Modeling Commenter Personas in Climate Misinformation Discourse

Anonymous ACL submission

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 stable speaker-level patterns, where the same individuals repeatedly employ characteristic narratives and rhetorical styles across interactions. We address this gap by introducing personas: data-driven groupings of commenters that capture recurring discourse strategies beyond the content of any single comment. Empirically, modeling persona improves misinformation classification performance beyond text-only baselines in low-capacity and resource-constrained settings, increasing accuracy from 0.79 to 0.81 and macro-F1 from 0.72 to 0.74 in a low-capacity MLP classifier (statistically significant, $p = 0.021$). Beyond detection, we find that personas shape how misinformation is received and propagated. Persona-conditioned language model agents exhibit systematic differences in misinformation acceptance that align with human persona-level susceptibility, and the relationship between misinformation prevalence and content persistence varies by the dominant conversational persona.

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., 2021; 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.

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 speaker-level regularities: 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.

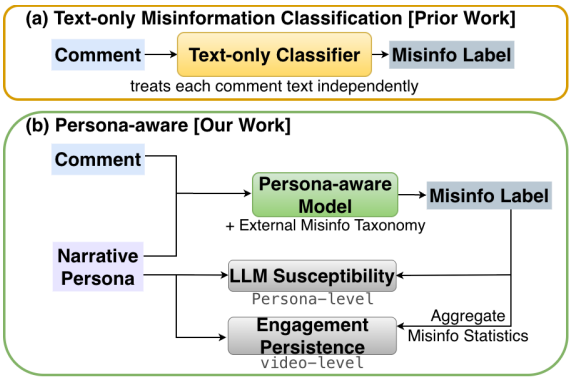


Figure 1: Comparison of text-only vs. persona-aware misinformation modeling. (a) Prior work (Islam et al., 2020; Coan et al., 2021; Upadhyaya et al., 2023; Choi et al., 2023; Rojas et al., 2024) typically predicts a misinformation category from comment text alone, treating each comment independently.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

We introduce *personas* to address this challenge: a compact, data-driven abstraction of recurring narrative frames and rhetorical styles in climate discourse. A growing body of work supports the importance of speaker and social context for misinformation and stance modeling. Incorporating user and interaction structure can improve misinformation detection beyond text alone, encouraging representations that aggregate signal across interactions rather than treating each instance independently (Nguyen et al., 2020; Mehta et al., 2022).

Stance research similarly argues that beliefs and framing are more faithfully modeled at the author level and are not reliably captured by lexical features alone (Allaway and McKeown, 2021; Zhang et al., 2024). Communication scholarship further emphasizes that disinformation propagates through identity linked, adversarial narratives that remain stable within groups and are expressed via characteristic rhetorical strategies (Díaz Ruiz and Nilsson, 2023). Together, these perspectives suggest that *who is speaking* provides a complementary axis for understanding climate misinformation beyond the words in a single comment.

Rather than relying on predefined demographics or static author embeddings, we induce personas directly from large-scale, in-the-wild TikTok climate comments by clustering semantic comment embeddings and consolidating clusters into interpretable persona types. This persona representation is then used in three ways (Figure 1b): (i) to augment fine-grained misinformation-type prediction, with a focus on whether persona provides additional signal in low-capacity or limited-supervision settings, (ii) as a scaffold for a controlled susceptibility probe using persona-conditioned LLM agents[1], and (iii) as an organizing variable for studying how misinformation relates to engagement persistence in comment threads.

Our main contributions are:

- We induce commenter personas from TikTok climate discourse and compress large-scale clusters into interpretable personas.
- We show that persona context improves fine-grained misinformation-type prediction in low-resource settings. For a low-capacity MLP classifier, adding persona features improves accuracy from 0.79 to 0.81 and macro-F1 from 0.72 to 0.74 ($p = 0.021$).
- We evaluate persona-conditioned LLM agents in a controlled susceptibility setting and show strong correlations with human persona-level misinformation susceptibility.
- We show that the relationship between misinformation and engagement persistence is persona-conditional: misinformation amplifies or suppresses persistence depending on the dominant conversational persona.

For the classification component, our goal is not to establish a new state-of-the-art classifier, but to isolate and quantify the contribution of persona-level signal under controlled modeling settings.

[1]In this work, we use GPT-5 as our LLM agent.

2 Related Work

2.1 Climate Misinfo/Obstruction Narratives

Prior work has documented that climate misinformation extends beyond outright denial to include more subtle obstruction strategies, such as questioning the feasibility of solutions, emphasizing economic trade-offs, or attacking the credibility of scientific messengers (Lamb et al., 2020; Coan et al., 2021; Holder et al., 2023). These studies motivate the use of fine-grained misinformation taxonomies rather than binary true/false labels. We adopt this perspective by leveraging the QUOTACLIMAT taxonomy (QuotaClimat, 2025), which captures diverse climate obstruction narratives. The fine-grained misinformation types in QUOTACLIMAT and detailed explanations of the categories are provided in Appendix B.1.

However, existing work primarily treats misinformation as isolated textual instances, without modeling stable speaker-level regularities in how these narratives are deployed. In contrast, our work explicitly models recurring rhetorical patterns through data-driven personas and examines how different personas preferentially express specific misinformation types in real-world discourse.

2.2 Persona/Author context in NLP

In social NLP tasks such as stance and misinformation detection, text alone is often insufficient as it lacks context about the author's underlying beliefs. Prior work by Benton and Dredze (2018a) shows that stance expressed in a tweet depends on latent author beliefs, and that incorporating auxiliary author embeddings improves stance classification.

A growing body of work shows that incorporating author or persona context improves stance and misinformation detection by capturing latent beliefs and communicative styles (Benton and Dredze, 2018a,b). These approaches demonstrate that who is speaking can provide signal beyond text alone.

Our work differs in two key ways. First, rather than relying on predefined author attributes or static user embeddings, we induce personas directly from large-scale, in-the-wild climate discourse. Second, we use persona not only as an auxiliary feature for prediction, but as an organizing abstraction for analyzing misinformation strategies, susceptibility, and engagement dynamics.

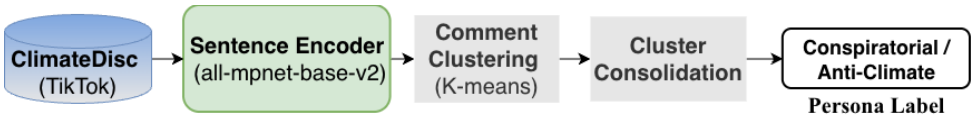


Figure 2: Offline persona induction: encode comments, cluster embeddings, and manually consolidate clusters to assign each comment a persona label p_i (Section 3.1).

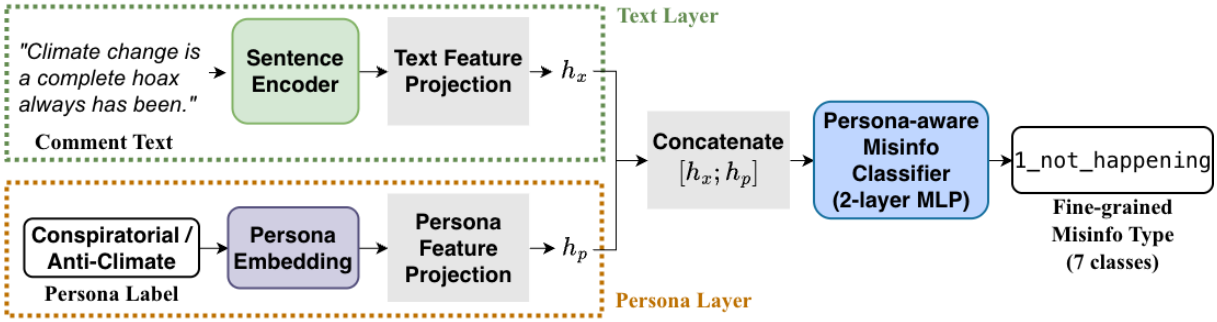


Figure 3: Persona-aware misinformation modeling: obtain fine-grained misinformation-type labels and evaluate prediction models that fuse text and persona representations (Section 3.2). Details of how we obtain comment-level misinformation labels are in Appendix A.

2.3 LLM Personas & Engagement Dynamics

Recent studies show that large language models can be conditioned to simulate behavior associated with specific personas or subpopulations (Park et al., 2023; Krishna et al., 2022; Argyle et al., 2023; Li et al., 2024). We build on this line of work by using persona-conditioned LLM agents as a controlled evaluation tool for climate misinformation susceptibility, rather than as a behavioral simulator per se. We show that persona-level susceptibility patterns in LLM agents align closely with those observed among real human commenters.

Separately, prior work has examined how misinformation diffusion and persistence depend on content properties, emotional framing, and network structure (Vosoughi et al., 2018; Liu et al., 2023; Grinberg et al., 2019; Sun and Xie, 2024). While these studies highlight substantial heterogeneity in engagement, they do not examine whether such dynamics are systematically conditioned on conversational personas. We address this gap by demonstrating persona-dependent persistence effects in climate misinformation discussions on TikTok.

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 stable speaker-level regularities. For each comment c_i authored by user u_i, we aim to (i) assign a narrative persona p_i capturing recurring rhetorical patterns, (ii) obtain a fine-grained misinformation type y_i when applicable, and (iii) analyze how persona conditions misinformation susceptibility and engagement persistence at both the comment and video levels. Dataset details are provided in Section 4.1.

Our pipeline proceeds in two stages. First, we induce narrative personas from large-scale climate comment discourse (Figure 2; Section 3.1). Second, we incorporate persona 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 a curated climate misinformation taxonomy; since the transfer procedure is not central to our main contributions, we defer its details to Appendix A.

3.1 Persona Creation

To capture stable rhetorical and ideological patterns across commenters used in section 3.2, 3.3, 3.4, we induce narrative personas from large-scale climate change comment discourse in CLIMATEDISC (Gao et al., 2025), a large-scale corpus of TikTok climate change videos and associated comments (section 4.1), which serve as a speaker-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

Local / Experiential Observers 1. Local Weather Anecdotes 12. Storms & Flooding 13. Geographical Self-Mentions	**Ideological / Political Combatants** 0. Anti-Trump / Anti-GOP / Anti-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. "Humans Won't Make It" Doomism	**Conspiratorial / Anti-Climate** 7. Conspiracy / Muddled Explainers 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: Narrative personas obtained by grouping the 21 k-means clusters of comment embeddings into semantically coherent higher-level categories. The Noise / Low-Signal group is excluded from downstream persona-sensitive analyses.

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 4.2.1).

To obtain a stable speaker-level abstraction for downstream modeling, we manually validate and consolidate the 21 fine-grained clusters into six higher-level *narrative personas* reflecting shared communicative stances toward climate change, along with a residual "Noise / Low-Signal" group excluded from persona-sensitive analyses. The resulting persona definitions are reported in Table 1.

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

3.2 Persona-aware Misinfo Classification

We compare a text-only baseline with a persona-aware model to test whether persona provides signal beyond comment text (Figure 3).

Text-only baseline. Model A is a lightweight feed-forward classifier operating over sentence-level semantic representations. Given a comment t in CLIMATEDISC, we first encode it using a pretrained sentence transformer based on MPNet (Reimers and Gurevych, 2020): $\mathbf{x} = \phi(t)$, where $\phi(\cdot)$ maps text to a fixed-dimensional dense vector representation. The encoded representation is then passed through a nonlinear projection, $\mathbf{h}_x = f_x(\mathbf{x})$, where $f_x(\cdot)$ denotes an affine transformation followed by a ReLU nonlinearity and dropout. Finally, a linear classifier produces logits over fine-grained misinformation types: $\hat{\mathbf{y}} = g(\mathbf{h}_x)$, where $\hat{\mathbf{y}}$ corresponds to a seven-class (excluding 0_not_relevant) misinformation taxonomy from QUOTACLIMAT.

Persona-aware model. Model B extends the text-only baseline by incorporating persona information as an additional latent signal. Each comment is associated with a persona label $c \in \{1, \ldots, 6\}$, which is mapped to a learnable persona embedding: $\mathbf{p} = E(c)$, where $E(\cdot)$ denotes a trainable embedding function. The persona embedding is transformed via a nonlinear projection, $\mathbf{h}_p = f_p(\mathbf{p})$, analogous in form to the text projection $f_x(\cdot)$.

The text and persona representations are combined via concatenation,

$$\mathbf{h} = [\mathbf{h}_x; \mathbf{h}_p]$$

and passed through a joint nonlinear transformation, $\mathbf{z} = f_j(\mathbf{h})$, before being mapped to output logits by the same classifier $g(\cdot)$: $\hat{\mathbf{y}} = g(\mathbf{z})$.

This late-fusion formulation isolates the contribution of persona information by introducing $\mathbf{p}$ as an additive latent variable alongside the fixed text representation $\mathbf{x}$.

We also apply the same formulation using RoBERTa-large representations. We include both lightweight and large pretrained models to examine how the utility of persona varies with text representational capacity, rather than optimize performance.

While this evaluation provides predictive value of persona, it does not address whether the induced personas capture meaningful differences in how misinformation is received or accepted.

3.3 Persona-Level Misinformation Susceptibility with LLM

We use persona-conditioned LLM agents to evaluate whether inferred personas capture systematic differences in misinformation susceptibility. We test whether an LLM prompted with persona descriptions (Appendix C.2) exhibits misinformation acceptance patterns mirroring user comments.

We construct a balanced evaluation set of 350 misinformation claims from the QUOTACLIMAT dataset, sampling 50 claims from each of the seven categories. For each inferred persona, we create a short persona card summarizing characteristic beliefs and rhetorical tendencies (Appendix C.1).

For persona p, we define an LLM susceptibility score as the proportion of misinformation claims not rejected by the persona-conditioned

LLM agent:

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

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

Human–LLM Persona Alignment We assess whether persona-conditioned LLM susceptibility patterns align with those of real human personas. For each persona p, we compute a human misinformation rate H_p, defined as the proportion of comments authored by p that are labeled as misinformation in CLIMATEDISC. We compute the LLM susceptibility score L_p previously defined. We then measure alignment between (H_p) and (L_p) using Pearson and Spearman correlations, where higher correlation indicates closer correspondence between human and LLM persona behavior.

We next examine whether these persona-level differences translate into distinct engagement trajectories.

3.4 Persona-Level Temporal Dynamics

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

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

We test whether the relationship between misinformation and persistence varies by persona by stratifying videos into low, mid, and high misinformation tertiles based on misinformation rate. Within each persona type, we conduct Kruskal–Wallis tests to determine whether persistence differs significantly across misinformation levels.

[2]We exclude videos whose dominant persona is classified as *Noise / Low-Signal*. We also exclude videos with zero or one comment, since persistence requires at least two comments to be defined.

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 persona-conditioned LLM agents using a balanced set of 350 claims (50 per category; Section 3.3).

CLIMATEDISC For persona 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) persona creation (Section 3.1), (ii) persona-aware misinformation classification on in-the-wild comments (Section 3.2), (iii) human persona-level misinformation rates H_p for LLM alignment (Section 3.3), and (iv) temporal dynamics and engagement persistence (Section 3.4).

For the persona-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.

4.2 Results

4.2.1 Persona Landscape

Figure 4 reports WCSS and Silhouette scores for $k \in [3, 30]$. WCSS decreases smoothly with k (no decisive elbow), while the Silhouette score attains a local maximum at $k = 21$, indicating comparatively strong separation at this granularity. We therefore set $k = 21$ as a practical trade-off be-

tween separation and interpretability. As a qualitative check, Appendix B.2 lists the ten comments nearest each centroid, which exhibit consistent topical and rhetorical patterns.

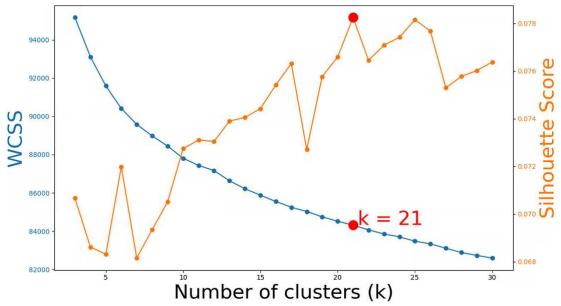


Figure 4: Within-Cluster Sum of Squares (WCSS; lower is better) and Silhouette Score (higher is better) across varying numbers of clusters k. The local Silhouette maximum and WCSS elbow around $k = 21$ motivate our choice of 21 clusters.

For interpretability in downstream analyses, we aggregate the 21 clusters into six higher-level narrative personas plus a residual "Noise / Low-Signal" group. We exclude the Noise group from persona-sensitive analyses (it consists largely of short acknowledgments, emojis, and @-mentions). Table 1 summarizes the mapping.

Figure 5 shows that cluster and persona sizes are relatively balanced: clusters range from 3,798–8,426 comments and personas from 3,992–19,019 comments. This reduces the risk that downstream effects are driven by a single dominant group.

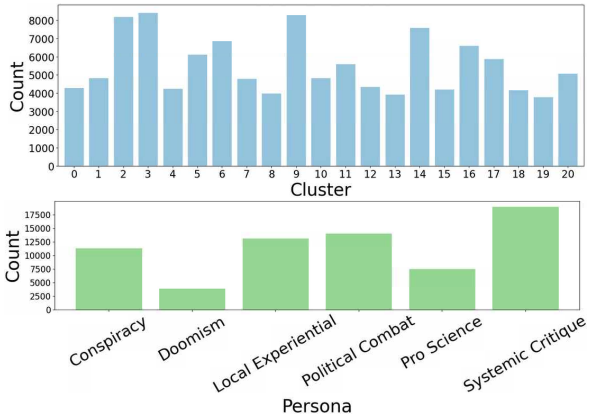
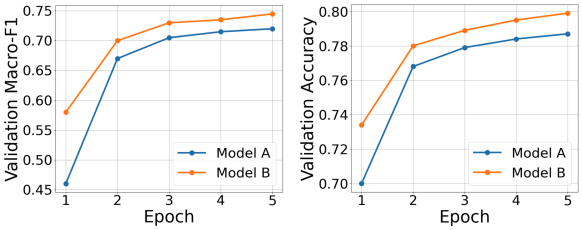


Figure 5: Distribution of cluster/persona sizes.

4.2.2 Text-Based Misinformation Classification with/without Persona

We evaluate whether persona information provides predictive signal beyond text alone (section 3.2).

For the MLP models, incorporating persona improves accuracy from 0.79 to 0.81 and macro-F1 from 0.72 to 0.74. The accuracy gain is statistically significant under an exact paired McNemar test ($p = 0.021$). As shown in Figure 6, the persona-aware model also converges more rapidly in early training, achieving substantially higher validation accuracy and macro-F1 after the first epoch.



F1 Accuracy

Figure 6: Validation accuracy and F1 across epochs for text-only (A) and persona-aware (B) MLP models.

Using RoBERTa-large, persona yields smaller improvements, increasing accuracy from 0.82 to 0.83 and macro-F1 from 0.76 to 0.78; however, this difference is not statistically significant ($p = 0.11$).

Overall, these results suggest that persona information provides complementary signal in lower-capacity settings, where the text representation alone may be insufficient to capture speaker-level regularities. In contrast, larger pretrained models such as RoBERTa appear able to implicitly encode persona-related cues from text, suggesting a representational ceiling beyond which explicit persona modeling yields diminishing returns.

4.2.3 Persona-Level Susceptibility Probing with LLM

Human vs. LLM Agents Susceptibility We observe a strong positive correlation between human and LLM persona susceptibility patterns. The Pearson correlation between H_p and L_p (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 personas by misinformation susceptibility is highly consistent between humans and LLM agents. This alignment suggests that persona-conditioned LLM agents capture meaningful, persona-specific structure in misinformation acceptance.

Susceptibilities Across Personas Table 2 summarizes overall susceptibility score (specified in section 3.3) per persona. *Conspiratorial / Anti-Climate* shows very high susceptibility (96.3%), while other personas 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 persona demonstrates comparatively low susceptibility (10.0%), 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.

Persona	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 2: Overall misinformation susceptibility rate by persona.

4.2.4 Persona-Stratified Engagement Dynamics

We investigate whether misinformation affects content persistence uniformly, or whether this relationship is conditioned on the dominant persona.

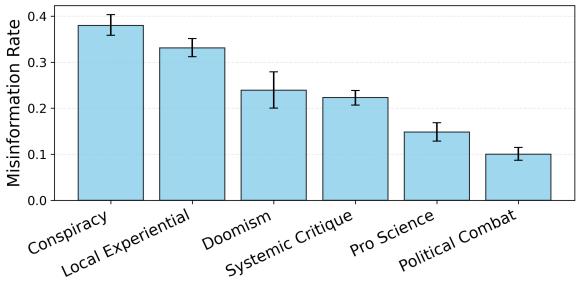


Figure 7: Misinformation rates across persona types (excluding Noise). Bars represent mean proportion of comments containing misinformation for each persona, with error bars showing 95% bootstrap confidence intervals (2,000 iterations).

Misinformation Rates Across Personas Before examining temporal dynamics, we first characterize the baseline misinformation production rates for each persona type. As shown in Figure 7, substantial heterogeneity emerges across persona types. Conspiracy personas exhibit the highest misinformation rate, indicating that approximately 38% of comments from this persona contain misinformation. The Pro Science persona, characterized by explicit appeals to scientific authority, still produces misinformation in 15% of comments. This finding indicates that scientific self-identification alone does not guarantee resistance to misinformation. The nearly fourfold difference between the highest (Conspiracy) and lowest (Political Combat) personas establishes that communicative style strongly predicts misinformation production.

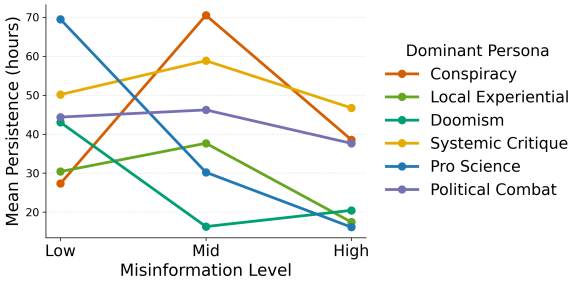


Figure 8: Mean persistence hours across misinformation level (Low, Mid, High) for each persona type.

Persistence Across Personas To examine persona-conditional relationships, we bin videos into three levels (Low, Mid, High) based on misinformation rate and compute mean persistence hours for each persona-by-bin combination. Figure 8 summarizes these persona-conditioned persistence patterns. Pro Science Personas show a strong negative relationship between misinformation and persistence. Videos dominated by Pro Science comments persist substantially longer when misinformation rates are low (69.5 hours) compared to high (16.1 hours). Conspiracy Personas exhibit the opposite pattern. Persistence increases from low-misinformation tertiles (27.3 hours) to high-misinformation tertiles (38.9 hours), even longer persistence at mid-misinformation rate. Systemic Critique Personas exhibit moderate variation across misinformation levels, with persistence remaining within a relatively narrow range compared to other personas. Doomism and Local Experiential Personas show negative relationships similar to Pro Science, with persistence declining at higher misinformation levels. Doomism drops from 43.1 hours (Low) to 20.3 hours (High), a 53% reduction. Local Experiential shows more variability, with an initial increase from Low (30.5 hours) to Mid (37.3 hours) before declining to High (16.8 hours).

To assess whether these observed patterns reflect statistically significant heterogeneity rather than sampling variability, we conduct within-persona Kruskal-Wallis tests examining whether persistence differs across misinformation levels. Table 3

Persona	H Statistic	p-value
Systemic Critique	20.501	$< 0.001^{***}$
Local Experiential	16.336	$< 0.001^{***}$
Conspiracy	9.237	0.010^{**}
Pro Science	4.875	0.047^{*}
Political Combat	3.850	0.146
Doomism	2.912	0.233

Table 3: Within-persona Kruskal–Wallis test results examining whether persistence differs across misinformation tertiles (Low, Mid, High). Significance levels are denoted as $^{***}p < 0.001$, $^{**}p < 0.01$, and $^{*}p < 0.05$.

presents these results. Four of six personas exhibit statistically significant differences in persistence across misinformation levels at $\alpha = 0.05$. These include Systemic Critique, Local Experiential, Conspiracy, and Pro Science personas, indicating that misinformation prevalence conditions engagement persistence for these communicative styles.

5 Discussion and Conclusion

This work demonstrates that persona constitutes a missing and consequential axis in climate change misinformation research. By modeling who is speaking in addition to what is said, we show that persona provides complementary structure that is not captured by text alone.

The relatively balanced distribution of personas across misinformation types suggests that climate misinformation discourse on TikTok is not dominated by a single narrative or speaker type. This diversity matters for modeling, as it implies that effective detection and intervention strategies must account for multiple recurring rhetorical patterns rather than targeting a single dominant narrative.

Our analysis of the predicted misinformation labels on CLIMATEDISC comments indicates that TikTok climate misinformation remains heavily skewed toward outright denial, but substantial volume also targets the credibility of messengers and promotes anti-solution narratives. These latter strategies focus on delegitimization and motive attacks rather than factual claims, making them more resistant to traditional fact-checking approaches and highlighting the need for alternative mitigation strategies.

Methodologically, our cross-domain matching pipeline from QUOTACLIMAT to CLIMATEDISC provides a practical mechanism for transferring fine-grained climate misinformation taxonomies to large, real-world platforms without requiring extensive human annotation. This scalability is critical for studying evolving misinformation narratives at platform scale.

Our findings indicate that persona information can aid misinformation classification, but its utility depends on model capacity. In lower-capacity settings, explicitly modeling persona yields consistent and statistically significant gains, suggesting that speaker-level regularities capture information not readily recoverable from individual comments alone. The faster early convergence of persona-aware MLP models further supports the view of persona as a structured inductive bias that helps exploit stable rhetorical patterns rather than sparse lexical cues. In contrast, gains are smaller and not statistically significant for RoBERTa-large, indicating that large pretrained models may already implicitly encode aspects of speaker style, stance, or narrative framing. Together, these results suggest that persona modeling is most beneficial when representational capacity or supervision is limited, offering diminishing returns as models become expressive. Rather than replacing strong text encoders, personas function as a complementary and lightweight inductive bias, particularly valuable in resource-constrained or efficiency-oriented settings.

Additionally, our persona-based LLM simulations demonstrate that such agents can serve as controlled probes of misinformation susceptibility: holding claims constant while varying persona produces systematic differences in acceptance that align with human persona-level vulnerability.

Our temporal dynamics analysis reveals that misinformation-persistence relationships vary substantially by persona type, with four of six personas showing statistically significant effects. These heterogeneous patterns suggest that different persona types are associated with distinct engagement trajectories when misinformation is present, potentially reflecting variation in moderation responses, amplification dynamics, and audience interaction patterns. This has direct implications for platform governance: effective content moderation requires differentiated strategies tailored to specific persona-misinformation combinations rather than uniform approaches across all content types.

Together, our findings suggest that persona-aware modeling offers a principled way to surface speaker-level structure in climate misinformation discourse, unifying prediction, susceptibility analysis, and engagement dynamics within a single abstraction. This perspective complements text-centric approaches and is especially valuable when model capacity or supervision is limited.

Limitations

Persona Induction Validity and Subjectivity Our personas are induced via embedding-based k-means clustering and then manually consolidated from 21 clusters into six higher-level types. While this improves interpretability, the resulting taxonomy is not uniquely determined: different embedding models, clustering algorithms, random seeds, or k selection criteria could yield alternative clustering results. Further, consolidation relies on researcher judgment and may introduce subjectivity in where boundaries are drawn. As such, personas should be interpreted as dataset-specific rhetorical groupings rather than definitive or exhaustive categories of users.

We infer persona 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, persona 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 applied to in-the-wild TikTok discourse. Although we validate transfer using semantic claim–comment matching, this validation relies on a cosine-similarity threshold selected via qualitative inspection and yields a relatively small set of matched pairs. Moreover, the QUOTACLIMAT taxonomy is claim-centric and may not perfectly align with how misinformation is expressed in conversational comments (e.g., sarcasm, implicature, partial agreement, or off-topic replies). As a result, predicted misinformation labels likely contain noise, which could affect downstream persona distributions, susceptibility estimates, and persistence analyses.

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 persona 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 While our analysis focuses on public comments without identifying individuals, the persona taxonomy could potentially be misused for profiling or targeting. We emphasize that personas describe rhetorical styles, not individual traits, and should not be used to make inferences about individual users' beliefs or susceptibility.

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 and clarity improvements throughout the manuscript, to refine the LaTeX formatting of several figures, and to support code refactoring and readability improvements in the analysis pipeline. The tool was used solely for assistance with writing, formatting, and code refactoring, and did not contribute to the research design, modeling decisions, data analysis, or empirical results.

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.

Corey Basch, Bhavya Yalamanchili, and Joseph Fera. 2021. climate change on tiktok: A content analysis of videos. *Journal of Community Health*, 47:163–167.

Adrian Benton and Mark Dredze. 2018a. 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.

Adrian Benton and Mark Dredze. 2018b. 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.

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.

Carlos Díaz Ruiz and Andreas Nilsson. 2023. Disinformation and echo chambers: Narrative and identity dynamics. *Journal of Marketing*.

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.

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.

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

S. Mehta and 1 others. 2022. Continually improving social context representations for fake news detection. In *ACL*.

Thanh Tam Nguyen, Ee-Peng Lim, and 1 others. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *CIKM*.

Joon Park, Joseph O'Brien, Carrie Cai, Meredith Morris, Percy Liang, and Michael Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. 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. pages 376–381.

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.

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.

Y. Zhang and 1 others. 2024. Modeling user stance with bipartite graph neural networks. In *ICWSM*.

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 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 misinformation classification task, indicating that the model is able to reliably distinguish among the fine-grained categories defined in QUOTACLIMAT.

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 4).

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 (Figure 5). 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 C.3), and we compare these annotations with our classifier's predictions.

Across these unmatched comments, agreement reaches 89%, with Cohen's $\kappa = 0.54$. The substantial raw agreement and moderate κ suggest that the classifier's predictions remain broadly aligned with independent judgments even when comments cannot be directly grounded to curated claims. Together, these results indicate that the model generalizes beyond claim-level paraphrases to more heterogeneous and implicitly framed misinformation in real-world discourse.

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 4: 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 5: Interpretation for different ranges of the Cohen's Kappa values.

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):

1. @sarah lewis
2. @psychicmediumjen
3. @Raymel Poche Jr.
4. @eliza koch
5. @MaKenzie Neas
6. @H.k.lynne
7. @emily xx
8. @Rachel Taylor
9. @aisela
10. @xottic_daa

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 LLM Prompts

C.1 Persona Card Construction

To construct persona cards, we prompt ChatGPT to synthesize coherent, reusable persona 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 persona 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 persona description.

Instructions:
- Describe this persona'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 persona description suitable for conditioning a language model.

Cluster description and example comments:
{cluster_description}

C.1.1 Personal Cards

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

Local / Experiential Observer
This persona 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 persona 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 persona 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 persona 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 persona 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 persona 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.'

C.2 Persona-Conditioned Stance Detection Prompt

We evaluate whether persona-conditioned LLM agents systematically agree with different types of climate misinformation using the following prompt template. The persona card (Appendix C.1.1) is injected into the `{persona}` field and the misinformation claim into the `{claim}` field.

```
Persona:
{persona}

Task:
Would *this persona* 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 persona-specific agreement patterns across misinformation categories.

C.3 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}
```