Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI and Social Good

Enhancing Online Climate Discourse: A Two-Stage Framework for Climate Content Categorization and Moderation

Apoorva Upadhyaya , Wolfgang Nejdl , Marco Fisichella
L3S Research Center, Leibniz University Hannover
{upadhyaya, nejdl, mfisichella}@l3s.de

Abstract

Climate change is one of the most pressing global challenges that requires urgent adaptation and resilience efforts, highlighting the need for both scientific solutions and effective communication. In the digital age, online content plays a key role in shaping climate narratives. Therefore, previous research has mainly focused on public perception or categorized content by topics such as impacts, mitigation, policy, etc. Despite these efforts, identifying discussions that address climate change adaptation is crucial for monitoring resilience and assessing public sentiment, while recognizing denial narratives helps combat misinformation. Moreover, the public's exposure to online climate content can either lead to or hinder climate action, emphasizing the need for climate content moderation. To address these issues, we propose a novel multi-stage framework where stage 1 categorizes climate-related content into adaptation, resilience, and denial while stage 2 moderates content by enhancing or intervening based on its alignment with climate goals. We present a novel dataset by manually annotating publicly available tweets and news articles into different climate categories with the help of a taxonomy developed by domain experts. Extensive experiments with benchmark climate and other domain datasets validate the efficacy of our prediction stage, while human and external evaluations confirm the relevance of our moderation stage.

1 Introduction

Climate change is one of the most critical challenges that our planet is facing today. As its impacts intensify, the need for climate adaptation and resilience has become more important than ever [Solecki *et al.*, 2024]. Climate adaptation is the process of adjusting to climate change to minimize risks, while climate resilience is the ability to anticipate, withstand, and recover from climate-related impacts [Hallegatte *et al.*, 2020]. These efforts are central to the United Nations Sustainable Development Goal (SDG) 13—Climate Action[1] (Targets 13.1 & 13.3), which emphasizes not only scientific and policy-driven solutions but also the critical role of effective communication and public engagement in driving meaningful change.

As in today's digital era, online content plays a crucial role in shaping public perception and influencing climate action [Pearce *et al.*, 2019]. Governments and policymakers rely on data-driven insights to formulate adaptation strategies, while climate scientists use digital discourse to spread awareness about scientific evidence. The general public, in turn, is exposed to diverse narratives on online platforms that can either promote climate action or spread misinformation [Treen *et al.*, 2020; Vivion *et al.*, 2024]. Due to this significant influence of online content, categorizing climate data into adaptation, resilience, and denial is critical not only for informing climate policies but also for combating misinformation.

While a plethora of research has analyzed public opinion on climate change [Upadhyaya *et al.*, 2023a; Upadhyaya *et al.*, 2023c], some of the prior studies have focused on classifying climate-related tweets into various categories such as root cause, impact, mitigation, politics, human intervention, and others [Vaid *et al.*, 2022; Duong *et al.*, 2022; Effrosynidis *et al.*, 2022]. Recently, [Islam *et al.*, 2023] released a theme-based dataset by exploring how organizations use climate campaigns to shape public perception. Despite these efforts, identifying discussions related to climate adaptation is crucial for tracking progress on resilience initiatives as well as assessing the public perception of climate policies [Woodruff *et al.*, 2022]. Furthermore, recognizing denial narratives could help combat misinformation. Hence, this diverse categorization will ensure that credible climate information reaches key stakeholders, ultimately supporting global climate adaptation efforts. This motivated us to focus on the classification of online content into climate adaptation, resilience, and denial as our primary objective.

To achieve this, different domain experts collaboratively developed a taxonomy and manually annotated publicly available climate-related tweets and news articles into adaptation, resilience, and denial categories (Section 3.1). We then propose a climate category prediction stage *(stage 1)* that utilizes both textual content and explicit cues extracted via LLMs (climate psychology values, hidden intent, and target stakeholders). These explicit and implicit features are processed through various model components to capture their in-

[1] https://sdgs.un.org/goals/goal13

teractions to finally predict climate categories (Section 2.2).

In addition, during the conduct of our classification experiments, we observe that some of the input posts contain misleading climate information such as *"Forget the f**k climateAction!! Eat more meat #ClimateHoax"*. Previous research has also shown that social media platforms, online news articles, and discussions serve as powerful tools to drive climate action or conversely hinder it [Pearce *et al.*, 2019]. Such information necessitates that existing content moderation systems not only react but proactively mitigate the harm of such content. Therefore, we consider climate content moderation as our secondary goal to ensure informed discussions.

To aim this, we design a climate content moderation stage *(stage 2)* that iteratively refines responses using a base-LLM (content generator) and climate-specific judge-LLM (content evaluator). Unlike prior works using LLMs for generating interventions [Jha *et al.*, 2024], our approach refines response generation through implicit token penalties and logits processing during training, without relying on ground truth while validating LLM responses in testing phase using external tools for robustness (Section 2.3).

Hence, in this study, our main contributions are as follows: ***(i.)***To the best of our knowledge, this is the first study to classify online content (tweets, news articles) into climate adaptation, resilience, and denial categories, followed by generating moderated, context-aware responses that either enhance or intervene based on climate mitigation alignment. ***(ii.)*** We present a novel dataset of publicly available climate-related tweets and articles, annotated using taxonomy collaboratively developed by domain experts. ***(iii.)***Our two-stage framework first classifies online posts using implicit and explicit contextual cues into different climate categories, then performs climate content moderation by generating refined responses through an iterative process involving a base-LLM and judge-LLM in the absence of ground-truth responses. ***(iv.)***Extensive experiments on benchmark climate and other domain datasets validate the significance of our prediction stage (stage 1), while human and external evaluations confirm the effectiveness of the LLM-based content moderation stage (stage 2). Code, Dataset, and Appendix are available here[2].

Task Alignment with UN SDGs This study aligns with UN SDG 13: Climate Action, specifically Targets 13.1 and 13.3, by categorizing online content into climate adaptation, resilience, and denial and introducing a climate content moderation system. This approach helps amplify credible climate discussions, promote resilience, foster awareness, and drive informed action among policymakers, scientists, and the public. Experimental analysis further demonstrates the effectiveness of targeted interventions and predictions across adaptation, resilience, and denial domains, strengthening climate communication and community preparedness and ultimately combating misleading information. Moreover, this interdisciplinary research, conducted in collaboration with computer science, biology, and science education researchers, school educators, and climate activists, aligns with SDG 17 (Partnerships for the Goals) by fostering cross-sector cooperation.

[2]https://osf.io/u4jnq/?view_only=8e706f57e9a7443b9fc6a9cc9222e26b

2 Methodology

Figure 1 shows the overall architecture of our approach, which consists of two main stages: *Climate Category Prediction* and *Climate Content Moderation*. We refer to our proposed method as **ClimaGuard** *(Climate Awareness and Guidance System)*. Next, we first describe the input features followed by the workflows of both stages.

2.1 Input Features

The textual content of the input post is considered as one of the input features (t). Since human behavior plays a crucial role in both causing and responding to climate change [Kikstra *et al.*, 2022; Steg, 2023], we query the LLM using the input text (t) to extract the user's climate psychology values (p) that influence their preferences and perceptions of climate actions (prompt in Figure 1 [Appendix A]), user's hidden intention that defines the purpose of the post (u), e.g. supportive, clarifying, provoking, satire, informative (Figure 2 [Appendix A]), and identify the key stakeholders/target groups of the post and assess the potential impact or perception of the post on each stakeholder group (s) using Figure 3 (Appendix A) to interpret how the post might influence or impact target audience, e.g, positive engagement, backlash, climate policy considerations. We consider p, u, and s as LLM-extracted features (see Figure 1).

2.2 Stage 1: Climate Category Prediction

This stage is responsible for classifying the given post into multiple climate categories. Implicit text (t) and LLM-extracted cues (p, u, s) are encoded and processed through attention mechanisms and a dynamic gated module to generate a context-aware representation. A feed-forward network then predicts multi-label categories for climate adaptation, resilience, and denial.

Embedding We initially pass all the input features (t, p, u, s) through separate embedding models. Following previous work [Nan *et al.*, 2024], we also use BAAI/bge-base-en-v1.5 [Xiao *et al.*, 2023] to generate high-quality meaningful representations of the input features, with dimension (d_e) followed by the dense layer of dimension d_f, leading to E_p, E_u, E_s, and $E_t \in R^{m \times d_f}$, where m is the maximum sequence length.

Self-Attention The embedded input text is then fed to the MultiHeadAttention (MHA) [Vaswani *et al.*, 2017] based on the concept of query (Q), key (K), and value (V) to uncover latent patterns in the input. We employ torch.nn.MultiheadAttention layer followed by dense layer of d_c, where embedded text (E_t) is fed as query, key, and value ($Q = E_t, K = E_t, V = E_t$), resulting in $C_i \in R^{m \times d_c}$ [$C_i = MultiHeadAttention([Q, K, V])$], representing implicit contextual information (refer Figure 1).

Co-Attention In parallel, we apply the co-attention mechanism to the LLM-extracted features to bridge the psychological values (E_p), intent (E_u), and real-world impacts (E_s) to achieve a more cohesive representation for predicting nuanced climate categories in text (Figure 1). Here, we first apply co-attention between the user's psychology (E_p) and intention (E_u), which captures the interplay between what the user's values and what they aim to achieve; resulting in *psychologically informed user intent*. To compute co-attention

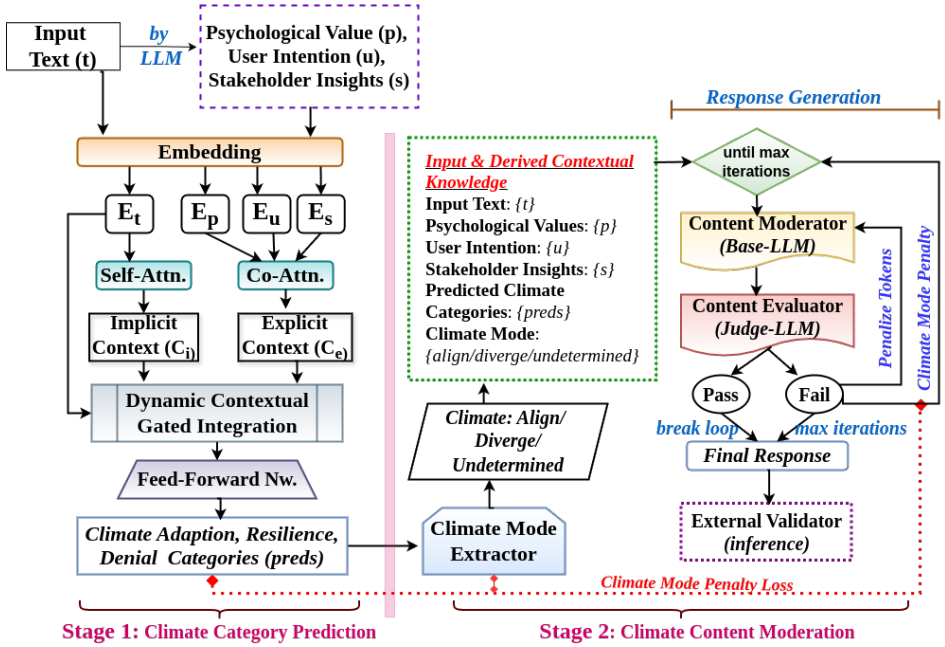

Figure 1: Architectural overview of our proposed ClimaGuard.

between E_p and E_u, we initially compute the affinity matrix M between E_p and E_u following [Xiong *et al.*, 2016] [$M = E_u E_p$]. The matrix M is normalized row and column-wise to obtain the attention weights $A_p = softmax(M)$ and $A_u = softmax(M^T)$. We then calculate the attention context of intention w.r.t psychology ($C_p = E_u A_p$) and vice-versa ($C_u = E_p A_u$). Similar to [Cui *et al.*, 2016], we capture $C_p A_u$ of the previous attention contexts in parallel, leading to the final co-dependent representation of psychology and intention as co-attention context, where $\in R^{m \times 2(d_f)}$ (Eq. 1).

$$Co - Atten(E_p, E_u) = [E_p A_u; C_p A_u] \tag{1}$$

Co-attention is again applied to integrate psychologically informed user intent with the stakeholder impact (E_s) [similar process is applied using Eq. 1]. This reflects how well the user's motivations match the external effects and identify potential conflicts or synergies between intent and audience reception.

$$C_e = Co - Atten(Co - Atten(E_p, E_u)), E_s) \tag{2}$$

The final co-attentive vector (Eq. 2), passed through a dense layer of d_c, encapsulates subtle interactions of external knowledge and ensures that both user-driven (values and intentions) and audience-driven (impact on stakeholders) factors are modeled holistically, which helps capture the *multi-faceted explicit context* $C_e \in R^{m \times d_c}$.

Dynamic Contextual Gated Integration is responsible for dynamic fusion strategy that combines the input with both implicit and explicit contexts. The implicit (C_i) and explicit (C_e) context, along with the encoded text (E_t), flow through this gating component (Figure 1). The main intuition is rooted in dynamically leveraging complementary perspectives to make a more nuanced and context-aware prediction. It allows the model to adjust its focus depending on the nature of the input. For example, if implicit context is clear, it may dominate; if explicit signals are strong (e.g., skepticism), they may carry more weight. To achieve this, we introduce two gates, C_i^{filter} and C_e^{filter}, which assess the interplay between the input and inferred context, producing a filtered implicit or explicit representation respectively.

$$u_{\text{i}} = \sigma(W_1 \cdot E_{\text{t}} + W_2 \cdot C_{\text{i}} + \text{b}_1) \tag{3}$$

$$u_{\text{e}} = \sigma(W_3 \cdot E_{\text{t}} + W_4 \cdot C_{\text{e}} + \text{b}_2) \tag{4}$$

where W_1, W_2, W_3, W_4, b_1, and b_2 are trainable weights and bias parameters and u_i and u_e dynamically decide how much weight to assign to the implicit or explicit context with respect to the encoded input. These vectors are then filtered using equations 5 and 6, ensuring that context is balanced against the input text while reducing the impact of irrelevant representations.

$$C_i^{\text{filter}} = u_{\text{i}} \cdot C_{\text{i}} + (1 - u_{\text{i}}) \cdot E_{\text{t}} \tag{5}$$

$$C_e^{\text{filter}} = u_{\text{e}} \cdot C_{\text{e}} + (1 - u_{\text{e}}) \cdot E_{\text{t}} \tag{6}$$

These filtered representations are then combined via a Hadamard product to capture their interactions, and this result is fused with the embedded input (E_t) leading to $F_{\text{final}} \in R^{m \times d_c}$ (Eq. 7), ensuring that core semantic information is preserved while leveraging the complementary strengths of implicit and explicit contexts.

$$F_{\text{final}} = E_{\text{t}} + C_i^{\text{filter}} \odot C_e^{\text{filter}} \tag{7}$$

Feed Forward Network The final fused representation (F_{final}) is flattened and then passed through the feed-forward network consisting of two dense layers with ReLU activation to classify the context-aware input representation into present climate adaption, resilience, or denial categories.

2.3 Stage 2: Climate Content Moderation

Once the multiple climate categories are extracted, this phase aims to moderate the content of the given input post by generating a response to either improve or intervene in the post, thus promoting ethical and engaging climate discourse. Algorithm 1, defining the workflow of iterative response generation and evaluation, is present in Appendix D.

Climate Mode Extractor serves as a critical intermediary between the category prediction and the content moderation stages, guiding how the system interprets and responds to a post based on its alignment with climate change narratives (Figure 1). In this module, we interpret predicted categories ($preds$) from Section 2.2 to determine the climate $mode$ of post (Step 1, Algorithm 1). We consider mode as $align$ towards climate change if the post consists of climate adaptation or resilience as predicted categories. If it contains counterproductive narratives (e.g., denial or misinformation), the mode is determined as $diverge$. When categories are ambiguous (e.g. a mix of denial and adaptation), the mode is $undetermined$.

Response Generation ensures moderated outputs by iteratively refining responses using a base-LLM (content moderator) and a judge-LLM (content evaluator) in the absence of ground truth responses. Here, base-LLM first generates a response [Step 4, Algorithm 1], which is then evaluated by a judge-LLM [Step 9, Algorithm 1]. Failed responses incur a penalty based on the climate mode, which adjusts token weights [Steps 13-16, Algorithm 1]. These adjusted weights influence token probabilities via logits processor during subsequent response generation by base-LLM [Steps 7 and 17, Algorithm 1]. This cycle continues iteratively, refining the response by base-LLM until it passes the evaluation or the maximum iterations are reached. Next, we detail the components of response generation.

Content Moderator (base-LLM) takes input text, LLM-extracted features, predicted climate categories, and the extracted climate mode to generate a moderated response using Prompt 5 (Appendix C) based on context while ensuring it aligns with user's intent or diverges as per the mode.

Content Evaluator (Judge-LLM) evaluates the base-LLM's response based on toxicity (0-1), persuasiveness (-3 to 3), factual accuracy (true/false/misleading/undetermined), and a binary pass/fail judgment (Step 9, Algorithm 1). As ground truth is unavailable, a climate-specific LLM serves as the judge, ensuring domain-relevant evaluation and structured feedback for refinement (supported by Section 4.2). The range for toxic and persuasive scores and different factual categories relies on external validation (Section 2.3).

Penalty for Failing Responses If judge-LLM rates a base-LLM response as "fail", a $penalty$ is applied based on the climate mode to enforce alignment (Step 13, Algorithm 1). A stronger penalty (0.7) in 'diverge' mode is applied to reflect the failure to counter anti-climate input or context, while a moderate penalty (0.5) in 'align' mode applies for inadequate alignment. A softer penalty (0.2) in 'undetermined' mode accounts for uncertainty.

Token Penalization for Failing Responses After applying the penalty, base-LLM is further penalized for generating a failed response by adjusting token weights based on the climate mode. Initially, all tokens have equal weight (1) (Step 5, Algorithm 1). If the response is judged as "fail", cosine similarity between generated tokens and the combined input context ($p, u, s, preds$) is calculated. In *align mode*, tokens dissimilar to the context ($cos\_sim < 0.5$) are penalized by reducing their weight to 0.5 ($1 - penalty$), encouraging alignment. In *diverge mode*, overly similar generated tokens with context ($cos\_sim >= 0.5$) are punished by reducing their weight to 0.3 ($1 - penalty$), preventing the use of tokens that deviated from climate change. In *undetermined* mode, a mild penalty of 0.8 ($1 - penalty$) is applied to all tokens. These adjustments guide the base LLM toward improved responses by penalizing specific tokens from the failed response in subsequent iterations (Steps 14-16, Algorithm 1). Please note that we reduce the weights by $1 - penalty$ to punish the tokens of a failed response, as lower weights decrease the probability of selecting those tokens in future iterations.

Logits Processor adjusts token selection during generation by applying the log of penalized token weights to the logits (Steps 16 and 17, Algorithm 1). This custom-made logits processor is then passed in the 'logits_processor' parameter of the base-LLM while generating response. Since the log of a reduced token weight is negative, it decreases the logit values for such penalized tokens, lowering their selection probability. This discourages the model from generating previously chosen tokens, guiding it toward a more appropriate response.

Iterative Refinement Entire process repeats iteratively, generating a new response, evaluating it, and penalizing tokens until an adequate response is produced or the maximum iterations are reached (Steps 2-19, Algorithm 1).

External Validation After the LLMs generate the final response, it is evaluated by external validators. Following prior research [Upadhyaya *et al.*, 2023c; Choi and Ferrara, 2024; Liu *et al.*, 2024], we use the Perspective API [Hosseini *et al.*, 2017] to assess the toxicity score (0-1), a Huggingface-based persuasiveness model [Pauli *et al.*, 2024] that provides a score between -3 to 3 to measure how persuasive the generated content is compared to the input, and the Google Fact-Check Explorer [Check, 2024] to determine the factual accuracy of the response in terms of true, false, misleading, or undetermined by analyzing the top 5 claims. These external tools help evaluate whether the LLM-produced content is toxic, persuasive, and factually accurate. This evaluation is performed during the testing phase to ensure transparency, however, we avoid external validation during training to reduce latency. During training, LLMs are guided by implicit token penalties, logits processing, and mode penalties to refine the response generation process without ground-truth (Algorithm 1).

2.4 Loss Functions

Category Prediction Loss (L_c) As category prediction is a multi-label classification, binary cross-entropy loss is used.

Contrastive Adversarial Loss (L_a) encourages the model to differentiate between correct and incorrect climate category predictions by adjusting the Euclidean distance between implicit (C_i) and explicit (C_e) context embeddings [$\mathrm{d} = \|\mathbf{C}_\mathrm{i} - \mathbf{C}_\mathrm{e}\|_2$]. For correct predictions, it increases the distance between these embeddings, promoting adversarial learning, while for incorrect predictions, it reduces the dis-

tance, helping the model better align with the correct labels (Eq. 8). This helps in training the model to distinguish between correct and incorrect predictions by manipulating the embeddings in an adversarial and contrastive manner.

$$\mathrm{L_a} = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{correct}_i \cdot \mathrm{ReLU}(\mathrm{margin} - \mathrm{d}_i) + (\mathrm{incorrect}_i) \cdot (-\mathrm{d}_i)) \tag{8}$$

Mode Penalty Loss (L_p) improves the model's category prediction accuracy while ensuring that generated responses align with the desired context, such as climate alignment. By penalizing failed responses based on the climate mode (Section 2.3), the model is guided to either stay aligned with or intentionally diverge from the context. This encourages the model to better categorize inputs into climate-specific categories and generate more contextually appropriate responses. The penalty, as described in Section 2.3, is normalized over the number of all failed responses of each input [$P_{\text{mean}} = \frac{1}{N_{\text{fail}}}\sum_{k=1}^{N_{\text{fail}}} penalty_k$], which is then normalized over all inputs and linked with the category prediction loss L_c (Eq. 9).

$$\mathcal{L}_{\mathrm{p}} = L_c . \frac{1}{N}\sum_{i=1}^{N}(P_{\mathrm{mean}\,i}) \tag{9}$$

Total loss: of our proposed approach is, $L = L_c + L_p + pL_a$.

3 Experimental Setup

3.1 Dataset

(i) ClimateTweets (CT): is a benchmark dataset of $8,881$ climate tweets with believe, deny, and ambiguous views regarding climate change [Upadhyaya *et al.*, 2023b]. We randomly select 2000 tweets for manual annotation for the climate category task. ***(ii) ClimateArticles (CA):*** To assess the efficacy of ClimaGuard across varying content lengths beyond tweets, we collect the 50 most recent climate news articles from 22 publicly available newspapers [Efficiency, 2024], yielding 1100 articles dated between 2023-05-29 and 2024-06-29. ***(iii) Climate & COVID Benchmark Datasets:*** To evaluate the generalizability of our ClimaGuard for *category prediction task*, we tested our model on different climate and COVID-related category classification datasets [Duong *et al.*, 2022; Upadhyaya *et al.*, 2024].

Climate Category Annotation To systematically define distinct climate-related categories, an interdisciplinary team comprising researchers from diverse fields—including computer science, biology, and science education—along with educators from secondary schools and activists from the Fridays for Future movement, collaboratively developed a taxonomy of adaptation, resilience, and denial categories. This categorization process of climate data is extensively reviewed using the information published by the United Nations[3], as well as prior research identifying climate change mitigation efforts and denial narratives on social media platforms [Duong *et al.*, 2022; Upadhyaya *et al.*, 2024; Gounaridis and Newell, 2024]. Few examples of "adaptation": plan for adaptation, implement adaptation measures, "resilience": emergency preparedness, early warning systems; "denial": denial of human impact, doubting scientific consensus. ***Manual Annotation:*** A team of five trained annotators with interdisciplinary backgrounds was assigned to annotate CT and CA datasets with the appropriate category labels according to the above schemas (multi-label classification). We gave clear instructions to the annotators to avoid any inherent bias towards the climate crisis and to annotate solely based on the meaning conveyed in the textual content. We obtained Cohen Kappa scores [Fleiss and Cohen, 1973] (inter-annotator agreement) of 0.79 (CT) and 0.77 (CA). These denote that the quality of annotations and the presented datasets are significantly productive. Dataset statistics are shown in Table 1 for both CT and CA datasets, where categories 1-4 represent adaptation, 5-10 resilience, 11-14 denial, and the last as none.

Category	CT	CA
Assess Impacts	25.80	65.55
Plan for Adaptation	10.70	25.27
Implement Adaptation	5.30	15.91
Evaluate Adaptation	0.10	3.45
Early Warning Systems	0.50	5.36
Emergency Preparedness	0.65	5.73
Slow Onset Events	2.05	29.45
Permanent Loss and Damage	2.40	17.45
Non-Economic Losses	2.00	18.36
Resilience of Communities	27.70	51.45
Denial of Human Impact	31.25	3.64
Resistance to Climate Action	32.50	10.73
Doubting Scientific Consensus	29.95	3.36
Spreading Information Pollution	30.15	2.64
None of these	4.55	6.18

Table 1: % distribution of different categories in CT & CA.

LLM-Generated Responses Evaluation (Content Moderation) After several rounds of discussion, an interdisciplinary team of five annotators (similar team who annotated climate category task) established the following evaluation criteria for the generated moderated response, based on recent research [Wang *et al.*, 2023] and other proposed parameters, which are rated on a 5-point Likert scale [1: not; 5: very]: *Informativeness* (accuracy and relevance of climate change information), *Relevance to Prompt* (focus on the specific climate-related query), *Responsible Communication* (avoidance of climate myths), and *Evaluating Impact* (clarity, feasibility, and novelty of climate solutions). The annotation team then scored the LLM-generated responses based on these parameters, as no ground truth was available.

3.2 Implementation Details

Evaluation metrics, hyperparameters, baselines, and environment details are covered in **Appendix E**.

4 Results

4.1 Climate Category Prediction

Comparison with Baselines

Table 2 shows that our ClimaGuard outperforms other baselines with an average weighted F1 score of 86.15 and 86.68,

[3]https://unfccc.int/topics/adaptation-and-resilience/the-big-picture/introduction#adaptation,https://climatepromise.undp.org/what-we-do/areas-of-work/adaptation-and-resilience,https://www.unep.org/topics/climate-action/adaptation

Model	ClimateTweets		ClimateArticles	
	Macro F1	Weighted F1	Macro F1	Weighted F1
	Avg./Std.dev	Avg./Std.dev	Avg./Std.dev	Avg./Std.dev
Large Language Models (LLMs)				
Mistral[zero-shot](t)	54.08/2.01	60.83/2.11	57.05/1.69	63.39/1.55
Mistral[few-shot](t)	55.14/3.10	61.59/2.59	59.31/1.72	64.25/2.01
Mistral[fine-tune](t)	60.02/1.04	66.4/0.64	61.68/0.59	68.91/0.61
Mistral[fine-tune](t,p,u,s)	61.36/0.77	69.35/0.35	65.27/0.49	71.69/0.61
Llama 3.2[zero-shot](t)	63.01/0.79	69.6/0.76	65.06/0.53	69.55/0.48
Llama 3.2[few-shot](t)	65.44/1.05	72.07/1.16	67.19/0.66	71.09/0.68
Llama 3.2[fine-tune](t)	69.09/0.51	76.32/0.57	72.55/0.49	78.76/0.81
Llama 3.2[fine-tune](t,p,u,s)	**72.62/1.13**	**79.05/0.91**	75.63/0.45	81.14/0.42
ClimateGPT[zero-shot](t)	59.45/0.41	65.63/0.29	62.18/0.25	68.08/0.31
ClimateGPT[few-shot](t)	64.11/0.58	69.17/0.26	67.57/0.39	71.17/0.41
ClimateGPT[fine-tune](t)	68.37/0.18	75.08/0.35	72.61/1.01	79.49/1.06
ClimateGPT[fine-tune](t,p,u,s)	71.28/0.62	78.14/0.49	**76.08/0.31**	**81.69/0.45**
Small Language Models				
BERT	55.26/2.04	60.15/2.10	56.15/1.31	61.57/1.69
RoBERTa	60.33/1.66	63.29/1.17	60.08/1.52	64.31/1.20
BERTweet	64.59/1.05	69.27/1.23	63.79/0.57	70.66/0.42
ClimateBERT	**64.67/0.38**	**70.15/1.02**	**67.15/0.39**	**72.43/0.61**
Our Proposed Variants				
Text (t)+im.context(self-atten.)	66.95/1.13	74.67/0.48	74.94/1.11	80.49/1.16
Text+ex. context(p,u,s)(concat)	70.58/0.69	77.06/0.83	76.12/1.39	81.39/0.69
Text+ex.(p,u,s)(co-atten.)	73.05/1.25	79.30/0.67	78.56/1.01	83.64/0.75
Text+im.+ex.(concat)	75.15/0.37	81.13/1.05	79.00/1.23	84.70/1.18
Text+im.+ex.(dynamic)	78.6/1.01	84.96/0.42	80.87/0.51	86.19/0.65
Text+im.+ex.(dynamic) +adv.+penalty (ClimaGuard)	**79.86/0.47**	**86.15/0.51**	**81.04/0.23**	**86.68/0.38**

Table 2: Results (Macro/Weighted F1) of baselines and ClimaGuard.[highlight :overall best; **bold**: best within category]

resulting in an average improvement of 15.89% and 12.89% compared to the best LLM and small language model (SLM) on CT and CA datasets for category prediction task. This demonstrates the strength of our approach which efficiently captures implicit and explicit contexts, enabling a richer, hierarchical understanding of climate-related content to identify the nuanced categories. It is also evident from Table 2 that LLMs perform better when fine-tuned with the training dataset. The addition of climate psychology values (p), intentions (u), and impact on stakeholders (s) along with the text further increases the performance of all LLMs with an average improvement of 3.91% and 3.27% in weighted F1 for CT and CA respectively, signifying the usefulness of contextual information for efficiently identifying multiple climate categories. Although Llama 3.2 achieves the highest weighted F1 score (79.05) among the baselines for CT, ClimateGPT performs comparably with 78.14 F1 for CT and better with 81.69 F1 for CA dataset, showcasing its effectiveness due to fine-tuning on curated climate documents and instruction-completion pairs by climate scientists, making it relevant for climate-specific tasks. However, our ClimaGuard performs better than these baselines, demonstrating its power of dynamically filtering and selecting the relevant fusion of implicit and multifaceted explicit context captured by effective learning of context embeddings in contrastive and adversarial manner. Compared with baselines, results of ClimaGuard are statistically significant (under t-tests ($p<0.05$).

Comparison with Different ClimaGuard Variants

Table 2 presents the different variants of our ClimaGuard. Adding explicit context to the text improves the weighted F1 by 3.20% for CT and 1.12% for CA datasets. This highlights the significance of external knowledge, particularly for tweets, where the shorter text length may lack the hidden cues. The concatenation of implicit and explicit context further improves ClimaGuard's performance. It is observed that integrating the context using the dynamic gating mechanism instead of concatenation enhances weighted F1 by 4.72% and 1.76% in CT and CA respectively, indicating the efficacy of our proposed gated mechanism to leverage the complementary strengths of both context and input. The addition of penalty and contrastive embedding loss functions further guides our ClimaGuard to efficiently identify different climate categories in the text.

Model	Impact	Miti.	P&P	R.C.	Oth.	Mic.F1
LDA	0.13	0.32	0.26	0.02	0.17	0.22
GloVe-LSTM	0.50	0.38	0.27	0.05	0.30	0.38
GloVe-GRU	0.37	0.45	0.28	0.07	0.34	0.36
BERT-FC	0.79	0.64	0.68	0.45	0.61	0.67
Ours	**0.85**	**0.71**	**0.73**	**0.61**	**0.68**	**0.76**

Table 3: Results of ClimaGuard on benchmark category data

Variant	Climate	COVID
GPT-3.5(FS+C,E)	70.25	68.80
Mistral(FS+C,E)	74.23/0.11	65.32/0.51
BERT	67.98/0.41	68.49/0.47
RoBERTa	70.15/1.04	70.22/1.31
CLIMATEBERT	72.43/1.41	69.19/0.53
BERTweet	75.84/0.56	74.56/1.15
COVID-Twitter-BERT	69.45/1.62	78.24/1.39
VirtuAI	83.25/0.72	87.03/0.59
Ours	**86.31/1.48**	**89.62/1.15**

Table 4: Weighted F1 (Mean/Std. dev) of ClimaGuard on benchmark climate and COVID relevance detection task.

Visualization of Category-Wise Results

Figures 7 (a) and (b) [Appendix F] present the category-wise precision, recall, and F1 score of our ClimaGuard for the best round of results on CT and CA. ClimaGuard performs well on both datasets across climate adaptation, resilience, and denial categories, with the CA achieving more balance in recall and F1 scores than CT due to better representation of categories within the CA dataset (Table 1). For ***adaptation categories (1–4)***, ClimaGuard on CA has a higher average F1 of 83.22 than on CT, which allows for better identification of posts about impacts, planning, and evaluation of adaptation measures; the categories that support resource allocation and collective action. For ***resilience categories (5–10)***, both models show strong recall, especially for slow-onset events, loss and damage, and community resilience categories. ClimaGuard on CA also captures posts on emergency preparedness more effectively, offering consistent performance across these categories, which are helpful in disaster preparedness and resource planning (Figure 7 (b)). For ***denial categories (11–14)***, the models on both datasets show high precision, with CT achieving higher F1 scores (Figure 7 (a)). However, ClimaGuard on CA compensates with a recall of 0.9091 by identifying false, misleading, or harmful information related to climate change (category 14) effectively. This capability is crucial for addressing misinformation patterns and ensuring accurate public discourse on climate change. Overall, the models' ability to classify these categories accurately supports targeted predictions across adaptation, resilience, and denial domains, enhancing communication strategies, promoting community preparedness, and combating climate misinformation effectively.

LLM	Input	Informat.		Rel. to Prompt		Respons. Comm.		Evaluate. Impact	
		CT	CA	CT	CA	CT	CA	CT	CA
Llama 3.2	**text only**	3.3	3.02	2.83	2.65	3.51	3.52	2.66	3.01
	text+context +categories	**4.15**	3.91	3.91	**3.66**	**4.27**	**4.16**	**4.35**	**4.29**
ClimateGPT	**text only**	2.87	3.17	3.16	2.74	3.04	3.32	3.16	3.10
	text+context +categories	3.84	**4.04**	**4.03**	3.5	4.05	4.03	3.81	**4.06**
Mistral	**text only**	3.1	2.87	2.72	2.12	2.98	3.05	2.66	2.58
	text+context +categories	3.69	3.25	3.05	3.56	3.55	3.87	3.75	3.31

Table 5: Human evaluation of LLM-Responses for CT & CA

Evaluation Across Benchmark Climate and COVID-Related Categories

Table 3 shows that our ClimaGuard outperforms baseline methods with 0.76 Micro-F1 in classifying climate-related tweets into five high-level categories: Root Cause, Impact, Mitigation, Politics or Policy, and Others [Duong *et al.*, 2022]. Moreover, the superior performance of ClimaGuard (Table 4) on the relevance detection task across climate and COVID-related categories [Upadhyaya *et al.*, 2024] highlights its robustness and adaptability by effectively capturing relevant categories in diverse and critical scenarios, for instance not only in climate but also in COVID. The results are statistically significant under t-tests ($p<0.05$).

4.2 Climate Content Moderation

Human Evaluation

Table 5 summarize the human evaluation of responses generated by different base-LLMs, rated on a scale of 1 to 5 across predefined parameters (refer Section 3.1) for both CT and CA datasets. The results indicate that responses generated by Llama are the most effective across most parameters, while ClimateGPT performs comparably, excelling in relevance to the prompt for CT and informativeness for CA. For example, Llama and ClimateGPT responses usually consist of sentences like: *I understand your skepticism, but let's take a closer look..Great initiative!, wake-up call for us, let's keep the conversation respectful!*. These findings suggest that both models are well-suited as base-LLMs for generating climate-appropriate responses. Furthermore, incorporating explicit knowledge, predicted categories and modes significantly enhances the informativeness, relevance, and ethical soundness of the generated climate responses compared to using text alone. This highlights the importance of leveraging external knowledge to provide LLMs with enriched context for response generation. Thus, we consider Llama 3.2 as best choice for base LLM to generate responses, and ClimateGPT which is domain-specific, became the better option for judge LLM in stage 2 (refer Section 4.2 [External Validation]).

External Validation

Figures 8 (a) and (b) and Figures 9 (a) and (b) [Appendix G] represent the cumulative distribution functions of persuasive and toxicity scores for both CT and CA datasets respectively. We conduct Kolmogorov-Smirnov [Hodges Jr, 1958] statistical method to compare the score distributions of both ClimateGPT and Llama with external validation tools by evaluating the maximum difference between their cumulative distribution functions (D-statistic). For persuasive scores (Figure 8 [Appendix G]), ClimateGPT (D = 0.159[CT]; 0.123[CA]) is closer to the external validation distribution compared to Llama (D = 0.208[CT]; 0.422[CA]), though both differences are statistically significant (p-values < 0.0001). For toxic scores, both LLMs deviate more substantially from the external validation data, with ClimateGPT (D = 0.367[CT]; 0.184[CA]) being closer than Llama (D = 0.579[CT]; 0.569[CA]) (Figure 9 [Appendix G]). We also perform the Jensen-Shannon Divergence statistical test (scipy.spatial.distance.jensenshannon) to measure the similarity between the factual categorical data of the two LLMs and Google Fact-Check (Figures 10 (a) and (b) [Appendix G]). For factual responses, ClimateGPT shows a lower JSD (0.0753[CT]; 0.0549[CA]) compared to Llama (0.1085[CT]; 0.0855[CA]) when evaluated against the external tool, suggesting that GPT is more consistent with the external tool's categorization of factual responses than Llama. Overall, *ClimateGPT* demonstrates better alignment with external validation data for persuasiveness, toxicity, and factual categorical data validating that ClimateGPT is a better choice as judge-LLM for evaluating base-LLM's responses.

Qualitative Analysis

Table 1 (Appendix H) shows the Llama responses for a sample tweet and article from both datasets. As can be seen from table, LLM-response quality improves significantly by incorporating context and category predictions. While a text-only response tends to be neutral and passive, focusing mainly on moderating tone, adding context allows the model to better understand the user's intent and psychological values, framing the conversation towards a more respectful and constructive exchange. The full input— with predicted categories and a divergence mode— allows the model to tackle misleading information proactively and provide both corrections and guidance toward solutions. Ultimately, the addition of context and predicted categories not only enhances the relevance and impact of the response but also ensures that it is more focused on promoting climate action, which is crucial for steering the public discourse toward positive change.

5 Related Work

Due to space limitations, Appendix I includes literature overview, research gaps, and motivation behind our approach.

6 Conclusion

In our work, we address the UN SDG 13-Climate Action (Targets 13.1 & 13.3) by proposing a novel multi-stage framework where stage 1 focuses on identifying climate adaptation, resilience, and denial categories, while stage 2 moderates online climate posts to either improve or intervene with respect to climate objectives and goals. We present a novel dataset of online tweets and news articles that were categorized into different climate categories with the help of a taxonomy developed by experts. Extensive experiments demonstrate the generalizability of our stage 1 and relevance of our stage 2 in the absence of ground-truth. Hence, our approach strengthens credible and responsible climate discourse, fosters resilience, enhances awareness, and thus empowers key stakeholders to take informed action while combating misinformation.

Acknowledgements

This work is partially supported by the research project SoBigData RI PPP funded by the European Commission with grant agreement number 101079043.

References

[Check, 2024] Google Fact Check. About fact check tools, 2024. https://toolbox.google.com/factcheck/. Accessed: 2024-02-04.

[Choi and Ferrara, 2024] Eun Cheol Choi and Emilio Ferrara. Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1441–1449, 2024.

[Cui *et al.*, 2016] Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. Attention-over-attention neural networks for reading comprehension. *arXiv preprint arXiv:1607.04423*, 2016.

[Duong *et al.*, 2022] Cuc Duong, Qian Liu, Rui Mao, and Erik Cambria. Saving earth one tweet at a time through the lens of artificial intelligence. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2022.

[Efficiency, 2024] ADG Efficiency. Climate news database, 2024. https://github.com/ADGEfficiency/climate-news-db. Accessed: 2024-02-04.

[Effrosynidis *et al.*, 2022] Dimitrios Effrosynidis, Alexandros Karasakalidis, Georgios K. Sylaios, and Avi Arampatzis. The climate change twitter dataset. *Expert Syst. Appl.*, 204:117541, 2022.

[Fleiss and Cohen, 1973] Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973.

[Gounaridis and Newell, 2024] Dimitrios Gounaridis and Joshua P Newell. The social anatomy of climate change denial in the united states. *Scientific Reports*, 14(1):2097, 2024.

[Hallegatte *et al.*, 2020] Stephane Hallegatte, Jun Rentschler, and Julie Rozenberg. Adaptation principles: a guide for designing strategies for climate change adaptation and resilience. 2020.

[Hodges Jr, 1958] JL Hodges Jr. The significance probability of the smirnov two-sample test. *Arkiv för matematik*, 3(5):469–486, 1958.

[Hosseini *et al.*, 2017] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.

[Islam *et al.*, 2023] Tunazzina Islam, Ruqi Zhang, and Dan Goldwasser. Analysis of climate campaigns on social media using bayesian model averaging. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 15–25, 2023.

[Jha *et al.*, 2024] Prince Jha, Raghav Jain, Konika Mandal, Aman Chadha, Sriparna Saha, and Pushpak Bhattacharyya. Memeguard: An LLM and vlm-based framework for advancing content moderation via meme intervention. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8084–8104. Association for Computational Linguistics, 2024.

[Kikstra *et al.*, 2022] Jarmo S Kikstra, Zebedee RJ Nicholls, Christopher J Smith, Jared Lewis, Robin D Lamboll, Edward Byers, Marit Sandstad, Malte Meinshausen, Matthew J Gidden, Joeri Rogelj, et al. The ipcc sixth assessment report wgiii climate assessment of mitigation pathways: from emissions to global temperatures. *Geoscientific Model Development*, 15(24):9075–9109, 2022.

[Liu *et al.*, 2024] Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, and Chenhao Tan. Literature meets data: A synergistic approach to hypothesis generation. *arXiv preprint arXiv:2410.17309*, 2024.

[Nan *et al.*, 2024] Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1732–1742, 2024.

[Pauli *et al.*, 2024] Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. Measuring and benchmarking large language models' capabilities to generate persuasive language. *arXiv preprint arXiv:2406.17753*, 2024.

[Pearce *et al.*, 2019] Warren Pearce, Sabine Niederer, Suay Melisa Özkula, and Natalia Sánchez Querubín. The social media life of climate change: Platforms, publics, and future imaginaries. *Wiley interdisciplinary reviews: Climate change*, 10(2):e569, 2019.

[Solecki *et al.*, 2024] William Solecki, Debra Roberts, and Karen C Seto. Strategies to improve the impact of the ipcc special report on climate change and cities. *Nature Climate Change*, 14(7):685–691, 2024.

[Steg, 2023] Linda Steg. Psychology of climate change. *Annual Review of Psychology*, 74(1):391–421, 2023.

[Treen *et al.*, 2020] Kathie M d'I Treen, Hywel TP Williams, and Saffron J O'Neill. Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 11(5):e665, 2020.

[Upadhyaya *et al.*, 2023a] Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. Intensity-valued emotions help stance detection of climate change twitter data. In *IJCAI*, pages 6246–6254, 2023.

[Upadhyaya *et al.*, 2023b] Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. A multi-task model for emotion and offensive aided stance detection of climate change tweets. In *Proceedings of the ACM Web Conference 2023*, pages 3948–3958, 2023.

[Upadhyaya *et al.*, 2023c] Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. Toxicity, morality, and speech act guided stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4464–4478, 2023.

[Upadhyaya *et al.*, 2024] Apoorva Upadhyaya, Wolfgang Nejdl, and Marco Fisichella. Harnessing empathy and ethics for relevance detection and information categorization in climate and covid-19 tweets. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4091–4095, 2024.

[Vaid *et al.*, 2022] Roopal Vaid, Kartikey Pant, and Manish Shrivastava. Towards fine-grained classification of climate change related social media text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Vivion *et al.*, 2024] Maryline Vivion, Valérie Trottier, Ève Bouhêlier, Isabelle Goupil-Sormany, Thierno Diallo, et al. Misinformation about climate change and related environmental events on social media: Protocol for a scoping review. *JMIR Research Protocols*, 13(1):e59345, 2024.

[Wang *et al.*, 2023] Han Wang, Ming Shan Hee, Md. Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. Evaluating GPT-3 generated explanations for hateful content moderation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 6255–6263. ijcai.org, 2023.

[Woodruff *et al.*, 2022] Sierra C Woodruff, Sara Meerow, Missy Stults, and Chandler Wilkins. Adaptation to resilience planning: Alternative pathways to prepare for climate change. *Journal of Planning Education and Research*, 42(1):64–75, 2022.

[Xiao *et al.*, 2023] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.

[Xiong *et al.*, 2016] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*, 2016.