the model to jointly learn the affective and semantic dimensions of conspiratorial language.

3 Data and Preprocessing

The official dataset was provided via Zenodo[2] and accessed through the starter scripts on GitHub[3].

3.1 Task Description

The SemEval 2026 Task 10 Subtask 2 focuses on detecting the presence of conspiracy thinking within social media discourse. Specifically, the goal is to determine whether a given Reddit submission statement expresses a conspiratorial belief (*Yes*), does not express one (*No*), or if the presence of conspiratorial content cannot be determined (*Can't tell*).

3.2 Data Split and Usage

Data splits. The dataset contains approximately 4,300 Reddit comments annotated for conspiracy presence.We used the script provided by the organizers to rehydrate the dataset. The training data contains labels that can be used for training, while the released `dev` and `test` sets remain unlabeled and are used only for inference and leaderboard submission. We constructed supervised splits for development by performing a stratified 90/10 train–validation split on the labeled data after class balancing to preserve the distribution across the three categories: *Yes*, *No*, and *Can't tell*.

Balancing data splits. Since the raw distribution was skewed toward non-conspiratorial posts, we applied random oversampling using `sklearn.utils.resample` (Pedregosa et al., 2011) to equalize the number of examples per class before splitting. This ensured that each label was represented with the same frequency in the training set, improving macro-F1 stability and preventing collapse toward the dominant class.

Preprocessing. Reddit comments were cleaned by lower-casing all text and removing URLs and redundant whitespace before being passed to the model. Each example was labeled with one of three conspiracy classes: `Yes`, `No`, or `Can't Tell`, encoded as integer labels 1, 0, and 2 respectively.

[2] https://zenodo.org/records/17065240

[3] https://github.com/hide-ous/semeval26_task10_starter-pack

Prompt Used for Paraphrase Augmentation

```
You are a data augmentation assistant.
Given a piece of text,
rewrite it as a natural paraphrase that
preserves the original
meaning and sentiment, but changes the
wording and phrasing.

Rules:
  - Keep the length roughly similar.
  - Do NOT add new information.
  - Do NOT remove important details.
  - Return ONLY the paraphrased text, no
explanations.

Original text:
{TEXT}

Paraphrased text:
```

Data Augmentation. To enrich the training set with both additional examples and emotion supervision, we employed GPT-5 (OpenAI, 2025) for two augmentation steps. First, we generated emotion annotations for all training examples using a chain-of-thought prompting strategy, where the model was instructed to reason over the text before assigning one of six canonical emotion labels: `Happy`, `Sad`, `Anger`, `Fear`, `Disgust`, and `Surprise`. Model outputs were parsed and normalized to these canonical labels. Second, we applied paraphrase-based augmentation by prompting GPT-5 using the template shown in Box 3.2 to generate a meaning-preserving rewrite for each of the approximately 4,300 original training examples, varying surface wording and phrasing while retaining the original sentiment, conspiracy label, and key semantic content. All original examples were retained unchanged alongside their paraphrased counterparts, effectively doubling the dataset to approximately 8,600 examples. This two-step augmentation strategy was motivated by the limited size of the training corpus and the high lexical overlap between conspiratorial and non-conspiratorial posts.

Hyperparameter Tuning. Text tokenization used the CT-BERT-v2 tokenizer (Müller et al., 2023) with a maximum input length of 128 tokens; longer sequences were truncated. The augmented dataset was split into 90% training and 10% validation sets. Fine-tuning used the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 8 for 3 epochs. The model was trained end-to-end without freezing any weights, optimizing a combined cross-entropy objective over both the

Table 1: Inter-annotator agreement on 150 sampled training examples. Agreement is reported as Cohen's κ and raw percentage agreement.

Comparison	Cohen's κ	% Agree
Annotator A vs. Annotator B	0.271	39.3
Annotator A vs. GPT-5	0.357	46.7
Annotator B vs. GPT-5	0.414	52.0
All three agree	—	28.0

conspiracy classification head and the emotion classification head.

3.3 Emotion Annotation Validation

To assess the reliability of the GPT-5-generated emotion labels used as auxiliary supervision, we conducted an inter-annotator agreement (IAA) study on a randomly sampled subset of 150 training examples. Two of the authors independently labeled each post with one of the six canonical emotion categories (`Happy`, `Sad`, `Anger`, `Fear`, `Disgust`, `Surprise`), blind to the GPT-5-generated labels. We report Cohen's κ for all pairwise comparisons, including each human annotator against GPT-5. Table 1 summarizes the results.

Agreement scores indicate fair pairwise alignment between annotators, consistent with prior findings that fine-grained categorical emotion classification is inherently difficult, particularly for short, context-poor social media texts (Demszky et al., 2020). Notably, GPT-5 agrees with each human annotator more strongly than the two human annotators agree with each other, suggesting that GPT-5 produces relatively central, regression-to-the-mean labels rather than capturing idiosyncratic human interpretations. Disagreements were concentrated between affectively adjacent categories such as *Anger* vs. *Disgust* and *Sad* vs. *Fear*, mirroring known confusions in the emotion-classification literature. Importantly, our framework treats the emotion head as a regularizing auxiliary signal rather than a source of gold-standard supervision, so moderate label noise is expected and does not undermine the primary classification objective.

4 System Description

In this work, we focus specifically on Subtask 2: conspiracy detection. Our objective is to determine whether a given Reddit post expresses a conspiratorial belief, does not express one, or if the presence of conspiratorial content cannot be determined. Accordingly, we adopt a three-way classification scheme with the labels *Yes*, *No*, and *Can't Tell*.

4.1 Model Selection

Prior to committing to our final architecture, we conducted a systematic comparison of encoder-based and zero-shot models on the official development set. As shown in Figure 1, encoder models generally outperformed zero-shot large language models on this task, with DistilRoBERTa achieving the highest F1 among baseline encoders at 0.66. Notably, GPT-5 in a zero-shot setting achieved only 0.45, suggesting that prompt-based inference alone is insufficient for this classification task. These findings motivated our selection of a domain-adapted encoder model fine-tuned on in-domain data rather than a generative or zero-shot approach.

4.2 Architecture

Our final system is built on CT-BERT-v2 (Müller et al., 2023), selected based on its domain alignment with the target corpus and its strong performance in our preliminary baseline comparison (Figure 1). Encoder-based models consistently outperformed zero-shot large language models on this task, and among encoder baselines, domain-adapted models proved particularly effective given the informal and emotionally charged nature of Reddit posts in the PsyCoMark dataset, offering a stronger inductive bias than general-purpose encoders such as RoBERTa or DeBERTa.

We fine-tuned CT-BERT-v2 with a multi-task dual-head architecture that jointly optimizes two objectives: (1) a conspiracy classification head that maps the `[CLS]` token representation to one of three labels (*Yes*, *No*, *Can't Tell*), and (2) an emotion classification head that simultaneously predicts one of six emotion categories (`Happy`, `Sad`, `Anger`, `Fear`, `Disgust`, `Surprise`). Both heads are implemented as linear layers operating on the shared `[CLS]` representation, with a combined cross-entropy loss optimized jointly during fine-tuning. This design is motivated by evidence that emotional tone is a defining feature of conspiratorial rhetoric (Uscinski et al., 2022), and that auxiliary affective supervision can improve performance in low-resource classification settings (Felbo et al., 2017).
We evaluated two input configurations:

1. **Text only**: the raw post text passed directly to CT-BERT-v2.

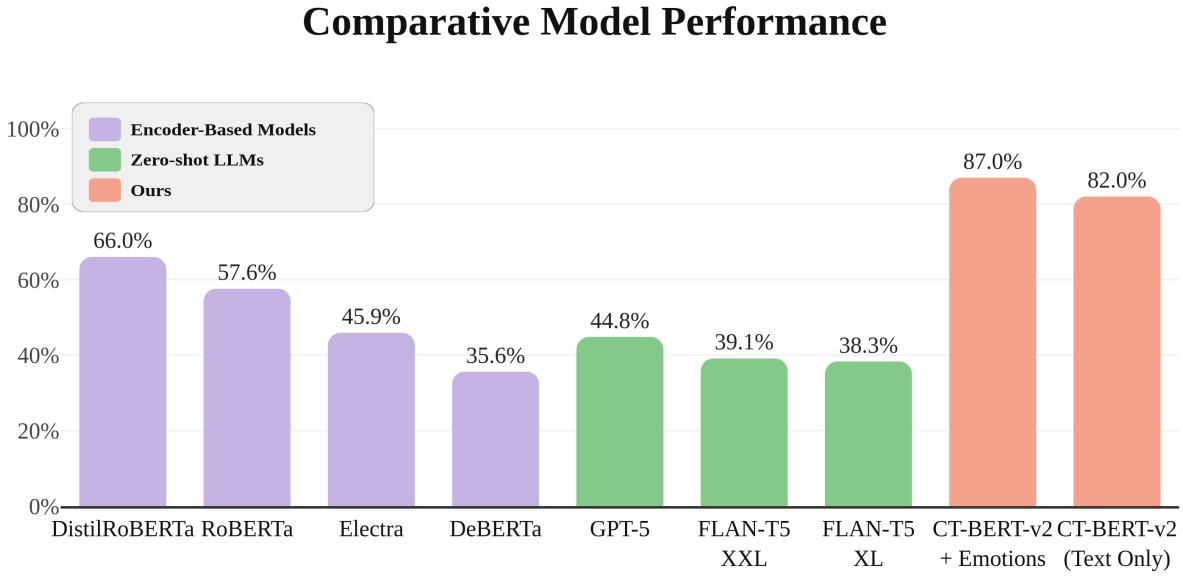

Figure 1: Comparative performance (F1 scores in %) of our proposed CT-BERT-v2 models against standard encoder-based baselines and zero-shot LLMs. Our emotion-aware configuration (87.0%) significantly outperforms all other evaluated models.

2. **Text + emotion annotations**: post text with GPT-5-generated chain-of-thought emotion labels used as auxiliary supervision signals during multi-task fine-tuning.

4.3 Training Setup

Fine-tuning was performed using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 2×10^{-5}, a batch size of 8, and 3 epochs of end-to-end training without weight freezing. Text tokenization used the CT-BERT-v2 tokenizer with a maximum input length of 128 tokens, beyond which sequences were truncated. The augmented dataset of approximately 8,600 examples was split into 90% training and 10% validation sets.

4.4 Evaluation

Each post is annotated with one of three labels indicating whether it expresses a conspiratorial belief: *Yes*, *No*, or *Can't Tell*. We evaluate model performance using the macro-averaged F1 score as the primary metric, consistent with the official task evaluation protocol, ensuring equal treatment of all three classes regardless of their frequency.

Overall Performance. Figure 1 presents the performance of all evaluated models on our internal validation split. Our text-only CT-BERT-v2 configuration achieves an F1 of **0.82**, substantially outperforming all baselines and demonstrating the effectiveness of domain-adapted pre-training. Incorporating emotion-aware multi-task learning further improves performance to **0.87**, a gain of five F1 points, confirming that GPT-5-generated emotion annotations capture discriminative affective signals that complement CT-BERT-v2's textual representations.[4]

4.5 Error Analysis

While our system achieves strong overall performance, a qualitative examination of model errors reveals consistent patterns that highlight the inherent difficulty of this task. The *Can't Tell* label represents the most challenging category for our model, as it encompasses posts that are ambiguous by definition, lacking the strong affective or lexical cues that characterize clearly conspiratorial or non-conspiratorial content. These posts often discuss conspiracy theories in a neutral, analytical, or ironic register, making it difficult for the model to distinguish genuine endorsement from skeptical engagement.

A second source of error arises from posts where conspiratorial framing is implicit rather than ex-

[4]All baseline and internal evaluation scores reported in this paper (including Figure 1) were computed using the original Codabench evaluation pipeline prior to a post-submission revision to the official evaluation protocol. Our locally evaluated macro-F1 on the public development set is 0.54, which we consider the more reliable estimate of system performance.

plicit. In such cases, the conspiratorial belief is embedded in presuppositions or indirect references rather than direct assertions, which challenges the model's ability to detect the underlying stance from surface-level representations alone. Similarly, posts that employ sarcasm or rhetorical questions may superficially resemble conspiratorial content while conveying the opposite intent, contributing to false positive predictions.

These observations suggest that future work should focus on improving the modeling of pragmatic and discourse-level features, particularly for the *Can't Tell* category. Incorporating conversational context, such as the surrounding thread or subreddit norms, may help resolve ambiguous cases that are difficult to classify from the post text alone.

5 Conclusion

This work demonstrates that a domain-adapted encoder model with emotion-aware multi-task learning can effectively detect conspiratorial content in social media text. We fine-tuned CT-BERT-v2 within a dual-head architecture that jointly optimizes conspiracy classification and emotion prediction, augmenting the training set through GPT-5-generated paraphrases and chain-of-thought emotion annotations. The *text + emotions* configuration achieves the strongest performance across all evaluated setups, reaching an F1 score of **0.54** on the official evaluation set, outperforming the text-only baseline by a meaningful margin. This finding aligns with psycholinguistic evidence that conspiratorial narratives are characterized by distinctive affective patterns, particularly elevated fear, anger, and moral outrage, and suggests that supervising models on these emotional dimensions enables them to capture deeper rhetorical cues beyond surface-level lexical overlap. Promising directions for future work include exploring richer affective representations beyond six-class emotion labels, improving model handling of ambiguous *Can't Tell* instances, and investigating the transferability of emotion-aware multi-task frameworks to broader misinformation detection settings.

6 Limitations

While our system achieves strong performance, several limitations warrant discussion. Our reliance on the [CLS] token representation limits the model's ability to handle implicit conspiratorial framing, presuppositions, and sarcasm, where surface affective patterns may mislead rather than inform the primary classifier. The auxiliary emotion supervision is derived from GPT-5 annotations rather than gold human labels, and while our IAA study suggests reasonable agreement, residual label noise may attenuate the multi-task signal. Our system addresses only Subtask 2 and is evaluated on English-language COVID-era Reddit data, leaving generalizability to other conspiracy domains, platforms, and languages untested. A fuller discussion of these limitations appears in Appendix A.

Acknowledgments

We thank the PsyCoMark organizers for releasing the dataset and evaluation platform, and for their support throughout the shared task. We also thank the anonymous reviewers for their constructive feedback.

References

Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment aware fake news detection on online social networks. In *ICASSP 2019 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2507–2511. IEEE.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Karen M Douglas and Robbie M Sutton. 2023. What are conspiracy theories? a definitional approach to their correlates, consequences, and communication. *Annual review of psychology*, 74(1):271–298.

Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A

Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Pavan Holur, Tianyi Wang, Shadi Shahsavari, Timothy Tangherlini, and Vwani Roychowdhury. 2022. Which side are you on? insider-outsider classification in conspiracy-theoretic social media. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4975–4987.

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.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*.

Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, and Anton van den Hengel. 2025. Deep learning for hate speech detection: a comparative study. *International Journal of Data Science and Analytics*, 20(4):3053–3068.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Martin Müller, Marcel Salathé, and Per E. Kummervold. 2023. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence*, 6:1023281.

OpenAI. 2025. GPT-5 system card. Technical report, OpenAI.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Lotte Pummerer, Theofilos Gkinopoulos, Karen M. Douglas, Daniel Jolley, and Kai Sassenberg. 2024. The appraisal model of conspiracy theories: Applying appraisal theories to understand emotional and behavioral reactions to conspiracy theories. *Psychological Inquiry*, 35(3-4):159–178.

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)*.

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.

Joseph Uscinski, Adam Enders, Amanda Diekman, John Funchion, Casey Klofstad, Sandra Kuebler, Manohar Murthi, Kamal Premaratne, Michelle Seelig, Daniel Verdear, and 1 others. 2022. The psychological and political correlates of conspiracy theory beliefs. *Scientific reports*, 12(1):21672.

Jan-Willem Van Prooijen and Karen M Douglas. 2017. Conspiracy theories as part of history: The role of societal crisis situations. *Memory studies*, 10(3):323–333.

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

A Limitations

Implicit conspiratorial framing. Our model relies primarily on the [CLS] token representation as input to both classification heads. While this design captures sentence-level affective and lexical signals effectively, it provides limited support for pragmatic or discourse-level reasoning. As a result, posts that express conspiratorial beliefs through implicit framing—presuppositions, indirect reference, or rhetorical structure—are systematically harder to classify than posts containing explicit assertions. For example, a post that presupposes the existence of a coordinated cover-up without ever stating one directly conveys a conspiratorial stance that our model frequently fails to detect from surface representations alone. Addressing this would likely require explicit modeling of presupposition and stance, or richer pooling strategies that aggregate evidence across multiple token positions rather than collapsing to a single [CLS] vector.

Sarcasm and rhetorical questions. A related but distinct challenge arises with sarcasm and rhetorical questions, which produce surface lexical and affective patterns closely resembling those of sincere conspiratorial assertions while conveying the opposite intent. Because our auxiliary emotion supervision is itself derived from surface text, it may reinforce rather than resolve this confusion: a sarcastically angry post and a sincerely angry conspiratorial post can receive similar emotion labels, offering little discriminative signal to the primary head. Robust handling of these cases likely requires conversational context (e.g., parent comments, thread structure, or subreddit norms) or explicit irony detection, neither of which our current architecture incorporates.

Reliance on automatically generated emotion labels. The auxiliary emotion supervision used in our multi-task framework is derived from GPT-5 chain-of-thought annotations rather than gold human labels. While our validation study (§3.3) suggests reasonable agreement between GPT-5 outputs and human judgments, residual label noise inevitably propagates into the auxiliary head and may attenuate the benefit of the multi-task signal. Disagreements were concentrated between affectively adjacent categories (e.g., Anger vs. Disgust, Fear vs. Sad), reflecting known difficulties in fine-grained categorical emotion classification. The auxiliary head serves primarily as a regularizer rather than a source of gold-standard emotion predictions. A more principled treatment, such as using soft label distributions over emotion categories or richer affective representations like continuous Valence-Arousal-Dominance dimensions (Mohammad, 2018), may yield further gains.

Scope of the system. Our work addresses Subtask 2 (binary conspiracy detection) and does not engage with Subtask 1 (psycholinguistic marker extraction), which requires a sequence-labeling architecture beyond the dual-head sentence classifier presented here. A unified system that jointly performs detection and marker extraction remains an open direction, particularly given that marker spans (Actors, Actions, Effects, Victims) plausibly carry information complementary to global affective signals.

Domain and platform specificity. Both our base encoder (CT-BERT-v2) and the PsyCoMark dataset itself are strongly oriented toward English-language, COVID-era social media discourse. The extent to which emotion-aware multi-task fine-tuning generalizes to other conspiracy domains (e.g., political conspiracies, climate denial), other platforms with different discourse norms, or other languages remains untested. The strong performance of CT-BERT-v2 on this task may partly reflect a fortunate overlap between its pre-training distribution and the dataset, and findings should be interpreted accordingly.

Inference-time threshold dependence. Our reported development F1 reflects a decision threshold ($\tau^* = 0.23$) selected post-hoc on the development split. While the F1 surface is relatively flat near the optimum, the chosen threshold is calibrated to a specific class distribution and may not transfer cleanly to deployment settings where the prevalence of conspiratorial content differs from the PsyCoMark training distribution.