Modeling Discourse Profiles in Climate Misinformation Discourse

Ge Gao, Zhengyang Shan, Bryan A. Plummer, Margrit Betke, Derry Wijaya
Boston University
{ggao02, shanzy, bplum, betke, wijaya}@bu.edu

Abstract

Prior work on climate misinformation detection typically classifies individual comments or claims using text alone. This framing overlooks a key property of online climate discourse: misinformation is often produced, reinforced, and sustained through recurring rhetorical patterns that reappear across comments and threads. Using a large-scale TikTok climate discourse dataset, we construct discourse profiles by clustering climate-related comments into recurring rhetorical patterns. We then incorporate these profiles into fine-grained climate misinformation classifiers using a late-fusion architecture. Under video-disjoint evaluation across three training seeds, discourse-profile-aware models significantly improve macro-F1 over text-only baselines for bag-of-words, TF-IDF, and MPNet representations. RoBERTa-large exhibits a positive mean improvement but substantially greater variability, and the difference is not statistically significant. Beyond classification, our results show discourse profiles shape misinformation reception and propagation. Discourse-profile-conditioned language model agents exhibit misinformation acceptance patterns aligned with profile-level misinformation prevalence in CLIMATEDISC, and the relationship between misinformation prevalence and engagement persistence differs across dominant discourse profiles.

1 Introduction

Short-video social media platforms such as TikTok have become major venues for public discussion of climate change, where accurate information, skepticism, and misinformation coexist in the same high-volume comment threads (Pearce et al., 2018; Sultana et al., 2024; Basch et al., 2022; Pera and Aiello, 2024; Baltasar et al., 2024). This setting is particularly challenging for automated analysis: short, algorithmically amplified videos drive rapid diffusion, while discussions are fast-moving, emotionally charged, allowing misinformation to persist and resist correction.

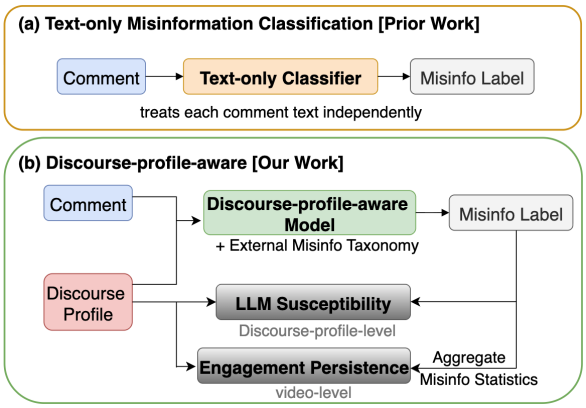


Figure 1: Comparison of text-only vs. discourse-profile-aware misinformation modeling. (a) Prior work (Islam et al., 2020; Coan et al., 2021; Upadhyaya et al., 2023; Choi et al., 2023; Rojas et al., 2024) typically predicts a misinformation category from comment text alone, treating each comment independently. (b) We construct discourse profiles by clustering comments and use discourse profile as additional context for misinformation-type prediction; the same discourse profile abstraction also supports discourse-profile-conditioned analyses of LLM susceptibility and engagement persistence.

As shown in Figure 1a, most prior work uses text-only classifiers that label each comment or claim in isolation (Islam et al., 2020; Coan et al., 2021; Upadhyaya et al., 2023; Choi et al., 2023; Rojas et al., 2024). However, climate discourse also exhibits recurring narrative frames and rhetorical styles that reappear across threads. This suggests that *how people consistently argue* can provide complementary signal beyond the words in a single comment.

We introduce *discourse profiles* to address this challenge: a compact, data-driven abstraction of recurring narrative frames and rhetorical styles in climate discourse. This perspective complements prior work showing that incorporating speaker and social context can improve misinformation detec-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2867–2887
November 6-10, 2026 ©2026 Association for Computational Linguistics

tion beyond text alone (Nguyen et al., 2020; Mehta et al., 2022), and that beliefs and framing are not fully captured by lexical features (Allaway and McKeown, 2021; Zhang et al., 2023). Rather than relying on predefined demographics or static author embeddings, we construct discourse profiles directly from large-scale, in-the-wild TikTok comments. Specifically, we cluster semantic comment embeddings and use LLM-based cluster consolidation to group the resulting fine-grained clusters into a small set of interpretable discourse profiles. Importantly, profiles are assigned at the *comment level*, not the user level, allowing the same user to exhibit different discourse patterns across contexts.

We then use discourse profiles in three ways (Figure 1b): (i) to augment fine-grained misinformation-type prediction, (ii) to construct a controlled susceptibility probe using discourse-profile-conditioned LLM agents, and (iii) to analyze how misinformation relates to engagement persistence across different discourse contexts.

Across video-disjoint evaluation with three training seeds, discourse-profile conditioning significantly improves macro-F1 for bag-of-words, TF-IDF, and MPNet classifiers, while RoBERTa-large shows a positive but statistically inconclusive mean difference. Discourse-profile-conditioned LLM agents also exhibit susceptibility patterns that align with profile-level misinformation prevalence in CLIMATEDISC. Finally, the relationship between misinformation prevalence and engagement persistence varies across dominant discourse profiles.

Our main contributions are:

- We introduce a data-driven discourse-profile abstraction for climate misinformation discourse, constructed from semantic clustering of large-scale TikTok comments and consolidation into interpretable profile types.
- We develop a discourse-profile-aware modeling and evaluation framework that incorporates profile information into fine-grained misinformation classification and uses matched text-only, shuffled-profile, and stylistic-variation controls to isolate when profile information contributes beyond text.
- We extend the same discourse-profile abstraction to controlled LLM susceptibility probing and video-level temporal analysis, revealing that the association between misinformation prevalence and engagement persistence differs across discourse profiles.

2 Related Work

2.1 Climate Misinformation Narratives

Climate misinformation includes not only outright denial, but also broader obstruction narratives such as questioning climate solutions, emphasizing economic trade-offs, or attacking scientific credibility (Lamb et al., 2020; Coan et al., 2021; Holder et al., 2023). We adopt the fine-grained QUOTACLIMAT taxonomy (QuotaClimat, 2025), which captures diverse climate obstruction narratives (Appendix B.1). Prior climate-misinformation and stance-detection methods discussed above typically predict labels from individual textual instances (Coan et al., 2021; Choi et al., 2023; Upadhyaya et al., 2023; Rojas et al., 2024). In contrast, we model recurring rhetorical patterns through discourse profiles and analyze how different profiles preferentially express specific misinformation types.

2.2 Author Context and Persona in NLP

Prior work shows that author context can improve stance and misinformation detection by capturing latent beliefs, interaction patterns, and communicative styles beyond text alone (Benton and Dredze, 2018; Qiu et al., 2015; Lynn et al., 2019; Barel et al., 2025). Related work has also explored dynamic and multi-persona user representations (Del Tredici et al., 2019; Marsden and Pröbster, 2019), while studies of climate denial rhetoric show that misinformation discourse often relies on recurring rhetorical and networking strategies rather than isolated claims (Bloomfield and Tillery, 2018). Existing approaches typically rely on predefined attributes, interaction structures, or static user embeddings. We instead construct discourse profiles directly from large-scale climate discourse and use them both for prediction and downstream behavioral analysis.

2.3 LLM Profiles and Engagement Dynamics

Recent work conditions LLMs on personas or discourse profiles to simulate subgroup behavior (Park et al., 2023; Krishna et al., 2022; Argyle et al., 2023; Li et al., 2024). We instead use discourse-profile-conditioned LLM agents as a controlled probe for misinformation susceptibility. Prior work has examined misinformation diffusion and persistence through content, emotion, and network structure (Vosoughi et al., 2018; Liu et al., 2023; Grinberg et al., 2019; Sun and Xie, 2024). We extend this line of work by showing that engagement

persistence varies across discourse profiles.

3 Method

Given short-form social media comments discussing climate change, our goal is to characterize and predict fine-grained climate misinformation by explicitly modeling recurring discourse-level regularities. For each comment c_i, we (i) assign a discourse profile d_i capturing recurring rhetorical patterns, (ii) obtain a fine-grained misinformation type y_i when applicable, and (iii) analyze how discourse profile conditions misinformation susceptibility and engagement persistence at both the comment and video levels.

Our pipeline proceeds in two stages. First, we construct discourse profiles from large-scale climate comment discourse (Figure 2; Section 3.1). Second, we incorporate discourse profile as an explicit signal for misinformation-type prediction (Figure 3) and for downstream analyses of susceptibility and engagement. Comment-level misinformation labels are obtained using a classifier trained on the curated QUOTACLIMAT taxonomy. Because this requires cross-domain transfer from curated claims to in-the-wild TikTok comments, we evaluate label quality using claim–comment matching, an independent LLM comparison, and a two-annotator human validation study on 100 randomly sampled CLIMATEDISC comments (Appendix A).

3.1 Discourse Profile Creation

To capture recurring rhetorical and discourse patterns used in Sections 3.2-3.4, we construct discourse profiles from large-scale climate change comments in CLIMATEDISC (Gao et al., 2025), a large-scale corpus of TikTok climate change videos and associated comments (Section 4.1), with the resulting profiles serving as a discourse-level abstraction across downstream tasks.

We represent comments as dense semantic embeddings using a pre-trained MPNet-based sentence transformer (Reimers and Gurevych, 2020). We then cluster the embeddings using k-means (Lloyd, 1982), selecting the number of clusters by jointly examining Within-Cluster Sum of Squares (WCSS) and Silhouette scores over a candidate range of $k \in [3, 30]$. This analysis identifies $k = 21$ as a practical trade-off between cluster compactness and separation (Section 4.3.1). We additionally compared against HDBSCAN on the same embeddings (Appendix C.2).

Local / Experiential Observers 1. Local Weather Anecdotes 12. Storms & Flooding 13. Geographical Self-Mentions	**Ideological / Political Combatants** 0. Anti-Trump/GOP/Biden 11. Government Blame / Anarchy 18. Electoral Mobilization
Systemic / Structural Critics 4. Consumption / Plastics Skepticism 17. Energy / Policy Debate 19. Agriculture / Diet 20. Anti-Corporation / Capitalism	**Pro-Science / Rationalist** 14. Pro-Science / Anti-Misinformation
Doomism / Nihilism 8. Doomism	**Conspiratorial / Anti-Climate** 7. Natural causes – not man-made 16. Denial / Hoax / Manipulation
Noise / Low-Signal (Excluded): 2. Emoji / Applause Spam; 3. Minimal Agreement Tokens; 5. Polite Thanks; 6. Short Emotional Bursts; 9. Excited Agreement; 10. Shared Anxiety; 15. @ Mentions.	

Table 1: Discourse profiles obtained by grouping the 21 k-means clusters of comment embeddings into semantically coherent discourse-profiles. The Noise / Low-Signal group is excluded from downstream discourse-profile-sensitive analyses.

To obtain a compact discourse-level abstraction for downstream modeling, we use GPT-5 for cluster consolidation. We prompt the model to group the 21 fine-grained semantic clusters into higher-level *discourse profiles* according to shared rhetorical patterns and communicative stances toward climate change (Appendix F.1). This procedure yields six higher-level discourse profiles, along with a residual 'Noise / Low-Signal' group excluded from discourse-profile-sensitive analyses. The resulting discourse profile definitions are reported in Table 1.

1. *"Pennsylvania, same dread."*
2. *"We're having an unseasonably warm winter here in Minnesota."*
3. *"This winter in Michigan there was only snow for a few weeks... now it's 50 degrees and sunny."*

While these comments differ in surface form, they all ground climate discussion in place-based lived experience. We therefore treat them as reflecting a shared discourse-level profile.

With discourse profiles defined, we next examine whether this abstraction provides predictive signal beyond comment text alone. Specifically, we evaluate whether incorporating discourse profile information improves fine-grained climate misinformation classification.

3.2 Profile-Aware Misinfo Classification

We compare text-only and discourse-profile-aware classifiers to test whether discourse profile provides predictive signal beyond comment text. Our goal is not to maximize benchmark performance, but to isolate when discourse profile contributes complementary information under different text represen-



Figure 2: Offline discourse profile construction: encode comments, cluster embeddings, and use GPT-5-based cluster consolidation to assign the resulting clusters to higher-level discourse profiles d_i (Section 3.1).

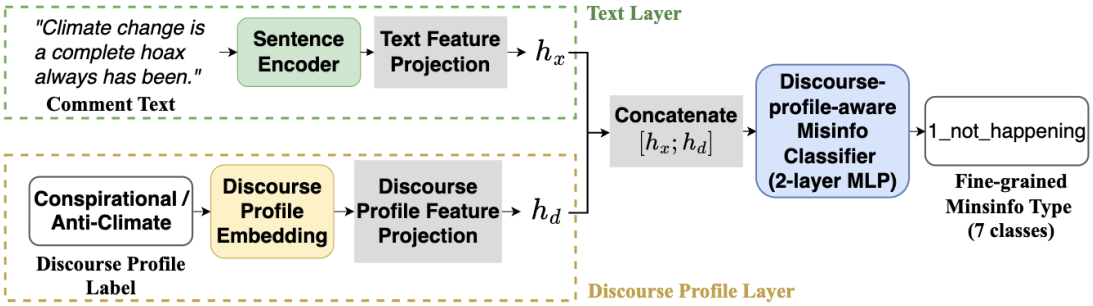


Figure 3: Late-fusion discourse-profile-aware misinformation modeling: obtain misinformation-type labels and evaluate prediction models that concatenate contextual text representations with discourse-profile embeddings before classification (Section 3.2). Details of how we obtain comment-level misinformation labels are in Appendix A.

tations and model capacities.

Text representations. Given a comment t in CLIMATEDISC, we encode it as a feature vector $\mathbf{x} = \phi(t)$, where $\phi(\cdot)$ may denote either a pre-trained dense sentence encoder or a lexical featurizer. In particular, we consider three families of text representations: (i) MPNet-based sentence embeddings (Reimers and Gurevych, 2020), (ii) non-contextual bag-of-words features, and (iii) TF-IDF (Jones, 2004) features. This allows us to examine the effect of discourse profile across both dense semantic and sparse lexical input spaces.

Text-only baseline. Model A is a text-only classifier that predicts fine-grained misinformation type from $\mathbf{x}$ alone. The text representation is passed through a nonlinear projection $\mathbf{h}_x = f_x(\mathbf{x})$, where $f_x(\cdot)$ denotes an affine transformation followed by ReLU and dropout, and then mapped to output logits $\hat{\mathbf{y}} = g(\mathbf{h}_x)$. The prediction space follows the seven-class misinformation taxonomy from QUOTACLIMAT, excluding `0_not_relevant`.

Discourse-profile-aware model. We incorporate discourse profile into misinformation prediction. Each discourse profile label $c \in \{1, \ldots, 6\}$ is mapped to a learnable embedding $\mathbf{d} = E(c)$.

We extend the text-only baseline by incorporating discourse profile as an additional latent signal. The discourse-profile embedding is projected as $\mathbf{h}_d = f_d(\mathbf{d})$, then concatenated with the text representation $\mathbf{h}_x$ to form $\mathbf{h} = [\mathbf{h}_x; \mathbf{h}_d]$. A joint nonlinear transformation $\mathbf{z} = f_j(\mathbf{h})$ is then applied before the final classifier, yielding $\hat{\mathbf{y}} = g(\mathbf{z})$. This formulation introduces discourse profile only at the classifier stage while keeping the contextual text representation unchanged.

Multi-seed evaluation. We use a fixed video-disjoint train/validation/test partition and repeat model training using three random seeds. We report accuracy and macro-F1 as mean across these runs. To assess whether correct-profile conditioning improves over the corresponding text-only model, we conduct a two-sided paired t-test on the three seed-level macro-F1 values.

Isolating profile signal. To distinguish whether improvements arise from discourse-profile information or merely increased model capacity, we conduct a controlled profile permutation test in which we randomly shuffle profile labels while holding the model architecture and training procedure fixed. For each of the three training seeds, we train ten shuffled-profile models using independently permuted profile assignments. We compare the shuffled-profile macro-F1 distribution with the corresponding correct-profile score using a two-sided one-sample t-test.

Robustness under stylistic variation. We additionally evaluate robustness under LLM-generated stylistic rewrites. Following Brian et al. (2025), we generate multi-round discourse-profile-conditioned variants that preserve the underlying misinformation category while varying rhetorical style and surface realization. We then evaluate both text-

only and discourse-profile-aware classifiers on the rewritten comments. Generation details, prompts, and examples are provided in Appendix E.

3.3 Discourse-Profile-Level Misinformation Susceptibility with LLM

We use discourse-profile-conditioned LLM agents as a controlled probe of whether the constructed discourse profiles recover systematic differences in misinformation susceptibility. We construct a balanced evaluation set of 350 misinformation claims from QUOTACLIMAT, sampling 50 claims from each of the seven categories. These claims are separate from the CLIMATEDISC comments used to construct the discourse profiles. We use GPT-5 as the language model for the discourse-profile-conditioned susceptibility probe. For each inferred discourse profile, we construct a natural-language profile card summarizing characteristic beliefs, goals, tone, and rhetorical tendencies (Appendix F.2). The cards do not contain profile-level misinformation rates, category-specific acceptance rates, or an expected ordering of profiles by susceptibility. Because the cards nevertheless encode characteristic beliefs and rhetorical tendencies, we interpret this experiment as a controlled profile-specific probe rather than a simulation of individual users.

As a specificity control, we randomly shuffle the profile-to-card mapping 100 times while keeping the claim set and evaluation procedure fixed. For each permutation, we recompute the correlation between the CLIMATEDISC profile-level misinformation rates and LLM susceptibility scores.

For discourse profile d, we define an LLM susceptibility score as the proportion of misinformation claims not rejected by the discourse-profile-conditioned LLM agent:

$$L_d = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{I}[r_{d,q} \neq \text{reject}],$$

where Q is the set of evaluation claims, $r_{d,q}$ is the LLM's response for discourse profile d on claim q, and $\mathbb{I}[\cdot]$ denotes an indicator function.

Human–LLM Discourse Profile Alignment. We assess whether discourse-profile-conditioned LLM susceptibility patterns align with those of real human discourse profiles. Let the comments in CLIMATEDISC be indexed by $i \in \{1, \ldots, N\}$, where d_i denotes the discourse profile assigned to comment i and y_i denotes its misinformation label. For each discourse profile d, we define the set of comments assigned to that profile as

$$C_d = \{i \in \{1, \ldots, N\} : d_i = d\}.$$

We then define the human misinformation rate for profile d as

$$H_d = \frac{1}{|C_d|} \sum_{i \in C_d} \mathbb{I}[y_i \neq 0],$$

where $y_i = 0$ denotes `not_relevant` and $y_i \in \{1, \ldots, 7\}$ denotes one of the seven misinformation categories. Thus, H_d represents the proportion of comments associated with discourse profile d that are labeled as misinformation. We compute the LLM susceptibility score L_d previously defined. We then measure alignment between (H_d) and (L_d) using Pearson and Spearman correlations, where higher correlation indicates closer correspondence between human and LLM discourse profile behavior. We next examine whether these discourse-profile-level differences translate into distinct engagement trajectories.

3.4 Discourse-Profile Temporal Dynamics

To examine how discourse profile conditions the downstream impact of misinformation, we analyze engagement persistence at the video level. For each video in CLIMATEDISC we compute (i) its dominant discourse profile, defined as the majority profile among its comments, (ii) its misinformation rate, and (iii) engagement persistence, measured as the time required to reach 60% of total comments (Liu et al., 2023). Our dataset for this engagement persistence experiment contains 2,925 videos[1].

Formally, for a video v with N_v total comments, let $t_{v,i}$ denote the timestamp of the i-th comment. We define the cumulative number of comments received by time t as

$$S_v(t) = \sum_{i=1}^{N_v} \mathbb{I}[t_{v,i} \leq t],$$

where $\mathbb{I}[\cdot]$ is the indicator function. Let $t_v^{(0)}$ denote the timestamp of the first observed comment for video v. The time at which the video reaches a fraction α of its total comments is defined as $T_v(\alpha) = \inf\{t : S_v(t) \geq \alpha N_v\}$. We define engagement persistence as $P_v = T_v(0.6) - t_v^{(0)}$, which measures the duration (in hours) required

[1]We exclude videos with zero or one comment, since persistence requires at least two comments to be defined.

for video v to accumulate 60% of its total comments. Larger values of P_v indicate more sustained engagement over time.

To examine whether persistence patterns differ across discourse contexts, we stratify videos into low, mid, and high misinformation groups by misinformation rate. Within each discourse profile type, we conduct Kruskal–Wallis tests to assess whether persistence distributions differ significantly across misinformation levels.

4 Experiments

4.1 Datasets

QUOTACLIMAT. To obtain fine-grained misinformation types, we use QUOTACLIMAT (QuotaClimat, 2025), a curated dataset of 6,091 climate-related statements annotated with one of the misinformation categories specified in Appendix B.1. We use QUOTACLIMAT for two purposes: (i) training a misinformation classifier that is transferred to CLIMATEDISC comments (Appendix A), and (ii) susceptibility probing with discourse-profile-conditioned LLM agents using a balanced set of 350 claims (50 per category; Section 3.3).

CLIMATEDISC. For discourse profile creation and downstream in-the-wild analyses, we rely on the CLIMATEDISC dataset (Gao et al., 2025), which comprises transcribed English speech from 7,110 TikTok videos and 116,256 corresponding comments tagged with "#climatechange," collected between January 2024 and November 2024. We use CLIMATEDISC in: (i) discourse profile creation (Section 3.1), (ii) discourse-profile-aware misinformation classification on in-the-wild comments (Section 3.2), (iii) human discourse-profile-level misinformation rates H_d for LLM alignment (Section 3.3), and (iv) temporal dynamics and engagement persistence (Section 3.4).

For the discourse-profile-aware misinformation classification experiments on CLIMATEDISC (Section 3.2), comments are divided into training, validation, and test sets in a 0.8/0.1/0.1 ratio. The splits are video-disjoint. All comments from a given video occur in exactly one partition, preventing video-level overlap between training and evaluation data. The resulting split contains 2,717 training videos, 344 validation videos, and 347 test videos. Noise / Low-Signal comments are excluded.

4.2 Implementation Details

For discourse-profile-aware misinformation classification, we use lightweight feed-forward models with MPNet embeddings, bag-of-words features, and TF-IDF features as input. For higher-capacity contextual encoders, we instantiate text-only and discourse-profile-aware variants using RoBERTa-large (Liu et al., 2019). We use the fixed video-disjoint train/validation/test split described above. Lightweight models are trained with Adam using learning rate 3×10^{-4}, batch size 32, and early stopping on validation macro-F1. RoBERTa-large models are trained with Adam using learning rate 2×10^{-5}, batch size 8, maximum sequence length 128, and early stopping on validation macro-F1. All models use cross-entropy loss and gradient clipping at 1.0.

4.3 Results

Table 2 summarizes misinformation classification performance under the video-disjoint evaluation setting. All reported accuracy and macro-F1 values are averaged across three training seeds. For the lightweight MLP-based classifiers, incorporating the correct discourse profile improves performance across all three text representations.

For bag-of-words features, discourse-profile conditioning increases mean accuracy from 0.690 to 0.709 and macro-F1 from 0.586 to 0.616, corresponding to a macro-F1 improvement of +0.030. The improvement is significant across training seeds under a paired t-test ($p = 0.028$). The largest gain is observed with TF-IDF features, where mean accuracy increases from 0.675 to 0.709 and macro-F1 from 0.522 to 0.601 (ΔF1 = +0.080, $p = 0.006$). With frozen MPNet sentence embeddings (Song et al., 2020), the profile-aware model improves mean accuracy from 0.799 to 0.808 and macro-F1 from 0.726 to 0.739 (ΔF1 = +0.014, $p = 0.015$). These results indicate that explicit discourse-profile information provides complementary predictive signal across both sparse lexical and dense sentence-level representations.

The effect is less conclusive for the higher-capacity contextual encoder. For RoBERTa-large, the profile-aware model achieves higher mean accuracy (0.856 vs. 0.845) and macro-F1 (0.810 vs. 0.794), yielding an average macro-F1 difference of +0.015. However, this difference varies substantially across training seeds and is not statistically significant ($p = 0.381$). We therefore interpret

Representation	Text Acc	Prof Acc	Δ Acc	Text F1	Prof F1	Δ F1	Paired p
[MLP]							
Bag-of-Words	0.690	0.709	+0.019	0.586	0.616	+0.030	0.028
TF-IDF	0.675	0.709	+0.034	0.522	0.601	+0.080	0.006
MPNet	0.799	0.808	+0.010	0.726	0.739	+0.014	0.015
RoBERTa-large	0.845	0.856	+0.011	0.794	0.810	+0.015	0.381

Table 2: Comparison of text-only (**Text**) and discourse-profile-aware (**Prof**) misinformation classification under video-disjoint evaluation. Accuracy and macro-F1 values are means across three training seeds. p-values are from two-sided paired t-tests comparing seed-level macro-F1 scores between the correct-profile and corresponding text-only models. Overall, correct-profile conditioning improves macro-F1 for bag-of-words, TF-IDF, and MPNet, while the RoBERTa-large difference is not statistically significant across seeds.

the RoBERTa-large result as a positive but statistically inconclusive average difference rather than evidence of a reliable improvement from discourse-profile conditioning.

To examine whether the aggregate gains are concentrated in only a small subset of misinformation types, we additionally conduct a category-level error analysis in Appendix D.

Contribution of Correct Profile Assignment. To isolate discourse-profile signal from added model capacity, we compare correct-profile conditioning against shuffled-profile controls averaged over 10 random permutations (Table 3). The shuffled-profile control reveals substantial differences across text representations. For bag-of-words, correct profiles outperform shuffled profiles in all three training seeds, with a mean macro-F1 advantage of $+0.034$; shuffled-profile performance is approximately equal to the text-only baseline. TF-IDF likewise shows a consistent correct-versus-shuffled advantage of $+0.040$ across all three seeds. These results provide the clearest evidence that the specific comment–profile correspondence contributes predictive information in the sparse lexical settings. The pattern is weaker for dense representations. For MPNet, correct profiles outperform shuffled profiles by only $+0.003$ macro-F1 on average, and the correct-versus-shuffled contrast is significant in only one of three seeds. Thus, although the correct-profile MPNet model significantly exceeds the text-only model across seeds, much of that improvement cannot be uniquely attributed to the semantic correctness of the profile assignment. RoBERTa-large similarly exhibits unstable shuffled-profile behavior, with a significant correct-versus-shuffled difference in only one of three seeds.

4.3.1 Discourse Profile Construction

As described in Section 3.1, we construct the initial discourse profiles using k-means clustering and select the number of clusters k by jointly examining WCSS and Silhouette scores over $k \in [3, 30]$. Figure 4 reports these cluster-selection diagnostics. WCSS decreases smoothly without a decisive elbow, whereas the Silhouette score reaches a local maximum at $k = 21$. We therefore set $k = 21$ as a practical trade-off between separation and interpretability. As a qualitative check, Appendix B.2 lists the ten comments nearest each centroid, which exhibit consistent topical and rhetorical patterns.

We aggregate the 21 clusters into six higher-level discourse profiles plus a residual "Noise / Low-Signal" group. After consolidation, the six higher-level discourse profiles remain relatively balanced (Appendix C), reducing the likelihood that downstream effects are driven by a single dominant profile. We exclude the Noise group from discourse-profile-sensitive analyses (it consists largely of short acknowledgments, emojis, and @-mentions). Table 1 summarizes the mapping.

Across ten independent k-means initializations at $k = 21$, mean pairwise adjusted Rand index is 0.741 and mean normalized mutual information is 0.815. Cluster-size coefficients of variation range from 0.301 to 0.448, and after consolidation into the higher-level discourse profiles, average profile-assignment agreement remains 82.8%. As a supplemental interpretability check, we asked two independent LLM raters, GPT-5.5 and Claude Opus 4.8, to reassign each of the 21 clusters. The raters were blind to the original cluster-to-profile mapping and assigned each cluster to one of the six discourse profiles or Noise / Low-Signal. The two raters agreed on 19 of 21 clusters ($\kappa = 0.88$). Their agreement with the original consolidation was $\kappa = 0.71$ and 0.64, respectively (Appendix C.1).

Representation	Text F1	Shuffled F1	Correct F1	Correct − Shuffled	Significant seeds
[MLP]					
Bag-of-Words	0.586	0.582	0.616	+0.034	3 / 3
TF-IDF	0.522	0.561	0.601	+0.041	3 / 3
MPNet	0.726	0.736	0.739	+0.003	1 / 3
RoBERTa-large	0.794	0.776	0.810	+0.033	1 / 3

Table 3: Permutation-test evaluation of discourse-profile conditioning across text representations. Shuffled-profile scores are first averaged across ten independent profile shuffles within each training seed; reported values are then averaged across the three training seeds. The last column reports how many of the three random seeds show a significant correct-vs-shuffled difference under the one-sample t-test. The correct-versus-shuffled advantage is most consistent for bag-of-words and TF-IDF, indicating that the specific profile assignment contributes most clearly in sparse lexical settings.

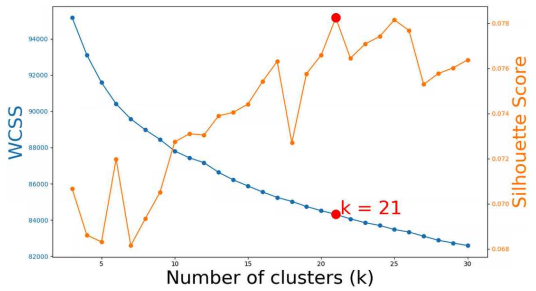


Figure 4: Within-Cluster Sum of Squares (WCSS; lower is better) and Silhouette Score (higher is better) across varying numbers of clusters k. The Silhouette score reaches a local maximum at $k = 21$.

Discourse Profile	Susceptible (%)
Local / Experiential Observers	29.4
Ideological / Political Combatants	4.9
Systemic / Structural Critics	8.0
Pro-Science / Rationalist	10.4
Doomism / Nihilism	22.0
Conspiratorial / Anti-Climate	**96.3**

Table 4: Overall misinformation susceptibility rate by discourse profile. Conspiratorial / Anti-Climate profile exhibits the highest acceptance rate.

4.3.2 Discourse-profile-Level Susceptibility Probing with LLM

Human vs. LLM Agents Susceptibility. We observe a strong positive correlation between human and LLM discourse profile susceptibility patterns. The Pearson correlation between H_d and L_d (specified in Section 3.3) is 0.82, indicating substantial linear agreement. The Spearman rank correlation is even higher at 0.94, suggesting that the relative ordering of discourse profiles by misinformation susceptibility is highly consistent between humans and LLM agents. This alignment suggests that discourse-profile-conditioned LLM agents capture meaningful, discourse-profile-specific structure in misinformation acceptance.

Profile card permutation control. To test whether this alignment depends on the correct profile card correspondence, we randomly permute the profile-to-card mapping 100 times while keeping the claim set and evaluation procedure fixed. Under permutation, the mean Spearman correlation is $\rho = -0.10$, whereas the true profile card assignment yields $\rho = 0.94$ and is significant under the permutation test ($p = 0.03$). This indicates that the observed correlation depends on the specific profile card correspondence rather than only on a generic tendency of the LLM to accept or reject the claims.

Susceptibilities Across Discourse Profiles. Table 4 summarizes overall susceptibility score (specified in Section 3.3) per discourse profile. *Conspiratorial / Anti-Climate* shows very high susceptibility (96.3%), while other discourse profiles range from 4.9% to 29.4%. Notably, *Ideological / Political Combatants* show the lowest overall susceptibility (4.9%), suggesting that strong political engagement does not necessarily translate to greater vulnerability to misinformation. Similarly, the *Pro-Science / Rationalist* demonstrates comparatively low susceptibility (10.4%), though this group is not immune to misinformation, indicating that scientific self-identification alone does not guarantee resistance. *Doomism / Nihilism* is comparatively higher (22.0%), highlighting the role of affective orientations such as resignation or fatalism in shaping misinformation vulnerability.

4.3.3 Engagement Dynamics Across Profiles

To examine how persistence co-varies with misinformation prevalence, we group videos into Low, Mid, and High misinformation bins and compute mean persistence hours for each profile-bin combination (Figure 5). For *Pro Science*, persistence

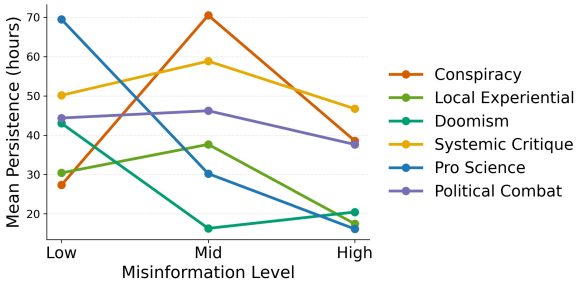


Figure 5: Mean persistence hours across misinformation level (Low, Mid, High) for each profile type.

Discourse Profile	H Statistic	Raw p	BH-FDR p
Systemic Critique	20.501	< 0.001	**0.003**
Local Experiential	16.336	< 0.001	**0.003**
Conspiracy	9.237	0.010	**0.020**
Pro Science	4.875	0.047	0.071
Political Combat	3.850	0.146	0.175
Doomism	2.912	0.233	0.233

Table 5: Within-profile Kruskal–Wallis tests examining whether persistence differs across misinformation tertiles (Low, Mid, High). BH-FDR denotes Benjamini–Hochberg correction across the six profile-level tests. Bold values remain significant at $\alpha = 0.05$. After BH-FDR correction, persistence differs significantly across misinformation tertiles for Systemic Critique, Local Experiential, and Conspiracy.

is higher in the low-misinformation group (69.5 hours) and lower in the high-misinformation group (16.1 hours). Within the *Conspiracy* profile, the high-misinformation group exhibits higher mean persistence than the low-misinformation group. *Systemic Critique* exhibits relatively small variation across bins. *Doomism* and *Local Experiential* show lower persistence in high-misinformation groups. For example, *Doomism* decreases from 43.1 hours (Low) to 20.3 hours (High), while *Local Experiential* increases from Low (30.5 hours) to Mid (37.3 hours) before decreasing at High (16.8 hours).

To assess whether these differences reflect statistically significant heterogeneity, we conduct within-profile Kruskal–Wallis tests comparing persistence across misinformation levels and apply Benjamini–Hochberg FDR correction across the six profile-level tests. Table 5 shows that, after correction, *Systemic Critique* ($p_{\text{adj}} = 0.003$), *Local Experiential* ($p_{\text{adj}} = 0.003$), and *Conspiracy* ($p_{\text{adj}} = 0.020$) remain significant. *Pro Science*, which is significant before correction ($p = 0.047$), no longer reaches significance ($p_{\text{adj}} = 0.071$), while *Political Combat* and *Doomism* remain non-significant. Thus, after multiple-comparison correction, we find evidence of persistence differences across misinformation levels for three profiles, while such differences are not detected for the remaining three.

5 Discussion and Conclusion

This work demonstrates that discourse profile is a useful, interpretable axis for climate misinformation research. By modeling how comments are framed in addition to what is said, we show that discourse profiles capture discourse-level regularities that complement comment text.

Our classification experiments show the clearest benefits of discourse-profile conditioning for lightweight text representations. Under video-disjoint evaluation, the correct-profile models significantly improve macro-F1 across three training seeds for bag-of-words, TF-IDF, and frozen MPNet representations. The largest improvement occurs with TF-IDF, while the MPNet improvement is smaller. RoBERTa-large shows a positive average difference, but the effect varies substantially across seeds and does not reach statistical significance. We also show that discourse-profile-conditioned LLM agents can serve as controlled probes of misinformation susceptibility. Holding claims constant while varying discourse profile produces systematic differences in acceptance behavior that align with human discourse-profile-level vulnerability. This suggests that discourse profiles provide a meaningful abstraction not only for classification, but also for analyzing susceptibility patterns.

Finally, our temporal dynamics analysis shows that engagement persistence co-varies with misinformation prevalence differently across discourse profiles. Some discourse contexts exhibit higher persistence in higher-misinformation settings, while others show the opposite pattern. These heterogeneous associations suggest that misinformation prevalence and engagement persistence are not uniformly distributed across discourse contexts, highlighting the importance of accounting for rhetorical variation when analyzing climate misinformation dynamics.

Together, these findings suggest that discourse-profile-aware modeling offers a principled way to surface discourse-level structure in climate misinformation discourse, unifying prediction, robustness analysis, susceptibility probing, and engagement dynamics within a single abstraction. Rather than replacing text-centric approaches, discourse profiles complement them by capturing stable patterns of expression and interaction that are difficult to recover from isolated comments alone.

Limitations

Discourse Profile Construction Validity and Subjectivity. Our discourse profiles are constructed via embedding-based k-means clustering followed by GPT-5-based consolidation of the 21 fine-grained clusters into six higher-level types. While this improves interpretability, the resulting taxonomy is not uniquely determined: different embedding models, clustering algorithms, random seeds, k-selection criteria, language models, or consolidation prompts could yield alternative groupings. In particular, the consolidation step is model- and prompt-dependent and should not be interpreted as producing a uniquely correct taxonomy. Accordingly, the resulting discourse profiles are best viewed as dataset-specific rhetorical groupings rather than definitive or exhaustive categories of users. The independent LLM reassignment analysis provides a supplemental reproducibility check, but it does not substitute for conventional human inter-annotator agreement.

We infer discourse profiles from observed comment text and do not have access to ground-truth user attributes (e.g., demographics, ideology, geography, or identity) or off-platform behavior. Consequently, discourse profile labels should not be interpreted as psychological traits or real-world subpopulations; they reflect recurring linguistic and rhetorical patterns expressed in public TikTok comments within our collection.

Cross-Domain Misinformation Label Transfer. Our misinformation labels on TikTok comments are produced by a classifier trained on QUOTACLIMAT and transferred to in-the-wild CLIMATEDISC discourse. In addition to claim–comment matching and LLM-based agreement checks, we evaluate this transfer on 100 randomly sampled comments independently labeled by two human annotators (Appendix A.3). Human–human agreement is substantial ($\kappa = 0.76$), whereas classifier–human agreement is moderate ($\kappa = 0.57$ and 0.55). These results provide direct evidence about cross-domain label quality but also confirm non-negligible label noise. The QUOTACLIMAT taxonomy is claim-centric and may not fully capture sarcasm, implicature, partial agreement, or conversational context in TikTok comments. We therefore treat individual transferred labels as imperfect and interpret downstream results primarily as aggregate profile-level associations.

Cross-Platform Generalization. The discourse profiles are constructed from a single TikTok climate-discourse corpus. We do not claim that the resulting six-profile taxonomy is platform invariant; rhetorical conventions and interaction patterns may differ across platforms, topics, and communities. Cross-platform profile construction and validation therefore remain important directions for future work.

LLM Agents Susceptibility. Our LLM-based susceptibility results should be interpreted as a controlled probe rather than a behavioral simulation of real users. Outputs can be sensitive to prompt wording, decoding settings, and the specific model used, and we evaluate only a single LLM configuration. Furthermore, our metric captures refusal/acceptance patterns under instruction rather than real-world exposure, incentives, or social context. We therefore emphasize alignment in relative discourse profile ordering rather than treating absolute LLM susceptibility rates as estimates of human vulnerability.

Temporal Dynamics and Causality. Our persistence analysis faces important methodological constraints. First, persistence hours conflate organic sharing velocity, algorithmic promotion, and content moderation decisions, precluding mechanistic attribution. Second, our cross-sectional design cannot establish whether misinformation density affects persistence or whether longer-persisting videos accumulate more misinformation over time. Third, video-level aggregation may obscure important dynamics at finer temporal scales.

Platform and Data Constraints. TikTok's proprietary recommendation algorithm and content moderation policies are opaque, limiting our ability to interpret persistence patterns mechanistically. We cannot observe deleted, shadowbanned, or algorithmically suppressed content, potentially biasing our sample toward platform-permitted content.

Ethical Considerations. Our analysis focuses on public comments and does not identify or track individual users. Nevertheless, discourse profile taxonomies could potentially be misinterpreted or misused as a form of user profiling. We emphasize that this is neither our intent nor what our method does. The discourse profiles in this work are comment-level abstractions constructed from recurring rhetorical patterns in public discourse,

not demographic, psychological, or persistent individual traits. A user may express different discourse patterns across comments or contexts, and our analysis does not observe or infer real-world user identities, beliefs, or susceptibility. Accordingly, discourse profiles should be used to study aggregate discourse patterns and model behavior, not to target or make claims about individual users.

Acknowledgments

Personally Identifiable / Offensive Information. We anonymize all potentially personally identifiable information in our released materials, including file paths and organization names in the code. In addition, examples of user comments shown in the paper have been sanitized to remove profane or offensive language.

AI Assistants in Writing. We used ChatGPT to assist with language polishing, LaTeX formatting refinement, and code refactoring. Separately, language models were used as methodological or experimental components in GPT-5-based cluster consolidation, susceptibility probing, stylistic-rewrite generation, annotation, and supplemental profile-reassignment analyses described in the corresponding sections.

References

Emily Allaway and Kathleen McKeown. 2021. Zero-shot stance detection: A dataset and model analysis. In *NAACL*.

Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31:1–15.

Clara Baltasar, Sergio D'Antonio Maceiras, Alejandro Martín, and David Camacho. 2024. Analysis of climate change misleading information in tiktok. In *Proceedings of the 1st Workshop on COuntering Disinformation with Artificial Intelligence (CODAI), co-located with the 27th European Conference on Artificial Intelligence (ECAI 2024)*, volume 3782 of *CEUR Workshop Proceedings*, pages 54–61, Santiago de Compostela, Spain. CEUR-WS.org. Creative Commons CC BY 4.0.

Guy Barel, Oren Tsur, and Dan Vilenchik. 2025. Acquired TASTE: Multimodal stance detection with textual and structural embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6492–6504, Abu Dhabi, UAE. Association for Computational Linguistics.

Corey Basch, Bhavya Yalamanchili, and Joseph Fera. 2022. #climate change on tiktok: A content analysis of videos. *Journal of Community Health*, 47(1):163–167.

Adrian Benton and Mark Dredze. 2018. Using author embeddings to improve tweet stance classification. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 184–194, Brussels, Belgium. Association for Computational Linguistics.

Emma Frances Bloomfield and Denise Tillery. 2018. The circulation of climate change denial online: Rhetorical and networking strategies on facebook. *Environmental Communication*, 13:1–12.

Chong Jun Rong Brian, Yixuan Tang, and Anthony Kum Hoe Tung. 2025. MPCG: Multi-round persona-conditioned generation for modeling the evolution of misinformation with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34030–34064, Suzhou, China. Association for Computational Linguistics.

YeonJung Choi, Lanyu Shang, and Dong Wang. 2023. Climatemist: Climate change misinformation and stance detection dataset.

Travis Coan, Constantine Boussalis, John Cook, and Mirjam Odile Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific Reports*, 11.

Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. 2019. You shall know a user by the company it keeps: Dynamic representations for social media users in NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4707–4717, Hong Kong, China. Association for Computational Linguistics.

Ge Gao, Zhengyang Shan, James Crissman, Ekaterina Novozhilova, YuCheng Huang, Arti Ramanathan, Margrit Betke, and Derry Wijaya. 2025. Insights into climate change narratives: Emotional alignment and engagement analysis on TikTok. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 128–143, Vienna, Austria. Association for Computational Linguistics.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on twitter during the 2016 u.s. presidential election. *Science*, 363(6425):374–378.

Faye Holder, Sanober Mirza, Namson-Ngo-Lee, Jake Carbone, and Ruth E. McKie. 2023. Climate obstruction and facebook advertising: how a sample of climate obstruction organizations use social media to disseminate discourses of delay. *Climatic Change*, 176:1–21.

Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10.

K. Jones. 2004. A statistical interpretation of term specificity in retrieval. *Journal of Documentation*, 60:493–502.

Ranjay Krishna, Donsuk Lee, Li Fei-Fei, and Michael Bernstein. 2022. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences*, 119.

William Lamb, Giulio Mattioli, Sebastian Levi, J. Roberts, Stuart Capstick, Felix Creutzig, Jan Minx, Finn Müller-Hansen, Trevor Culhane, and Julia Steinberger. 2020. Discourses of climate delay. *Global Sustainability*, 3.

Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2024. The steerability of large language models toward data-driven personas. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7290–7305, Mexico City, Mexico. Association for Computational Linguistics.

Maggie Liu, Jing Wang, and Daniel Preotiuc-Pietro. 2023. Analyzing and predicting persistence of news tweets. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–732, Nusa Dua, Bali. 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. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

S. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

Veronica Lynn, Salvatore Giorgi, Niranjan Balasubramanian, and H. Andrew Schwartz. 2019. Tweet classification without the tweet: An empirical examination of user versus document attributes. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 18–28, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicola Marsden and Monika Pröbster. 2019. Personas and identity looking at multiple identities to inform the construction of personas. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Nikhil Mehta, Maria Leonor Pacheco, and Dan Goldwasser. 2022. Tackling fake news detection by continually improving social context representations using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1380, Dublin, Ireland. Association for Computational Linguistics.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information &; Knowledge Management*, page 1165–1174. ACM.

Joon Park, Joseph O'Brien, Carrie Cai, Meredith Morris, Percy Liang, and Michael Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Warren Pearce, Sabine Niederer, Suay Ozkula, and Natalia Sanchez-Querubin. 2018. The social media life of climate change: Platforms, publics, and future imaginaries. *Wiley Interdisciplinary Reviews: Climate Change*, 10.

Arianna Pera and Luca Aiello. 2024. Shifting climates: Climate change communication from youtube to tiktok. In *Proceedings of the 16th ACM Web Science Conference*, pages 376–381.

Minghui Qiu, Yanchuan Sim, Noah A. Smith, and Jing Jiang. 2015. Modeling user arguments, interactions, and attributes for stance prediction in online debate forums. In *SDM*.

QuotaClimat. 2025. Frugal ai challenge text train dataset. Dataset for the Frugal AI Challenge 2025, focused on climate disinformation narratives.

Nils Reimers and Iryna Gurevych. 2020. all-mpnet-base-v2: Sentence-transformers model. Fine-tuned for sentence embeddings on 1B+ sentence pairs.

Cristian Rojas, Frank Algra-Maschio, Mark Andrejevic, Travis Coan, John Cook, and Yuan-Fang Li. 2024. Hierarchical machine learning models can identify stimuli of climate change misinformation on social media. *Communications Earth & Environment*, 5.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *ArXiv*, abs/2004.09297.

Bebe Chand Sultana, Md Prodhan, Edris Alam, Md Sohel, A. B. M. Mainul Bari, Subodh Pal, Md Islam, and Abu Islam. 2024. A systematic review of the nexus between climate change and social media: present status, trends, and future challenges. *Frontiers in Communication*, 9:1301400.

Yanqing Sun and Juan Xie. 2024. Who shares misinformation on social media? a meta-analysis of individual traits related to misinformation sharing. *Computers in Human Behavior*, 158:108271.

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023. Towards sentiment and temporal aided stance detection of climate change tweets. *Information Processing & Management*, 60:103325.

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

Chong Zhang, Zhenkun Zhou, Xingyu Peng, and Ke Xu. 2023. Doubleh: Twitter user stance detection via bipartite graph neural networks. *Preprint*, arXiv:2301.08774.

A Taxonomy Transfer Details

This appendix describes how we obtain fine-grained misinformation-type labels for in-the-wild comments by transferring a curated taxonomy to the CLIMATEDISC corpus.

A.1 Misinformation Classifier Training

We fine-tune a RoBERTa-large classifier on QUOTACLIMAT to predict one of eight climate misinformation categories from text. Training is performed on a single NVIDIA A100 GPU for four epochs with batch size 128 and learning rate 2×10^{-5}. Our model reaches a test accuracy of 0.75 on the eight-class QUOTACLIMAT classification task. This provides a source-domain performance reference; because the downstream application is to in-the-wild TikTok comments, we separately evaluate cross-domain agreement below.

A.2 Cross-domain Validation via Claim-Comment Matching

To validate transfer from curated claims to in-the-wild comments, we embed both QUOTACLIMAT claims and CLIMATEDISC comments using the same sentence transformer and identify high-similarity claim–comment pairs via cosine similarity. Pairs exceeding a fixed threshold of $\tau = 0.85$ are treated as referring to the same underlying claim. We select $\tau = 0.85$ via qualitative inspection: it is the lowest value that consistently yields pairs about the same underlying claim with aligned stance; lower thresholds increasingly admit topic-level matches with stance mismatches. This produces 208 matched comment–claim pairs (Table 6).

On these pairs, the classifier's prediction on the TikTok comment matches the ground-truth label of the paired QUOTACLIMAT claim 79% of the time (majority baseline 38%), with Cohen's $\kappa = 0.72$ indicating strong agreement beyond chance (Table 7). Manual inspection and quantitative agreement metrics further support that matched pairs refer to the same underlying claim.

Based on this validation, we apply the classifier to all comments in CLIMATEDISC to characterize misinformation types in in-the-wild discourse.

To assess label quality beyond the high-similarity regime, where explicit claim alignment is unavailable, we conduct an additional agreement analysis using an independent LLM annotator. We randomly sample 200 comments from CLIMATEDISC that do not exceed the similarity threshold with any QUOTACLIMAT claim. Each comment is annotated independently by ChatGPT using a fixed prompt (Appendix F.5), and we compare these annotations with our classifier's predictions.

Comment (ClimateDisc)	Matched Claim (QuotaClimat)
Climate change is a HOAX	Climate change is a complete hoax always has been.
Yes, climate change is real to what extent is arguable, but definitely changing	Climate change is definitely real. It changes every year from winter to autumn to spring to summer.
Climate change is very real. The climate has always changed, but in recent years it is changing too fast	Climate change is definitely real. It changes every year from winter to autumn to spring to summer.

Table 6: Representative matched comment–claim pairs obtained using a cosine-similarity threshold of 0.85.

Cohen's Kappa	Quality
>0.8	Almost Perfect Agreement
>0.6	Substantial Agreement
>0.4	Moderate Agreement
>0.2	Fair Agreement
0-0.2	Slight Agreement
<0	Almost No Agreement

Table 7: Interpretation for different ranges of the Cohen's Kappa values.

Across these unmatched comments, agreement reaches 89%, with Cohen's $\kappa = 0.54$. The high raw agreement but moderate κ provides an additional cross-domain agreement check while also indicating that substantial uncertainty remains. Because an LLM comparison does not substitute for human annotation, we complement this analysis with an independent human validation study below.

A.3 Cross-Domain Human Validation

To directly assess label quality on in-the-wild TikTok discourse, we randomly sample 100 comments from CLIMATEDISC. Two human annotators independently assign each comment one of the eight QUOTACLIMAT categories using the same category definitions as the classifier. We assess both human–human agreement and agreement between each human annotator and the transferred classifier predictions using Cohen's κ.

The two human annotators show substantial agreement ($\kappa = 0.76$), indicating that the taxonomy can be applied with reasonable consistency to the sampled in-the-wild TikTok comments. Agreement between the classifier and each human an-

Comparison	Cohen's κ
Human 1 $\leftrightarrow$ Human 2	0.76
Human 1 $\leftrightarrow$ classifier	0.57
Human 2 $\leftrightarrow$ classifier	0.55

Table 8: Cross-domain agreement on 100 randomly sampled CLIMATEDISC comments independently labeled by two human annotators.

notator is moderate ($\kappa = 0.57$ and 0.55). These results provide direct human evidence about cross-domain label quality while also confirming that non-negligible label noise remains. We therefore treat the CLIMATEDISC misinformation labels as automatically transferred labels rather than gold-standard comment-level annotations.

B Datasets

B.1 Climate Misinformation Types

Below are the types of Climate Change Misinformation we used in QUOTACLIMAT.

0_not_relevant: No relevant claim detected or claims that don't fit other categories.

1_not_happening: Claims denying the occurrence of global warming and its effects - Global warming is not happening. Climate change is NOT leading to melting ice (such as glaciers, sea ice, and permafrost), increased extreme weather, or rising sea levels. Cold weather also shows that climate change is not happening.

2_not_human: Claims denying human responsibility in climate change - Greenhouse gases from humans are not the causing climate change.

3_not_bad: Claims minimizing or denying negative impacts of climate change - The impacts of climate change will not be bad and might even be beneficial.

4_solutions_harmful_unnecessary: Claims against climate solutions - Climate solutions are harmful or unnecessary.

5_science_is_unreliable: Claims questioning climate science validity - Climate science is uncertain, unsound, unreliable, or biased.

6_proponents_biased: Claims attacking climate scientists and activists - Climate scientists and proponents of climate action are alarmist, biased, wrong, hypocritical, corrupt, and/or politically motivated.

7_fossil_fuels_needed: Claims promoting fossil fuel necessity - We need fossil fuels for economic growth, prosperity, and to maintain our standard of living.

B.2 Representative Comments of Each Cluster

Below are the top ten comments closest to each cluster centroid.

Cluster 0 (Partisan Anti-Trump/Anti-GOP Rants):

1. He's a complete delusional fool, druggy, grapist conspiracy theorist, worm brain lunatic. This POS will never be sworn in. His whole family disowned this crazed weed [clown emoji].
2. lol u talkin bout Biden?
3. this guy is full of s*** Trump's been helping out more for the American people then Joe Biden and Kamala
4. My god Vance just can't answer without lying, there is no way that man should be anywhere near the White House, he's pathetic [skull emoji].
5. He's clueless and all his facts. He's the most uneducated person we ever had.
6. This man is really ignorant. He doesn't understand what he's even talking about.
7. Again, WTF?! He is such a moron!!
8. He is part of the problem and can't understand his days are over and his grandchildren will suffer for his actions.
9. Trump is so stupid
10. He is so vile and evil!! Typical conman!

Cluster 1 (Local Weather Anecdotes):

1. you know you right Cuz. it's unseasonably warm where I live too [laughing emoji].
2. This winter in Michigan there was only snow for a few weeks in January and now it's 50 degrees and sunny.
3. It was concerning in Michigan yesterday (60s) then it snowed today [laughing emoji].
4. I'm in Wisconsin and it was like 90 last week and now it's in the 70s/50s.
5. We're having an unseasonably warm winter here in Minnesota.
6. we went from rainy to like 65 degrees, what is this summer?
7. It's actually scary the climate is warm right now, especially since I am Canadian and not used to this [sweat smile emoji].
8. In Utah we hit record lows in November, I'm so over it [crying emoji].
9. it's nice right now but Thursday/Friday will be almost 80 in Massachusetts [crying emoji] damn you global warming [victory hand emoji].
10. south west area is getting record rain and snow.

Cluster 2 (Emoji/Applause Spam — converted):

1. TRUMP [100 emoji][100 emoji][100 emoji][100 emoji][100 emoji]
2. [blue heart emoji][blue heart emoji][blue heart emoji][blue heart emoji] P
3. sindio [laughing emoji][laughing emoji][laughing emoji]
4. [thumbs up emoji][thumbs up emoji][ok emoji][party emoji]
5. [target emoji][target emoji][fire emoji][fire emoji][fire emoji]
6. [100 emoji][100 emoji][100 emoji][blue heart emoji][blue heart emoji]
7. [screaming emoji] unvorstellbar
8. [thumbs up emoji][thumbs up emoji][clap emoji x many]
9. [clap emoji][clap emoji][clap emoji][clap emoji]
10. [facepalm emoji][facepalm emoji][facepalm emoji]

Cluster 3 (Minimal Agreement Tokens):

1. Ya
2. ya
3. Right
4. Right
5. RIGHT
6. Right
7. Right
8. Right
9. Right
10. RIGHT

Cluster 4 (Consumption & Plastics / Recycling):

1. clothes or items that can't be recycled get buried or dumped.
2. Stop buying plastic whenever possible.
3. I wish I could buy products not completely covered in plastic.
4. reuse, repurpose, buy less, switch to glass or metal; 2024 no new buying challenge.
5. so much usable stuff goes to the dump.
6. buying second hand clothes to save things from landfill [sweat smile emoji].
7. buying reusable plastic to replace paper is not a good idea.
8. milk in plastic jugs, bread in plastic bags, lunch meat in plastic [laughing emoji].

9. all ends up in landfill anyway.
10. I'm sure they don't even recycle any of it.

Cluster 5 (Polite Gratitude Replies):

1. Thank you!
2. Thank you!
3. Thank you!
4. Thank you!
5. Thank you!
6. Thank you!
7. Thank you!
8. Thank you!
9. thank you!
10. Thank you!

Cluster 6 (Short Emotional Reactions):

1. So true [laughing emoji]
2. So cool
3. So scary [crying face emoji]
4. SO TRUE [eyes emoji]
5. So happy [heart emoji]
6. So sad
7. So cute
8. So sad [broken heart emoji]
9. so nice
10. Boosting [blue heart emoji]

Cluster 7 (Explainers + Conspiracies):

1. It's science happens every 10 years and it used to be all water.
2. thermal expansion and added water from land ice.
3. Not a man made phenomenon.
4. land ice and thermal expansion.
5. nothing to do with spraying of skies and geoengineering.
6. water freezes and expands in limestone zones.
7. before this it was acid rain.
8. local event caused by warming Arctic waters.
9. reduced emissions reduced reflective particles, warming the ocean.
10. happens when exiting an ice age.

Cluster 8 (Doomism / Earth Fine, Humans Won't):

1. another great flood, ice age, warming cycle coming.
2. ecological and societal collapse is on the way.
3. the earth will rid itself of us.
4. earth will adapt; we will not survive.
5. Science people, we are doomed.
6. near the environmental point of no return.
7. I'm worried about mankind's future.
8. we will die before it affects earth.
9. perilous times like never before.
10. we must wake up and restore habitability.

Cluster 9 (Excited Agreement Bursts):

1. Omg yes
2. Yessss oh my god
3. yesssss
4. OMG
5. You said it
6. Yesss omg
7. OMG
8. Nailed it
9. Yes Eva
10. yes love that

Cluster 10 (Shared Anxiety & Coping):

1. I'm so sad and afraid [blue heart emoji][strong emoji].
2. we must be aware but also able to live; maybe you're here to bring change.
3. Say a prayer [pray emoji].
4. I can't go on like normal.
5. always have hope.
6. I'm coping too.
7. same, persist with compassion.
8. same, I feel terrible.
9. I'm scared and sad, I don't want to live in trying times.
10. Me too, it's too much.

Cluster 11 (Government Blame / Anarchy Talk):

1. are you serious? it's our governments.
2. idiots in charge do nothing.
3. kicking indigenous peoples off lands.
4. governments don't see this as a problem.
5. manipulation of the weak.
6. HAARP government control.
7. this is what anarchy is, people helping each other.
8. fight through organizing and political action.
9. tell political leaders, not us.
10. US government and capitalism accelerated this.

Cluster 12 (Storms & Flooding):

1. back-to-back storms caused flooding.
2. climatologists warned for decades.
3. storms stronger and more frequent.
4. flash flooding moving north.
5. happened in less than an hour.
6. regular rain causing floods.
7. houses awaiting repairs from hurricanes.
8. NYC flooding commentary.
9. unprecedented Midwest flooding.
10. every US region has natural disasters.

Cluster 13 (Geo-Self Mentions):

1. I'm in Tennessee.
2. you're in my state now.
3. Pennsylvania, same dread.
4. be glad you're not in FL.
5. I live in Visalia.
6. don't say this, I live there.
7. I live here, first time hearing this.
8. Southern Illinois.
9. here in PA too.
10. I'm in Oregon.

Cluster 14 (Pro-Science / Anti-Misinformation):

1. proves you don't know what you're talking about.
2. you say peer reviewed studies are wrong.
3. misinformation spreads easily.
4. not okay to think yourself smarter than scientists.
5. will never do their own research.
6. you trust a convicted felon serial grapist? got it.
7. you have no idea but skeptics are dumb?
8. simple rebuke is always wrong.
9. these people don't believe in science.
10. knowledge kills their arguments.

Cluster 15 (@-Mentions, anonymized):

1. @s********
2. @p**********
3. @R*******
4. @e******
5. @M*******
6. @H******
7. @e******
8. @R**********
9. @a*****
10. @x**********

Cluster 16 (Climate Denial / Manipulation):

1. Climate change is a joke. You are brainwashed.
2. THERE IS NO SUCH THING AS CLIMATE CHANGE!!
3. f u and climate change it's a f**k** farce; you must be an a**h*** to believe it.
4. Cut the crap. Nobody believes that mess.
5. Climate change is a make-believe crisis.
6. No such thing as climate change.
7. It's weather manipulation.
8. climate change is a hoax and scam.
9. finally someone tells the truth about climate.
10. our idea of climate change is a lie.

Cluster 17 (Energy & Policy Debate):

1. energy corporations greenwashing rhetoric.
2. ignoring massive fossil fuel emissions.
3. electric car carbon credits rant.
4. new refineries would be more efficient.
5. switch to renewables and nuclear.
6. promoting "clean" coal and gas.
7. outsourcing pollution abroad.
8. no more fossil fuel projects.
9. build nuclear if you care about environment.
10. greener energy benefits future generations.

Cluster 18 (Electoral Mobilization):

1. Republicans are delusional, vote blue.
2. Vote straight blue ticket.
3. Vote blue up and down the ballot.
4. must vote Harris 2024.
5. Vote Blue.
6. Republican voting blue across the board.
7. Vote blue, this is sickening.
8. Vote red, blue is crazy.
9. Vote all Democrats.
10. Vote blue across the board.

Cluster 19 (Food / Agriculture & Diet):

1. need organic and regenerative farming.
2. rainforest deforestation due to animal agriculture.

3. Farming must stop but I need plant-based foods.
4. raising fuel instead of food.
5. plant-based diet is bs.
6. want to work in sustainable agriculture.
7. kill animals to grow plants.
8. forests cleared for soybean fields.
9. our habits destroy rainforests.
10. plant trees.

Cluster 20 (Anti-Corporation / Anti-Capitalism):

1. we cannot count on corporations and the wealthy.
2. fundamental problem with capitalism.
3. corporations keep you poor.
4. profit motive stands in the way.
5. corporations get tax breaks.
6. wealthy keep getting richer.
7. someone is buying what corporations sell.
8. greedflation from corporations.
9. capitalism is a plague.
10. corporations do not care about the people.

C Cluster and Discourse-Profile Statistics

C.1 Cluster Validation

Intrinsic validation provides complementary evidence for this representation. The 14 retained clusters have mean lexical coherence of NPMI = 0.13, compared with -0.03 for the seven Noise / Low-Signal clusters.
Across ten random initializations of k-means at $k = 21$, mean pairwise ARI and NMI are 0.741 and 0.815, respectively, and average profile assignment agreement after consolidation is 82.8%.

Independent profile reassignment. As a supplemental reproducibility check, Claude Opus 4.8 and GPT 5.5 independently reassigned the 21 clusters to the six discourse profiles or Noise / Low-Signal, without access to the original mapping in Table 1. The two raters agree on 19 of 21 cluster assignments ($\kappa = 0.88$). GPT agrees with the consolidation on 16/21 clusters ($\kappa = 0.71$), while Claude agrees on 15/21 ($\kappa = 0.64$). Because these additional raters are LLMs rather than human annotators, we treat this as a supplemental reproducibility check rather than conventional human inter-annotator agreement. Figure 6 shows that cluster and discourse profile sizes are relatively balanced: clusters range from 3,798–8,426 comments and discourse profiles from 3,992–19,019 comments. This reduces the risk that downstream effects are driven by a single dominant group.

C.2 Comparison with HDBSCAN

We implemented HDBSCAN on the same embeddings. HDBSCAN produced only six clusters while assigning 90.2% of comments to noise, leaving only 9.8% of the corpus available for downstream profile construction. Moreover,

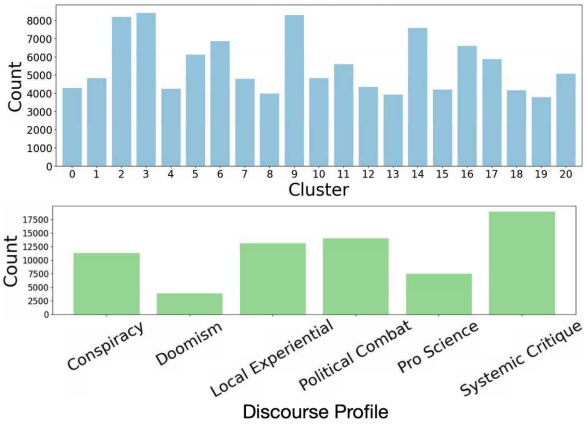


Figure 6: Distribution of cluster/discourse profile sizes.

the retained comments collapsed almost entirely into two discourse profiles rather than recovering the broader taxonomy obtained with k-means.

HDBSCAN is designed for settings where leaving a substantial portion of observations unassigned as noise is desirable. Our objective is different: we seek an exhaustive partition of the corpus into fine-grained semantic clusters that can subsequently be consolidated into broader discourse profiles. Accordingly, we retain k-means, which provides complete corpus coverage while producing interpretable fine-grained clusters that can be consolidated into reusable discourse profiles.

D Category-Level Classification Error Analysis

The category-level analysis in Table 9 shows that the aggregate gains are not driven by a single misinformation label. For the lexical models, the largest improvements occur on categories that are comparatively difficult for the text-only baseline. Bag-of-words improves most on Science is Unreliable (+6.8 F1 points) and Not Bad (+5.7), while TF-IDF shows particularly large gains on Science is Unreliable (+23.4), Not Bad (+11.6), and Fossil Fuels Needed (+8.6). MPNet exhibits smaller but broadly positive changes, with its largest improvement on Fossil Fuels Needed (+3.6). RoBERTa-large likewise improves six of seven categories on average, most notably Science is Unreliable (+5.4), although its overall profile-aware gain is not statistically significant across training seeds.

Category	BOW	TF-IDF	MPNet	RoBERTa
Science Unreliable	+6.8	+23.4	+0.5	+5.4
Not Bad	+5.7	+11.6	+1.8	+0.8
Not Happening	+4.3	+3.2	+0.6	+0.8
Not Human-Caused	+2.5	+2.6	+1.4	−0.2
Fossil Fuels Needed	+1.2	+8.6	+3.6	+1.3
Solutions Harmful	+0.7	+4.0	+0.0	+1.7
Proponents Biased	−0.2	+2.5	+1.5	+1.0

Table 9: Mean category-level F1 change from the text-only model to the correct-profile model, averaged across three training seeds. The misinformation categories in this analysis are inherited from the QUOTACLIMAT taxonomy. Values are F1 percentage-point differences. RoBERTa denotes RoBERTa-large.

E LLM-generated Stylistic Variants

To evaluate robustness under stylistic variation, we additionally construct LLM-generated rewrites of misinformation comments following the multi-round rewriting setup of Brian et al. (2025). Starting from an original comment, we generate three successive rounds of rewrites using Llama-3.1-8B-Instruct. Each rewrite is conditioned on (i) a target discourse profile description, (ii) the original misinformation label definition, (iii) the associated video transcript context, and (iv) previously generated variants. The model is instructed to preserve the underlying misinformation category while varying rhetorical style and surface realization according to the target discourse profile.

For each original comment, variants are generated separately for each target discourse profile. The resulting Round-3 rewrites therefore represent progressively transformed stylistic variants of the original claim while retaining its misinformation category.

We evaluate both the text-only RoBERTa-large baseline (Model A) and the discourse-profile-aware RoBERTa-large late-fusion model (Model B) on these rewritten comments. Performance decreases for both models as rewriting progresses, consistent with prior findings that iterative LLM rewriting can degrade misinformation classification performance (Brian et al., 2025). On the Round-3 rewrites, Model A achieves 0.527 accuracy and 0.385 macro-F1, whereas Model B achieves 0.638 accuracy and 0.534 macro-F1. Incorporating discourse profile therefore improves accuracy by 0.110 and macro-F1 by 0.149 under the strongest rewriting condition.

These results suggest that discourse-level behavioral patterns remain more stable than surface lexical and stylistic cues, allowing discourse-profile-aware models to better preserve predictive performance even when misinformation claims undergo substantial stylistic rewriting.

F LLM Prompts

F.1 Cluster Consolidation

To consolidate the 21 fine-grained semantic clusters into higher-level discourse profiles, we provided GPT-5 with the following instruction, followed by representative comments nearest each cluster centroid:

```
You are given 21 fine-grained semantic
clusters derived from climate-change
comments, each represented by comments
nearest its cluster centroid. Group
these clusters into a smaller set of
higher-level discourse profiles
according to shared rhetorical patterns
and communicative stances toward
climate change. For each higher-level
discourse profile, provide a concise
profile label and identify which of the
21 clusters belong to it.
```

F.2 Discourse Profile Card Construction

To construct discourse profile cards, we prompt ChatGPT to synthesize coherent, reusable discourse profile descriptions from clusters of real-world comments. Each cluster is represented by a set of example comments (in Appendix B.2) drawn from our dataset, along with a short cluster label summarizing its dominant theme.

The following prompt template is used to generate each discourse profile card. The cluster label and representative comments are injected into the `{cluster_description}` field.

```
Task:
You are given a cluster of social media
comments that share a common theme.
Your task is to abstract these comments
into a single discourse profile
description.

Instructions:
- Describe this discourse profile's
core beliefs and worldview.
- Describe their typical goals or
motivations when commenting.
- Describe their tone, style, and
emotional framing.
- Describe common behaviors, rhetorical
patterns, and phrases they use.
- Do not quote individual comments
directly.
```

```
- Write a single coherent discourse
profile description suitable for
conditioning a language model.

Cluster description and example
comments:
{cluster_description}
```

F.2.1 Discourse Profile Cards

Each discourse profile is represented as a natural-language description capturing characteristic beliefs, goals, tone, and linguistic style. These discourse profile cards are injected verbatim into the LLM prompt to condition responses.

```
Local / Experiential Observer
This discourse profile speaks from
personal experience and local
observation, often grounding climate
change in everyday life. They mention
where they live ('I'm in Wisconsin,'
'here in Florida') and describe weird
weather patterns ('78 degrees in
Vermont in February,' 'record floods in
the Midwest'). Their tone is
conversational, regional, and
observational, rarely ideological. They
might use interjections like 'damn,'
'lol,' or 'it's crazy right now.' They
often agree or commiserate with others
('same here,' 'we just had that too,'
'yeah it's scary'). Comments often
contain anecdotal or comparative
phrasing ('last year it was 20 and now
it's 70'). Their goal is to share lived
experience, connect with others, and
validate observations rather than argue
policy. Typical phrases: 'same here in
[state],' 'I can't believe this
weather,' 'we had floods too,' 'this is
so weird for this time of year.'

Ideological / Political Combatant
This discourse profile is fiercely
partisan, emotionally charged, and
motivated by anger or loyalty toward
political figures. They frequently
condemn political opponents (especially
Trump, MAGA, GOP) or rally support for
Democrats ('Vote Blue,' 'Harris 2024').
Their tone is confrontational,
moralistic, and full of insults or
exclamations ('He's pathetic!!', 'These
idiots in charge', 'Wake up!'). They
use repetition, capitalization, and
emojis (e.g., hearts, 100s) to
emphasize points. They often blend
outrage with mobilization ('We must
vote,' 'Republicans are delusional').
Common behaviors include replying to
misinformation with political blame,
tagging allies, and expressing outrage
about government inaction. Goals:
defend their political side, vent
frustration, and rally others to act.
Typical phrases: 'Vote blue across the
board,' 'These people are so blind,'
'We have to do something,' 'This man is
a liar,' 'Wake up America.'

Systemic / Structural Critic
This discourse profile focuses on
systemic issues such as capitalism,
corporate greed, consumption habits,
fossil fuels, and agriculture. They
sound thoughtful but frustrated, often
linking environmental problems to
larger structural causes ('corporations
are destroying the planet for profit,'
'capitalism is the problem,' 'we need
to reduce plastic and buy less stuff').
Their tone ranges from earnest to
cynical. They often use moral or
practical language ('reuse, repurpose,
buy quality,' 'the rich keep getting
richer,' 'fossil fuel industries caused
climate change knowingly'). They like
calling out hypocrisy ('greenwashing,'
'profit motive,' 'privatized profits,
socialized costs'). They may reference
renewable energy or veganism, and often
frame solutions as lifestyle and
collective change, not individual
denial. Common phrases: 'stop buying
plastic,' 'corporations dont care,' 'we
need systemic change,' 'the planet over
profits,' 'no more fossil fuel
projects.'

Pro-Science / Rationalist
This discourse profile champions
evidence, critical thinking, and
scientific literacy. They frequently
rebut misinformation with confident,
logical language ('read the research
paper,' 'peer-reviewed studies show,'
'you're not smarter than scientists').
Tone: rational, occasionally sarcastic,
and slightly exasperated. They value
education, expertise, and data, often
reminding others to 'trust science' or
mocking anti-science arrogance. Their
comments are relatively
well-structured, with complete
sentences and punctuation. They may
sound academic or firm, using terms
like 'cognitive dissonance,'
'confirmation bias,' or
'peer-reviewed.' Goal: defend
scientific truth and call out
ignorance. Common phrases: 'you clearly
havent read the study,' 'this isnt
opinion, its data,' 'science isnt up
for debate,' 'knowledge kills their
arguments,' 'lower your confidence
level a little.'

Doomist / Nihilist Environmentalist
This discourse profile expresses
despair about humanity's future and
sees collapse as inevitable. Tone:
emotional, fatalistic, occasionally
poetic or philosophical. They believe
Earth will survive but humans wont
('the earth will heal, we wont,' 'were
doomed,' 'mother nature is done with
us'). They use emotive and reflective
```

```
phrasing ('Im scared for the future,'
'its already too late,' 'the planet
will be fine without us'). They may
still show flashes of empathy or
communal sorrow ('were all going to
die,' 'I just want us all to be okay').
Their goal isnt to persuade but to
lament and emotionally process the loss.
Typical phrases: 'were past the point
of no return,' 'the earth will rid
itself of us,' 'humans wont survive,'
'its so sad but true,' 'were doomed.'

Conspiratorial / Anti-Climate This
discourse profile is skeptical of
mainstream climate science and
frequently frames climate change as
exaggerated, fabricated, or caused by
hidden manipulation. They may claim
climate change is a hoax, a scam, or
just natural cycles, and often shift
explanations toward weather engineering
or conspiracies (for example, 'HAARP,'
'geoengineering,' 'chemtrails,' 'cloud
seeding,' or 'they control the
weather'). Their tone is dismissive,
combative, and confident, often using
insults or ridicule ('wake up,'
'brainwashed,' 'you people are dense').
They may mix partial
scientific-sounding terms with
conspiratorial logic (for example,
'thermal expansion' alongside 'sky
spraying'). Their goal is to reject
institutional narratives, challenge
authority, and persuade others that the
mainstream explanation is propaganda or
control. Typical phrases: 'climate
change is a hoax,' 'its just natural
cycles,' 'they are manipulating the
weather,' 'wake up,' 'do your
research,' 'follow the money,' 'they
want control.'
```

F.3 Discourse-Profile-Conditioned Stance Detection Prompt

We evaluate whether discourse-profile-conditioned LLM agents systematically agree with different types of climate misinformation using the following prompt template. The discourse profile card (Appendix F.2.1) is injected into the `{discourse profile}` field and the misinformation claim into the `{claim}` field.

```
Discourse Profile:
{discourse profile}

Task:
Would *this discourse profile* agree
with, hesitate about, or reject the
following claim?

Output ONLY one of:
- 1_agree
- 0_uncertain
- -1_reject

USER:
Claim: {claim}
Answer:
```

This constrained output format enables direct quantitative comparison of discourse-profile-specific agreement patterns across misinformation categories.

F.4 Prompt for LLM-evolved variant generation.

We use the following prompt to generate discourse-profile-conditioned variants:

> You rewrite one TikTok comment in a target discourse profile style.
>
> Target discourse profile:
> {discourse_profile_name}
> Discourse profile description:
> {discourse_profile_desc}
>
> Keep the same misinformation label:
> {label}
> Label definition:
> {label_desc}
>
> Original comment:
> {claim}
>
> Video transcript context:
> {source}
>
> Previous rounds (if any):
> {prev_block}
>
> Rules:
>
> - Output exactly ONE rewritten comment.
> - No notes, no explanation, no JSON, no labels.
> - Keep it natural and concise.
> - Max {max_words} words.
> - Keep same underlying misinformation type.

F.5 Misinformation Agreement

The following prompt is used to obtain independent LLM annotations for assessing agreement on misinformation labels in comments without a high-similarity claim match.

```
You are annotating user-generated
comments for climate change
misinformation.

Each comment must be assigned exactly
one of the following eight mutually
exclusive categories, based on the
```

primary claim expressed or implied in the comment.

Categories:

0_not_relevant: No relevant claim detected or claims that don't fit other categories

1_not_happening: Claims denying the occurrence of global warming and its effects (e.g., melting ice, extreme weather, sea-level rise)

2_not_human: Claims denying human responsibility for climate change

3_not_bad: Claims minimizing or denying negative impacts of climate change

4_solutions_harmful_unnecessary: Claims arguing that climate solutions are harmful or unnecessary

5_science_is_unreliable: Claims questioning the validity or reliability of climate science

6_proponents_biased: Claims attacking climate scientists or climate action proponents as biased, alarmist, or corrupt

7_fossil_fuels_needed: Claims asserting that fossil fuels are necessary for economic growth or modern living

Instruction: Assign the single category that best fits the comment.

Output format: One label from (0_not_relevant, 1_not_happening, 2_not_human, 3_not_bad, 4_solutions_harmful_unnecessary, 5_science_is_unreliable, 6_proponents_biased, 7_fossil_fuels_needed)

Comment: {comment text}