dangphuduy at SemEval-2026 Task 10: Span-based Conspiracy Marker Extraction and Emotion-Aware Detection via Gated Fusion

Duy Dang Phu and **Son Bui Hong** and **Dang Van Thin**
University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam
Vietnam National University HoChiMinh City, Vietnam
{24520010,22521246}@gm.uit.edu.vn, thindv@uit.edu.vn

Abstract

Conspiracy theories on social media pose significant societal risks, making it essential to detect both conspiracy-related content and the textual spans that serve as conspiracy markers. In this work, we propose two effective methods to address these challenges. For marker extraction, we develop a span-based sliding window framework that improves efficiency and accuracy by focusing on localized context. In addition, inspired by the distinctive emotional patterns in conspiracy texts, we design a dynamic gating mechanism to integrate emotional and semantic representations. We evaluate our methods on the SemEval 2026 Task 10, where our team (dangphuduy) achieved competitive results, ranking 4th in Task 1 (Span Extraction) and 3rd in Task 2 (Conspiracy Detection). Experimental results demonstrate that both proposed methods significantly enhance model performance.

1 Introduction

Online conspiracy theories pose a growing challenge due to their rapid spread on social media and their role in amplifying misinformation (Shahsavari et al., 2020; Galende et al., 2022). Addressing this problem requires both identifying conspiracy-related content and extracting the linguistic signals that characterize conspiratorial thinking. SemEval 2026 Task 10 (Samory et al., 2026; Ghosh et al., 2026) targets this issue through two subtasks: *Conspiracy Marker Extraction*, which aims to identify textual spans expressing core conspiracy markers grounded in evolutionary psychology, and *Conspiracy Detection*, which classifies Reddit comments as conspiracy-related or not.

Using SemEval 2026 Task 10 as our evaluation benchmark, we propose task-specific modeling strategies for both subtasks. For marker extraction, we introduce a sliding window span-based framework that focuses on effectively capturing the surrounding contextual information of candidate spans. For conspiracy detection, we develop a binary classification model that dynamically fuses semantic and emotional representations through a gating mechanism, motivated by the observation that conspiracy-related texts exhibit distinctive emotional patterns.

2 Background

This section provides the conceptual and research context for our study. We first introduce the problem setting and clarify the objectives of the PsyCoMark task. We then review prior work related to conspiracy discourse analysis.

2.1 Task Description

The PsyCoMark task aims to foster the development of interpretable and generalizable models for conspiracy theory understanding, grounded in evolutionary psychology. It is composed of two synergistic subtasks that can be addressed independently or jointly.

Subtask 1: Conspiracy Marker Extraction focuses on identifying text spans that express core psycholinguistic markers of conspiracy thinking. Each document may contain zero, one, or multiple markers, which are allowed to overlap or be nested. The predefined marker categories include: *Actor* (individuals or groups alleged to hold malicious intent), *Action* (alleged actions or plans carried out by the actors), *Effect* (negative consequences resulting from these actions), *Victim* (entities harmed by the conspiracy), and *Evidence* (arguments, claims, or reasoning used to support the conspiracy narrative). Performance is evaluated using an overlap-based macro-averaged F1-score for each marker type.

Subtask 2: Conspiracy Detection is a document-level classification task that requires determining whether a Reddit comment is conspiracy-related or not. Although this subtask can be solved independently, it is designed to

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 793–802

benefit from the explicit modeling of conspiracy markers extracted in Subtask 1. Evaluation is conducted using macro-averaged F1-score.

Overall, PsyCoMark encourages models that not only achieve high predictive performance but also provide fine-grained, interpretable insights into the linguistic and psychological structure of conspiracy discourse.

2.2 Related Work

Conspiracy Detection and Analysis Recent research in Natural Language Processing has increasingly focused on the automatic detection of conspiracy-related content in news and social media (Shahsavari et al., 2020; Galende et al., 2022). The PAN@CLEF 2024 shared task (Korenčić et al., 2024) represents a notable shift by emphasizing not only conspiracy classification from a critical perspective but also the identification of standard narrative elements within conspiracy discourse. This shift highlights the need for more fine-grained and structured analytical approaches.

Span-based Modeling for Conspiracy Markers Span-based modeling has been widely adopted in Named Entity Recognition (NER) and has demonstrated strong effectiveness in handling nested, overlapping, and complex spans (Fu et al., 2021; Yu et al., 2022). These advantages make span-based methods particularly suitable for capturing fine-grained textual phenomena in challenging extraction tasks. However, despite their success in NER, span-based approaches remain largely underexplored for conspiracy marker extraction, leaving their potential in this domain insufficiently investigated.

Emotion Signals in Conspiracy Narratives Although emotional and psycholinguistic cues have been shown to improve performance in related tasks such as fake news and misinformation detection (Zhang et al., 2021), their role in conspiracy analysis has received limited attention (Liu et al., 2024). Given that conspiracy-related discourse particularly on social media often exhibits distinctive emotional patterns, incorporating emotion-aware representations constitutes a promising direction.

In this work, we address these gaps by introducing a span-based framework for fine-grained marker modeling and a dynamic emotion-aware mechanism to enhance robustness in emotionally charged narratives.

3 System Overview

3.1 Subtask 1

We adopt a span-based modeling framework inspired by span-based named entity recognition (Yu et al., 2022). Given an input sequence, it is first encoded into contextualized token representations using a pretrained RoBERTa (Liu et al., 2019) encoder. Based on these representations, each candidate span is modeled by combining its boundary information with global and structural features. Formally, for a span $s = (i, j)$, its representation is defined as:

$$\mathbf{v}_s = \left[\mathbf{h}_i; \mathbf{h}_j; \mathbf{h}_{\text{CLS}}; \mathbf{e}_{(j-i+1)}\right]. \quad (1)$$

where $\mathbf{h}_i$ and $\mathbf{h}_j$ are the contextualized hidden states of the span's start and end tokens, $\mathbf{h}_{\text{CLS}}$ denotes the global sequence representation, and $\mathbf{e}_{(j-i+1)}$ is a learned embedding that encodes the span width. We do not apply additional pooling operations (e.g., max pooling or mean pooling) over the span tokens, as such operations would significantly increase computational and memory costs under practical hardware constraints.

To handle long sequences, we employ a sliding window approach that splits the input into overlapping chunks, which are encoded independently and later aggregated for final prediction. This design allows the span-based model to attend to the most relevant local context around each candidate span. The window size is a critical hyperparameter: overly short windows may break semantic coherence and limit contextual coverage, while excessively long windows can dilute salient information and introduce noise. Hence, an appropriate window length is essential to balance contextual completeness and focus.

For spans that appear in multiple overlapping windows, we resolve prediction conflicts by selecting the maximum score across all occurrences of the span. Formally, let s be a candidate span that appears in multiple windows $\{w_1, w_2, \ldots, w_k\}$, and let $p(s \mid w_t)$ denote the predicted score for span s in window w_t. The final score is computed as:

$$p(s) = \max_{t \in \{1,\ldots,k\}} p(s \mid w_t). \quad (2)$$

This strategy prioritizes the most confident prediction for each span while mitigating boundary effects introduced by window segmentation, and avoids score dilution that may arise from averaging across heterogeneous contexts.

3.2 Subtask 2 - Emotion-Aware Module

To capture emotional cues in conspiracy discourse, we use `SamLowe/roberta-base-go_emotions`, a RoBERTa-based classifier fine-tuned on GoEmotions (Hugging Face Model Hub, 2024; Demszky et al., 2020), as a fixed emotion feature extractor. The model outputs probability scores over 28 emotion categories, which are incorporated as auxiliary inputs, while keeping its parameters frozen to preserve the intrinsic emotional representations and prevent task-specific fine-tuning from distorting the original affective semantics.

We first validate the discriminative power of these features. As reported in Appendix A, empirical analysis reveals clear divergences: emotions such as *curiosity*, *annoyance*, *amusement*, and *surprise* are prevalent in conspiracy content, whereas *approval*, *joy*, and *admiration* characterize non-conspiracy texts. Consequently, we retain only a subset of the most discriminative emotions to construct the input emotion vector.

To ensure numerical stability and mitigate the impact of outliers, the raw scores of the selected emotions are normalized using a quantile-based Min-Max scaling strategy. Specifically, values are scaled based on the $(q_{0.05}, q_{0.95})$ quantiles and clipped to the $[0, 1]$ range, yielding the normalized emotion vector E_{raw}.

Inspired by the gating mechanism in Gated Recurrent Units (GRUs) (Cho et al., 2014), which balances information flow, we design a streamlined fusion strategy to combine semantic and emotional representations. First, E_{raw} is projected into the same latent space as the RoBERTa `[CLS]` representation H via a linear transformation and ReLU activation:

$$E_{proj} = \text{ReLU}(W_p E_{raw} + b_p) \tag{3}$$

To effectively fuse these modalities, we concatenate them and employ a sigmoid-based gating function to compute a scalar fusion coefficient z:

$$z = \sigma(W_g[H; E_{proj}] + b_g) \tag{4}$$

where $[\,;\,]$ denotes vector concatenation. The final fused representation is obtained via element-wise interpolation:

$$V_{fusion} = z \odot H + (1 - z) \odot E_{proj} \tag{5}$$

This mechanism enables the model to adaptively regulate the influence of emotional cues relative to semantic context for each specific instance. In particular, a lower value of z biases the fusion process toward emotional representations, which is especially effective for capturing emotionally charged social media discourse, such as anger-driven or provocative comments. Conversely, a higher value of z places greater emphasis on semantic features, allowing the model to focus on analytically structured and seemingly neutral narratives that may nevertheless convey implicit conspiratorial intent. Unlike the standard GRU which models temporal dependencies, our formulation adopts a non-recurrent, feed-forward gating design tailored for static representation fusion. Finally, V_{fusion} is fed into a dense classification layer to estimate the conspiracy probability $y \in [0, 1]$.

4 Experimental Setup

For both subtasks, we merge the official training and development sets provided by the organizers and randomly split the combined data into 80% for training and 20% for validation. The original development set is publicly available and intended for model development, rather than serving as a hidden test set. This re-splitting strategy allows us to maximize the use of available data while maintaining a validation set for model selection.

4.1 Subtask 1

For Subtask 1, we use *RoBERTa-large* (Liu et al., 2019) as the backbone encoder. To handle long sequences, we apply a sliding window strategy with a chunk length of 64 and a stride of 32. The maximum span width is set to 20. Each span is represented by concatenating the embeddings of its start token, end token, and the [CLS] token.

The model is trained using focal loss with $\alpha = 0.8$ and $\gamma = 2.0$ to mitigate class imbalance. We optimize the model using AdamW with a learning rate of 2×10^{-5}, a batch size of 32, and train for 5 epochs. All experiments are conducted on a single GPU.

Additionally, we employ an ensemble strategy by training the model independently seven times with different random seeds. At inference time, we average the prediction scores of the seven models for each candidate span, and retain spans whose averaged scores exceed a threshold of 0.45.

4.2 Subtask 2

For Subtask 2, we fine-tune *RoBERTa-large* as the backbone model. The training is conducted with a batch size of 16, a learning rate of 2×10^{-6}, and the *BCEWithLogitsLoss* objective for 3 epochs.

To enhance model robustness, we adopt a prompt-based few-shot data augmentation strategy, where three randomly sampled training examples are used to construct conspiracy and non-conspiracy prompts for generating additional samples. We further apply ensemble learning by training 12 models with different random seeds and aggregating their predictions via majority voting to improve prediction stability and generalization.

Vanilla RoBERTa Baseline To provide a standard point of comparison, we include a vanilla *RoBERTa-large* baseline without emotion features, data augmentation, or ensemble techniques. The model is fine-tuned using the same training configuration as above, but only on the original training data. This baseline serves to isolate the contribution of the proposed emotion-aware gating mechanism and ensures a fair comparison with standard approaches.

5 Results

5.1 Subtask 1

In Subtask 1, we investigate the effectiveness of the sliding window strategy under different context lengths. As reported in Table 2, increasing the context length from 64 to 128 leads to a noticeable degradation in performance, while a further increase to 256 results in out-of-memory (OOM) errors. These results suggest that selecting an appropriate context length can simultaneously improve model performance and reduce computational cost. In our experiments, a context length of 64 achieves a favorable balance between accuracy and resource efficiency.

Furthermore, we evaluate our system under different ensemble sizes and decision thresholds. As shown in Table 2, increasing the ensemble size consistently improves precision while reducing recall. This trend can be attributed to the variability in span predictions among models trained with different random seeds, whereby only spans favored by multiple models achieve sufficiently high averaged scores to be retained as candidates. However, since span detection is inherently more challenging than binary classification, achieving agreement across models is difficult. As a result, larger ensembles produce more conservative predictions, leading to higher precision but lower recall. To alleviate this issue, we reduce the decision threshold to 0.45 when performing inference with larger ensemble sizes.

Table 1: Performance under different context lengths on Subtask 1 evaluated on the test set at IoU 0.5.

Context Len	P	R	$F1_{Mi}$	$F1_{Ma}$
64	23.34	29.82	26.18	23.62
128	22.12	26.07	23.93	21.92
256	Out of Memory (OOM)			

Table 2: Effect of ensemble size on Subtask 1 test set performance in terms of macro F1 at IoU 0.5.

Ens	Thres	P	R	$F1_{Mi}$	$F1_{Ma}$
1	0.5	23.34	29.82	26.18	23.62
5	0.45	23.91	29.80	26.53	24.24
7	0.45	25.45	28.08	26.70	24.19
10	0.45	26.24	27.53	26.87	24.00
10	0.5	32.57	20.16	24.91	21.08

As shown in Table 3, the *Actor* and *Victim* categories achieve better performance than the other argument types, mainly due to their shorter average span lengths (13.02 and 12.51), which facilitate more accurate boundary detection, and their relatively consistent entity-centered linguistic patterns. In contrast, the remaining categories typically involve longer and more complex spans with diverse semantic information, making extraction more challenging. Notably, *Effect* shows the lowest performance: although both *Effect* and *Action* depend on the *Actor*, *Action* is usually realized through explicit subject–verb structures, whereas *Effect* is often expressed implicitly and lacks stable syntactic patterns. Meanwhile, *Evidence* relies more on context-dependent expressions and is less constrained by such dependencies. Overall, these performance differences are consistent with the intrinsic characteristics of each argument type.

Overall, our proposed span-based sliding window framework, combined with an optimized ensemble strategy, achieved 4th place in the official SemEval 2026 Task 10 - Subtask 1 leaderboard. This competitive ranking underscores the robustness of our localized context approach in handling complex conspiracy markers.

Table 3: Per-type statistics (Cnt, Len) computed on the combined training and development sets, and per-type performance ($F1_{Ma}$) on Subtask 1 evaluated on the test set at IoU 0.5.

Type	Cnt	Len (mean±std)	$F1_{Ma}$
Action	4945	26.96 ± 30.23	18.63
Actor	6552	13.02 ± 19.47	38.51
Effect	3810	35.48 ± 38.16	14.92
Evidence	3727	35.41 ± 52.99	18.88
Victim	3387	12.51 ± 14.71	30.28
Agg.	-	-	24.24

5.2 Subtask 2

In Subtask 2, we assess the effectiveness of the proposed *dynamic gating mechanism* by comparing it against the base model and several fixed-gate variants. Specifically, we investigate a range of constant gating coefficients z to analyze the contribution of emotion and semantic representations under different fusion ratios. The corresponding results are presented in Table 4. Our findings indicate that incorporating a fixed gating coefficient consistently enhances performance relative to the vanilla RoBERTa baseline. Notably, setting $z = 0.3$ improves the weighted F_1 score from 75.52 to 76.20, demonstrating that explicitly integrating emotional and semantic features yields tangible benefits. Moreover, enabling the model to dynamically learn the gating coefficient further boosts performance, achieving the best weighted F_1 score of 76.38. This result suggests that adaptive, instance-level modulation of feature contributions facilitates more flexible and effective representation fusion, leading to superior predictive capability.

We further analyze the effect of data augmentation size on model performance. Moderate augmentation yields slight improvements, with the best performance achieved at an augmentation size of 850, while excessive augmentation leads to degradation due to increased noise, indicating that overly aggressive augmentation may hinder model generalization. Detailed results are reported in Appendix B.

To further boost robustness, we employ ensemble learning by aggregating models trained with different random seeds. As shown in Table 7, ensemble methods consistently surpass single-model baselines, with the performance peaking at an ensemble size of 12. This configuration achieved our best weighted F_1 score of 77.68, securing 3rd place in the official Subtask 2 leaderboard. We observe that including more than 12 models slightly degrades performance, suggesting that the introduction of weaker models can negatively affect the overall ensemble quality. Further details are provided in Appendix C.

Table 4: Ablation study of the gating coefficient z on Subtask 2 evaluated on the test set. All models are trained without data augmentation or ensemble methods.

Model variant	$F1_{No}$	$F1_{Yes}$	$F1_{weighted}$
Vanilla RoBERTa	77.84	72.81	75.52
Fixed gate $z = 0.3$	78.20	73.86	76.20
Fixed gate $z = 0.5$	75.59	73.41	74.59
Fixed gate $z = 0.7$	76.73	74.59	75.75
Dynamic gating	77.66	74.90	**76.38**

To gain further insights into model errors, we analyze the distributions of emotion scores for samples misclassified by models trained with three different random seeds. Specifically, we compare the emotional characteristics of incorrectly predicted test instances with those of the full training set. The detailed distribution patterns are reported in Appendix D. Overall, the misclassified samples exhibit markedly different emotion distribution patterns between conspiracy and non-conspiracy classes. This observation suggests that unstable or highly variable emotional signals introduce additional uncertainty into the prediction process, thereby contributing to performance degradation.

6 Conclusion

In this paper, we propose a sliding-window strategy to enhance span-based conspiracy marker extraction, along with a gating mechanism to effectively fuse emotional and semantic representations. Experimental results demonstrate that the sliding-window approach significantly improves model performance by enabling better utilization of local contextual information while simultaneously reducing memory consumption. Moreover, incorporating emotional features proves beneficial for binary conspiracy classification, leading to consistent performance gains. The proposed dynamic gating mechanism further allows the model to adaptively balance the contributions of emotion and semantic representations, yielding additional improvements. The effectiveness of our approach is validated by our performance in the SemEval 2026 Task 10 com-

petition, where our team (dangphuduy) achieved a ranking of 3rd in Task 2 (Conspiracy Detection) and 4th in Task 1 (Span Extraction). These results underscore the robustness and state-of-the-art potential of our framework in tackling the complexities of conspiracy-related content.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

Limitations and Future Work

For Subtask 1, a key limitation of our approach lies in its reduced robustness when handling long argument spans with highly diverse and complex structural patterns. This indicates that modeling based mainly on local contextual information has not yet been able to adequately address robust boundary detection and effective modeling of long-range dependencies. Future work could investigate a broader range of potential strategies to address these challenges, such as incorporating structured linguistic knowledge or adopting hierarchical extraction mechanisms, among other possible directions.

For Subtask 2, although a dynamic gating mechanism is employed to fuse emotional and semantic representations, the model may struggle under distributional shifts in emotional signals between conspiracy and non-conspiracy instances, which can lead to degraded detection performance. To address this limitation, future work could incorporate mechanisms for identifying such distributional anomalies and proactively adjusting the fusion strategy by prioritizing semantic features, or even relying solely on semantic representations, rather than depending entirely on the dynamic gating mechanism.

In addition, we do not extensively explore auxiliary techniques such as ensemble learning, data augmentation, alternative span representation strategies, or different emotion extraction models, all of which could potentially further improve performance. Investigating these complementary directions is left for future work.

Ethics Statement

This study utilizes the official dataset provided by SemEval-2026 Task 10, strictly adhering to user privacy and data usage regulations. The proposed model is intended to assist in the analysis of conspiracy narratives for research purposes, not to enable automated censorship or infringe on free speech. We acknowledge the potential for biases and false positives; therefore, the system should be used with human-in-the-loop oversight rather than as a standalone decision-maker for content moderation.

References

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Jinlan Fu, Xuan-Jing Huang, and Pengfei Liu. 2021. Spanner: Named entity re-/recognition as span prediction. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long papers)*, pages 7183–7195.

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

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.

Hugging Face Model Hub. 2024. Samlowe/roberta-base-go_emotions. `https://huggingface.co/SamLowe/roberta-base-go_emotions`.

Damir Korenčić, Berta Chulvi, X Bonet Casals, Mariona Taulé, Paolo Rosso, and Francisco Rangel. 2024. Overview of the oppositional thinking analysis pan task at clef 2024. *Working Notes of CLEF*.

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

Zhiwei Liu, Boyang Liu, Paul Thompson, Kailai Yang, and Sophia Ananiadou. 2024. Conspemollm: Conspiracy theory detection using an emotion-based large language model. *arXiv preprint arXiv:2403.06765*.

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)*.

Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R Tangherlini, and Vwani Roychowdhury. 2020. Conspiracy in the time of corona: automatic detection of emerging covid-19 conspiracy theories in social media and the news. *Journal of computational social science*, 3(2):279–317.

Jie Yu, Bin Ji, Shasha Li, Jun Ma, Huijun Liu, and Hao Xu. 2022. S-ner: A concise and efficient span-based model for named entity recognition. *Sensors*, 22(8):2852.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476.

A Detailed Emotion Score Analysis

Table 5 presents the mean probability scores for all 28 emotion categories across conspiracy-related and non-conspiracy comments in the merged training and development data.

As discussed in Section 5.2, noticeable differences are observed in emotions such as *curiosity* and *approval*, whereas most other emotions exhibit relatively small variations.

Overall, the distribution is dominated by the *neutral* category, with mean probabilities exceeding 0.57 in both classes, indicating that fine-grained emotional signals are generally subtle and sparse.

B Effect of Data Augmentation Size

Table 6 reports the impact of varying augmentation sizes on Subtask 2 performance. We observe marginal improvements with moderate augmentation, with the highest weighted F_1 score achieved at an augmentation size of 850. However, further increasing the augmentation size to 980 slightly degrades performance, likely due to the introduction of additional noise in the augmented data.

C Effect of Ensemble Size

Table 7 presents the effect of ensemble size on Subtask 2 performance. Ensemble learning consistently improves performance over a single model, with performance peaking at an ensemble size of 12, achieving a weighted F_1 score of 77.68. Increasing the ensemble size to 15 leads to a slight performance drop, suggesting diminishing returns from additional models.

Table 5: Full comparison of mean emotion scores across all 28 categories on the original training and development sets released by the organizers after the development phase.

Emotion	Mean_{yes}	Mean_{no}	Diff_{yes-no}
curiosity	0.0857	0.0669	0.0187
approval	0.0904	0.1063	-0.0159
joy	0.0086	0.0201	-0.0115
annoyance	0.0434	0.0319	0.0115
admiration	0.0369	0.0452	-0.0083
amusement	0.0194	0.0124	0.0070
surprise	0.0160	0.0094	0.0066
gratitude	0.0092	0.0151	-0.0059
confusion	0.0588	0.0539	0.0049
fear	0.0133	0.0086	0.0047
disappointment	0.0250	0.0293	-0.0043
sadness	0.0137	0.0180	-0.0042
realization	0.0396	0.0439	-0.0042
z‘ remorse	0.0038	0.0073	-0.0035
anger	0.0096	0.0064	0.0033
desire	0.0074	0.0101	-0.0028
optimism	0.0282	0.0305	-0.0023
excitement	0.0063	0.0084	-0.0021
neutral	0.5730	0.5709	0.0021
disgust	0.0080	0.0065	0.0015
caring	0.0089	0.0102	-0.0013
love	0.0078	0.0089	-0.0011
disapproval	0.0312	0.0323	-0.0011
relief	0.0017	0.0024	-0.0007
nervousness	0.0026	0.0031	-0.0005
pride	0.0010	0.0014	-0.0004
grief	0.0009	0.0010	-0.0002
embarrassment	0.0030	0.0029	0.0001

Table 6: Effect of data augmentation size on model performance onSubtask 2 evaluated on the test set.

Aug. size	$\mathbf{F1}_{\text{No}}$	$\mathbf{F1}_{\text{Yes}}$	$\mathbf{F1}_{\text{weighted}}$
0	77.66	74.90	76.38
850	77.14	75.80	76.52
980	77.56	74.73	76.25

D Emotion Distribution Analysis on Misclassified Samples

We analyze the differences in emotion score distributions between the original data and misclassified samples. Specifically, we compare (i) the full merged dataset (original training data combined with the development set provided by the organizers) and (ii) two sets of misclassified samples obtained from a validation split constructed by randomly sampling 20% of the merged data.

For each emotion, we compute the score differ-

Table 7: Effect of ensemble size on Subtask 2 performance evaluated on the test set.

Ensemble size	$\mathbf{F_1}_{\text{No}}$	$\mathbf{F_1}_{\text{Yes}}$	$\mathbf{F_1}_{\text{weighted}}$
1	77.14	75.80	76.52
5	77.97	75.95	77.04
8	78.57	76.36	77.55
12	78.72	76.46	**77.68**
15	78.33	76.09	77.29

ence between label *yes* and label *no*. Diff_{yes-no} denotes the difference on the full dataset, while Diff v1 and Diff v2 correspond to two independent training runs evaluated on the validation split. It is important to note that the emotions reported in this table correspond exactly to the subset of emotion categories selected as input features for our final model.

The observed distribution shifts indicate that misclassified samples exhibit substantially different emotional patterns compared to the overall dataset, suggesting that unstable or conflicting emotional signals may contribute to prediction errors.

Table 8: Emotion score difference (Label yes minus Label no): full dataset vs. wrong predictions on the validation split (20% of merged train+dev) across two runs

Emotion	Diff_{yes-no}	Diff v1	Diff v2
neutral	0.0021	0.0986	0.0954
approval	-0.0159	0.0483	0.0507
annoyance	0.0115	-0.1182	-0.0995
admiration	-0.0083	-0.0285	-0.0341
realization	-0.0042	-0.0056	0.0007
excitement	-0.0021	-0.0082	-0.0111
disappointment	-0.0043	-0.0213	0.0135
anger	0.0033	-0.0925	-0.0865
joy	-0.0115	0.0225	0.0081
curiosity	0.0187	-0.0780	-0.0783
amusement	0.0070	-0.0036	-0.0226
confusion	0.0049	-0.1252	-0.0975
optimism	-0.0023	0.0101	0.0153
sadness	-0.0042	0.0097	0.0052
fear	0.0047	-0.0195	-0.0071
desire	-0.0028	-0.0127	0.0033
surprise	0.0066	-0.0348	-0.0354
gratitude	-0.0059	-0.0096	-0.0116
remorse	-0.0035	-0.0324	-0.0076

E Data Augmentation Details

We employ a large language model, `openai/gpt-oss-120b`, accessed via the Groq API, to generate augmented training samples. We use a temperature of 0.3 to balance diversity and stability in generation.

Few-shot Sampling For each augmentation instance, we randomly sample three sentences from the training set as in-context examples. These examples are used to guide the generation process while encouraging diversity across augmented samples.

Conspiracy Prompt

System: You are an expert in linguistics, NLP, and the study of conspiracy narratives. You understand how conspiracy beliefs are expressed in natural language, including themes such as hidden power, secret agendas, manipulation, and distrust of authorities.

User: Given three example sentences related to conspiracy theories, generate one new sentence that:

- Clearly reflects a conspiracy belief
- Is semantically different from the examples
- Does not reuse entities, phrases, or sentence structures
- Sounds natural and realistic
- Mentions hidden control, secret coordination, or powerful unseen actors
- Has a length roughly comparable to the examples

Examples:

1. `<example_1>`
2. `<example_2>`
3. `<example_3>`

Output: Return exactly one sentence. Do not include explanations or quotation marks.

Non-Conspiracy Prompt

System: You are an expert in linguistics and NLP, specializing in factual, neutral, and evidence-based discourse.

User: Given three example sentences that are not related to conspiracy theories, generate one new sentence that:

- Is clearly non-conspiratorial
- Does not suggest hidden control or secret coordination
- Is semantically different from the examples
- Does not reuse entities, phrases, or sentence structures
- Sounds natural and realistic
- Has a length roughly comparable to the examples

Examples:

1. `<example_1>`
2. `<example_2>`
3. `<example_3>`

Output: Return exactly one sentence. Do not include explanations or quotation marks.

F Appendix: SemEval 2026 Task 10 Leaderboard

This appendix provides the comprehensive leaderboard for the two subtasks of the SemEval 2026 Task 10 competition. Our team, **dangphuduy**, is highlighted in bold. **Note that the F1 scores presented in the tables have been rounded to two decimal places.**

F.1 Task 1: Span Extraction Results

Table 9: Official Leaderboard for Task 1 (Span Extraction). Our team ranked 4th.

Team	F1
HU	0.26
CredenceAI	0.24
CCNU	0.24
dangphuduy	**0.24**
UMUTeam	0.24
UCSC NLP	0.22
NUST PsyAI	0.21
AILS-NTUA	0.21
YNU-HPCC	0.21
Team A	0.19
CuriosAI	0.18
CUET_320	0.18
wangkongqiang	0.15
zhangpeng	0.14
Jia	0.08
AGAI	0.07

F.2 Task 2: Conspiracy Detection Results

Table 10: Official Leaderboard for Task 2 (Conspiracy Detection). Our team ranked 3rd.

Team	**F1**
NJUST_KMG	0.89
mdok-style	0.78
dangphuduy	**0.78**
VARH-AI	0.78
UMUTeam	0.77
HU	0.76
CuriosAI	0.76
NUST PsyAI	0.76
CSECU-DSG	0.75
LATE-iimas	0.75
davidinfotec	0.75
psy_detectives	0.75
AILS-NTUA	0.75
TruthGradient	0.75
wangkongqiang	0.74
YNU-HPCC	0.74
Unibuc-NLP	0.74
Hidetsune	0.73
Team A	0.73
CCNU	0.73
UCSC NLP	0.73
zhangpeng	0.73
Macaroni	0.73
TTLab	0.72
AGAI	0.70
CUET_320	0.59
Jia	0.37
GUNLP	0.34