

# CliME: Evaluating Multimodal Climate Discourse on Social Media and the Climate Alignment Quotient (CAQ)

Abhilekh Borah<sup>\*1</sup>, Hasnat Md Abdullah<sup>\*2</sup>, Kangda Wei<sup>2</sup>, Ruihong Huang<sup>2</sup>

<sup>1</sup>Manipal University Jaipur, India

<sup>2</sup>Texas A&M University, USA

abhilekh.229301149@uj.manipal.edu, hasnat.md.abdullah@tamu.edu,  
kangda@tamu.edu, huangrh@cse.tamu.edu

## Abstract

The rise of Large Language Models (LLMs) has raised questions about their ability to understand climate-related contexts. Though climate change dominates social media, analyzing its multimodal expressions is understudied, and current tools have failed to determine whether LLMs amplify credible solutions or spread unsubstantiated claims. To address this, we introduce **CliME (Climate Change Multimodal Evaluation)**, a first-of-its-kind multimodal dataset, comprising 2579 Twitter and Reddit posts. The benchmark features a diverse collection of humorous memes and skeptical posts, capturing how these formats distill complex issues into viral narratives that shape public opinion and policy discussions. To systematically evaluate LLM performance, we present the **Climate Alignment Quotient (CAQ)**, a novel metric comprising five distinct dimensions: **Articulation, Evidence, Resonance, Transition, and Specificity**. Additionally, we propose three analytical lenses: **Actionability, Criticality, and Justice**, to guide the assessment of LLM-generated climate discourse using CAQ. Our findings, based on the CAQ metric, indicate that while most evaluated LLMs perform relatively well in Criticality and Justice, they consistently underperform on the Actionability axis. Among the models evaluated, Claude 3.7 Sonnet achieves the highest overall performance. We publicly release our *CliME* dataset and *code* to foster further research in this domain.

## 1 Introduction

Climate change has emerged as one of the most pressing challenges of our time, not only in scientific and policy circles but also in the public imagination (Change, 2018; Baste et al., 2021; Yusoff and Gabrys, 2011). Gómez-Casillas and Gómez Márquez, 2023 found that YouTube had the

<sup>\*</sup>These authors contributed equally to this work.

Image:



Text: It hits so hard.

Description:

The image and caption convey a powerful message about the urgency of addressing climate change. The visual metaphor of a train colliding with a bus represents the collision between efforts to curb climate change and the reality of the climate crisis. ... The caption "It hits so hard" emphasizes the impact of the climate crisis, ...

Figure 1: CliME sample data: Each data point includes a climate change-related image from a Reddit or Twitter post, the accompanying post text, and a generated description integrating both the image and text.

most significant positive effect on climate awareness in Latin America, followed by Instagram and Twitter, demonstrating the role of these platforms in disseminating climate-related information. Social media platforms like Twitter (now X) and Reddit have emerged as prime spaces for climate discourse, shaping public opinion, mobilizing activism, and influencing policy. Viral campaigns such as **#FridaysForFuture**, **#ShowYourStripes** and **#EarthHour** illustrate the power of social networks in transforming grassroots efforts into global movements. However, these online platforms propagate misinformation and polarization, vastly associated with skepticism, contrarianism, and denial (Treen et al., 2020). In 2023 alone, posts denying climate change on X tripled compared to previous years, highlighting the persistence of climate denial narratives online (Jessica Guynn, 2024). The rise of multimodal content in this domain further complicates this landscape. A flooded city image paired with "Climate policies harm the economy" spreads faster than peer-reviewed data, exploiting visual-emotional resonance (O'Neill, 2020). The rising hype of LLMs show their outstanding capabilities at text generation but their proficiency in

grounding responses in visual-textual context in climate communication, remains unproven. Early studies reveal concerning trends: when prompted with climate-skeptical posts, models like GPT-4 often default to vague, non-committal language. For instance, a study by *NewsGuard* found that GPT-4 (Hurst et al., 2024) was more likely than its predecessor GPT-3.5 (OpenAI, 2022) to generate misinformation when prompted with false narratives, including those about climate change, without sufficient disclaimers or corrections (Sara Fischer, 2023; Stella Levantesi, 2024). This mirrors the "false balance" tactic often used to undermine scientific consensus, where both sides of an issue are presented as equally valid despite overwhelming evidence favoring one side. In addition, automated fact-checking systems for climate-related content, such as CLIMATE-FEVER (Diggelmann, 2021) face challenges while evaluating complex claims. To solve this, recent LLM based frameworks like Climinator (Leippold et al., 2024) integrates multiple scientific viewpoints in the verification process. However, these approaches often assumes that LLMs contain comprehensive understanding of climate science from all relevant perspectives. As a result, there remains a gap in evaluating how well LLMs can assess climate-related contents and generate informed suggestions alongside their fact-checking verdicts.

To address these challenges and assess whether LLMs can interpret and generate credible climate communication, we introduce **CLiME**, a benchmark comprising 2579 pairs of posts scraped from Reddit and Twitter, with most of the data focused on memes, infographics, and skeptic content (see Fig. 1). We then generate descriptions of these images and texts using DeepSeek Janus Pro (Chen et al., 2025), as it has demonstrated proven capability to understand memes and skeptic contents (Sahota, 2025), followed by human annotation (see Section 3.2). These descriptors serve as the basis for evaluating LLMs' capacity to comprehend and address climate change, guided by our proposed three key lenses: **(i) Actionability**, **(ii) Criticality**, and **(iii) Justice** (see Section 4.1). To augment the existing climate-fact-checking pipelines, with the measurement of the LLM responses generated through these lenses, we introduce the **Climate Alignment Quotient (CAQ)**, a novel metric that quantifies gaps across five critical axes: **(i) Articulation**, **(ii) Evidence**, **(iii) Resonance**, **(iv) Transition**, and **(v) Specificity** (see Section 4.2), thereby determining

the extent to which LLMs capture intrinsic climate knowledge. Figure 2 illustrates the entire workflow. In summary, our contributions are as follows:

- (i) A first-of-its-kind multimodal benchmark, **CLiME**, primarily featuring climate change related memes and skeptic content from social media.
- (ii) Three analytical prompting paradigms: **Actionability**, **Criticality**, and **Justice** lenses, designed to investigate LLMs and assess their ability to interpret and generate credible climate discourse.
- (iii) The **Climate Alignment Quotient (CAQ)**, a metric to measure the intrinsic alignment of LLMs in climate communication.

## 2 Related Works

**Climate Communication in Social Media.** (Grundmann and Krishnamurthy, 2010) pioneered the analysis of climate change discourse through textual content, comparing climate-related word frequencies in news articles between Europe and the USA. With the rise of multimedia platforms like Twitter (X), Instagram, TikTok, WhatsApp, and YouTube, social media has been shown to enhance public awareness of climate issues (Farooq et al., 2024). (Abdallah and Youssef, 2023) further demonstrated a positive correlation between climate-related social media content and increased public awareness. Studies highlight the effectiveness of personalized, relatable content in engaging audiences (León et al., 2022), and social media is increasingly viewed as a trusted information source (Hamed Mussa, 2023). However, misinformation and echo chambers remain significant barriers to active engagement (Abdallah and Youssef, 2023). Hence, the analysis of climate change discourse on social media remains a rapidly evolving field that requires further attention.

**LLMs in Climate Change Discourse Analysis.** Recent work has applied LLMs to analyze climate change discourse. ChatREPORT (Ni et al., 2023), using ChatGPT (Hurst et al., 2024) with expert-designed prompts, examined 9,781 corporate sustainability reports to evaluate climate action. (Thulke et al., 2024) introduced ClimateGPT, a domain-specific LLM trained on 300 billion tokens (4.2 billion climate-related), validated through benchmarks and human evaluation. (Zhou et al., 2024) leveraged GPT-4 to uncover latent narratives in climate-related news from North American and Chinese sources.

While (Ni et al., 2023) and (Zhou et al., 2024)

adapt existing LLMs for climate discourse through prompt engineering, they assume the models’ inherent capability to address climate issues. In contrast, Joe et al., 2024 evaluated GPT-4o on climate tasks of varying expertise levels using GAMI literature (Berrang-Ford et al., 2021), revealing limitations in handling high-expertise tasks, especially those involving stakeholder identification and nuanced analysis (Hurst et al., 2024). There remains a gap in multi-perspective assessments of LLM responses across various climate change discourse sources, including news articles and social media.

**Multimodal Understanding in Climate Context.** Multimodal data is increasingly used in climate research for tasks like stance detection, predictive modeling, and video analysis (Dancygier, 2023; Wang et al., 2024a; Mohan and Sinha, 2023; Bai et al., 2024; Wang et al., 2024b). Social media platforms, especially Twitter (X) and Reddit, are key sources of such content, including humorous posts (Kovacheva, 2022; Bai et al., 2024). Recent work has adapted vision-language models (VLMs) like CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) to climate-specific tasks, achieving superior performance in stance detection and misinformation detection (Wang et al., 2024b). Benchmarks like MultiClimate (Wang et al., 2024a) and GreenScreen (Sharma et al., 2024) evaluate models on image–text alignment, narrative coherence, and visual rhetoric in climate-skeptic content. However, they focus mainly on YouTube data, overlooking fast-spreading formats like memes and infographics.

**Assessment Frameworks for Climate Communication.** Three main frameworks have emerged for evaluating climate communication: (i) human-rated with AI support (Bulian et al., 2023), (ii) AI-based evaluation using ChatGPT (Gursesli et al., 2023), and (iii) expert annotation with Likert scales (Nguyen et al., 2024). Bulian et al., 2023 offers a qualitative approach, assessing text style, clarity, tone, and epistemological elements like accuracy and uncertainty. Gursesli et al., 2023 provides a quantitative method focused on narrative quality, measuring coherence, inspiration, and fluency. Nguyen et al., 2024 presents a domain-specific framework using expert ratings to evaluate LLM-generated climate advice for agriculture. However, existing frameworks overlook climate communication via social media and humorous content.

### 3 CliME Dataset

To understand how Large Language Models (LLMs) engage with and comprehend climate discourse, specifically in addressing climate change, we introduce the first-of-its-kind multimodal benchmark, **CliME\*** (**Climate Change Multimodal Evaluation**) dataset. Comprising **2,579** data points, sourced and filtered from real-world Twitter and Reddit posts, primarily memes, skeptics, and infographics, CliME shifts the focus toward evaluating models based on their ability to generate credible, actionable, and equitable climate-related communication with multimodal (image-text) contexts.

#### 3.1 Dataset Creation

The CliME dataset was developed by systematically scraping posts from both Twitter (now X) and Reddit to capture a diverse spectrum of climate related discourse, including both memes and skeptical content. Data collection from X was conducted using Selenium and Twikit (Twikit, 2025), enabling the extraction of posts tagged with climate-related hashtags such as **#climatechange**, **#climatememes**, **#globalwarming**, **#netzero**, and **#climateskeptics**. Reddit data was obtained via the Yet Another Reddit Scraper (YARS) (YARS, 2024) library, specifically targeting posts and comments from climate-focused subreddits such as **r/climatememes** and **r/climateskeptics**. The initial raw dataset comprised approximately  $\sim 4K$  posts, which subsequently underwent a rigorous filtering pipeline to enhance relevance and quality. First, language filtering was applied using the langdetect (Shuyo, 2014) library to exclude non-English texts, ensuring linguistic consistency for downstream analysis. Next, we conduct relevance verification with manual review to eliminate off-topic content, including unrelated memes and spam. Lastly, de-duplication was performed to remove redundant entries; perceptual hashing via ImageHash (Buchner, 2013) was utilized to generate unique fingerprints for images, identifying and flagging near-identical visuals, such as reposted memes with minor modifications. After completing the filtering process, the final dataset comprised **2579** entries, each meeting the criteria for linguistic consistency, relevance, uniqueness, and quality.

\*<https://huggingface.co/datasets/climedataset/CliME>

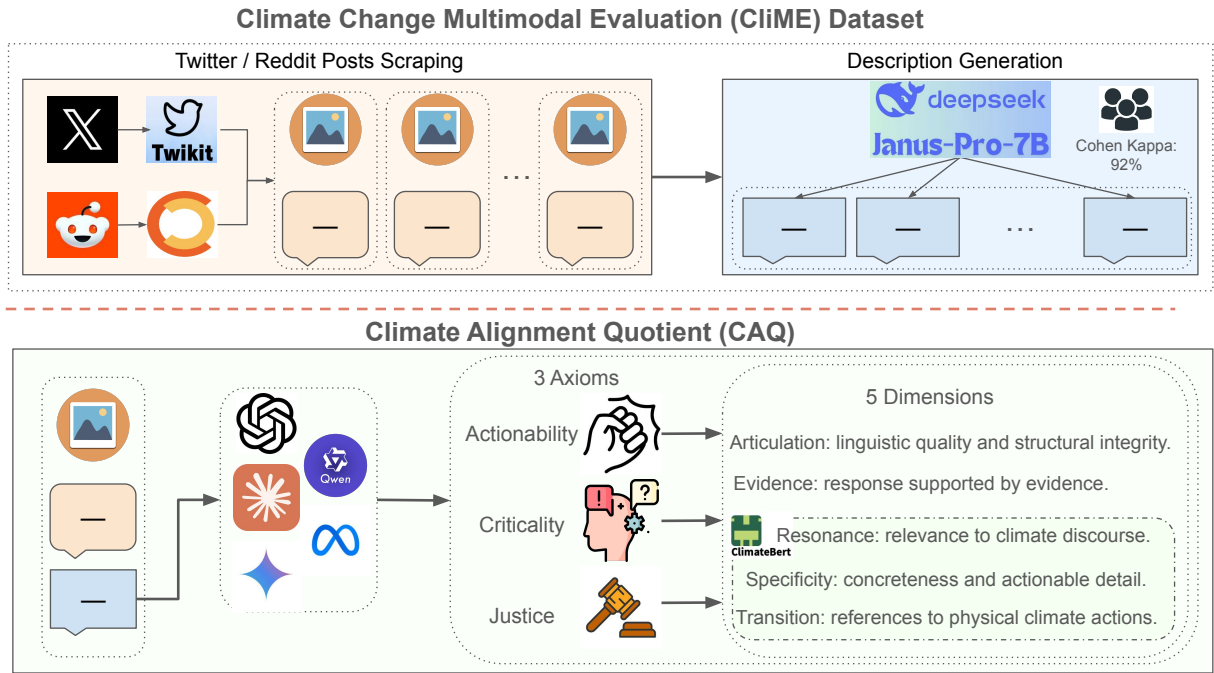


Figure 2: Overview of the *Climate Change Multimodal Evaluation (CliME)* dataset and the *Climate Alignment Quotient (CAQ)* workflow. The upper section illustrates the data collection process from Twitter and Reddit posts, utilizing multimodal sources (text and images) and description generation through the Janus-Pro-7B model followed by human annotations. The lower section demonstrates the CAQ evaluation framework, integrating multimodal data and analytical lenses (Actionability, Criticality, Justice) to assess climate communication across five dimensions: Articulation, Evidence, Resonance, Transition, and Specificity.

### 3.2 Descriptor Generation and Human Annotations

A critical challenge lies in how LLMs process visual content: directly feeding raw pixels to text-based LLMs risks stripping away the nuanced, context-dependent narratives that images convey. This exploits the primal impact of visuals to lend credibility to false claims. In addition, (Sahota, 2025) demonstrated that most VLMs still struggle to understand internet memes in a way that aligns with human interpretation, with the exception of DeepSeek Janus Pro. Hence, to bridge this gap, we employ **DeepSeek Janus Pro**, an autoregressive VLM that unifies multimodal understanding and generation, to process both the text and its accompanying image, generating structured, context-aware descriptions. We configure DeepSeek with a temperature of 0.3 to balance specificity and creativity; max token length of 512. These detailed text descriptors help in assessing LLMs to distinguish fact from manipulation and ensure transparency in how visual context shapes their outputs. Subsequently, for robust verification, we manually annotated the generated descriptors. In this process, we carefully examined whether the combined interpretation of the post’s image and text aligned semanti-

cally with the descriptor’s intended meaning, assigning a score of 1 for a match and 0 for no match. The data was evaluated by two human annotators, and we filtered the data based on their evaluations. The inter-annotator agreement is 92.0% measured by Cohen’s kappa (Cohen, 1960), showing near perfect agreement. Our final dataset comprises triples of image, original text, and generated descriptor, helping in systematic evaluation of assessing LLMs to distinguish fact from manipulation and ensure transparency in how visual context shapes their outputs (cf. Appendix A.2 for examples).

## 4 Climate Alignment Quotient (CAQ)

### 4.1 Analytical Lenses

We evaluate LLMs’ ability to interpret adversarial multimodal narratives, rebut misinformation with scientific rigor, and center justice for marginalized communities in climate descriptors through three novel analytical lenses: (i) **Actionability**, (ii) **Criticality**, and (iii) **Justice** (cf. Appendix A.1). Descriptors are first passed through carefully crafted analytical prompts, which serve as interpretive lenses designed to assess how effectively LLMs engage with crucial climate discourse elements. For instance, consider the descriptor of a

Twitter post as shown in Figure 1: "The image and caption highlight the environmental policy of the Trump administration, specifically focusing on the withdrawal from the Paris Climate Agreement. The post emphasizes the risks associated with this decision, such as increased greenhouse gas emissions and the potential for climate-related disasters. It also highlights the Green New Deal as a potential solution, emphasizing the need for a comprehensive and immediate response to climate change. The post encourages action and emphasizes the urgency of addressing climate issues." Our analytical lenses would evaluate this as follows:

**Actionability** assesses the translation of climate discourse into concrete interventions. In our example, the post's endorsement of the "Green New Deal" as a solution is scrutinized for actionable components. While the proposal advocates for broad objectives like transitioning to "renewable energy", its practical feasibility may be rated as medium due to potential political gridlock and ambiguous funding mechanisms. The message emphasizes urgency but lacks specifics, such as identifying legislative bodies that would *champion the policy*, setting deadlines for key milestones like grid "de-carbonization", or outlining strategies to mobilize workforce training programs. Unaddressed gaps include potential economic strains on industries reliant on fossil fuels and the absence of contingency plans for technological bottlenecks, revealing a disconnect between aspirational goals and practical roadmaps.

**Criticality** examines the structural roots and foundations of climate narratives. The post links the Trump administration's withdrawal from the "Paris Agreement" to increased emissions and climate related disasters. A critical analysis would question this causal oversimplification, noting that while the withdrawal weakened global cooperation, emissions were already rising due to entrenched systems like "fossil fuel" lobbying and inadequate clean energy incentives. Framing the Green New Deal as a singular solution might sidestep debates over its scalability, such as conflicting estimates about job creation or its silence on nuclear energy's role. This uncritical portrayal risks reinforcing partisan divides rather than addressing systemic barriers like corporate influence on *climate policy*.

**Justice** centers on marginalized voices and systemic inequities. The post's U.S. (United States)-centric focus overlooks how the withdrawal from the Paris Agreement undermined climate financing

for "Global South" nations, exacerbating vulnerabilities in regions least responsible for emissions. By positioning the Green New Deal as a domestic fix, the narrative neglects historical U.S. accountability for global emissions and fails to address how transition costs might disproportionately affect low income communities, such as rising energy prices or displacement from renewable infrastructure projects. A justice lens would highlight absent voices, such as Indigenous groups advocating for land sovereignty in solar farm expansions, and question whether the policy redistributes power or perpetuates existing inequities.

Following evaluation through these analytical lenses, the resulting outputs provide essential context and data to progress into our quantitative evaluation step: calculating the **Climate Alignment Quotient (CAQ)**.

## 4.2 Climate Alignment Quotient (CAQ)

The **CAQ** is a composite metric specifically designed to quantify the effectiveness and alignment of climate-related communication generated by LLMs. Integrating and systematically evaluating outputs derived from the analytical lenses (Actionability, Criticality, Justice), the CAQ assesses alignment across five critical dimensions: **(i) Articulation, (ii) Evidence, (iii) Resonance, (iv) Transition, and (v) Specificity**. The CAQ score is calculated as a weighted sum of the five core metrics, mathematically can be described as:

$$\begin{aligned} \text{CAQ} = & w_1 \cdot \text{Articulation} + w_2 \cdot \text{Evidence} \\ & + w_3 \cdot \text{Resonance} + w_4 \cdot \text{Transition} \\ & + w_5 \cdot \text{Specificity} \end{aligned} \quad (1)$$

where  $w_1 + w_2 + w_3 + w_4 + w_5$  aggregates to 1. Empirically, we set  $w_1 = 0.3$ ,  $w_2 = 0.2$ ,  $w_3 = 0.25$ ,  $w_4 = 0.15$ , and  $w_5 = 0.1$ .

**(i) Articulation:** Our articulation score measures the linguistic quality and structural integrity of climate communications through a dual component analysis. The articulation score is calculated as an equally weighted combination of coherence and completeness, thereby ensuring each component contributes equally to the measure, given by:

$$\text{Articulation} = 0.5 \cdot \text{Coherence} + 0.5 \cdot \text{Completeness} \quad (2)$$

The *coherence* component assesses how well sentences connect and flow together and is further broken down into:

$$\text{Coherence} = 0.6 \cdot \text{Syntactic} + 0.4 \cdot \text{Semantic} \quad (3)$$

*Syntactic* coherence quantifies discourse markers (e.g., coordinating/subordinating conjunctions, "mark" dependencies) relative to sentence count, indicating structural connectedness. *Semantic* coherence uses the all-mpnet-base-v2 Sentence Transformer (Song et al., 2020) to compute embedding similarity between adjacent sentences, capturing thematic continuity. We found that weights of 0.6 (semantic) and 0.4 (syntactic) best aligned with human coherence judgments. *Completeness* assesses grammatical integrity by identifying valid constructions: complete clauses (subject + predicate) and imperatives (action-verb starters). Using SpaCy (Honnibal and Montani, 2017), we detect subjects via "nsubj" or "nsubjpass" and predicates via ROOT verbs. The completeness score is the fraction of sentences matching these patterns, reflecting grammatical consistency in climate communication.

**(ii) Evidence:** The Evidence metric distinctly assesses the extent to which the climate communications in the LLM’s responses are substantiated by verifiable data and concrete examples. We employ the specialized NLP model "*climate-nlp/longformer-large-4096-1-detect-evidence*" (Morio and Manning, 2023) to compute this score. A high Evidence score indicates that the language model’s claims are supported by detailed, traceable evidence, such as explicit data points, clear references, and documented instances of climate action demonstrating a robust and genuine commitment to addressing climate challenges. In contrast, a low score reveals that the communications lack sufficient backing, potentially suggesting superficial engagement or greenwashing.

The remaining three CAQ components are derived via a suite of models from ClimateBERT (Webersinke, 2021), pretrained on a text corpus comprising climate related research paper abstracts, corporate and general news and reports from companies. The measures derived from this collection are as follows:

**(iii) Resonance:** The Resonance metric leverages the "*climatebert/distilroberta-base-climate-detector*" model to quantify how strongly content engages with climate context. This model performs binary classification, outputting a probability score

that represents the likelihood of climate relevance. Higher scores indicate content that more directly addresses climate change concepts, terminology, and themes. The detector performs well at identifying subtle climate references while filtering out environmental content unrelated to climate change.

**(iv) Transition:** The Transition metric uses the "*climatebert/transition-physical*" model to evaluate references to physical climate transition processes. It identifies content related to tangible climate adaptation and mitigation practices, infrastructure changes, and transitions in physical systems. It detects mentions of renewable energy implementation, carbon capture technologies, climate-resilient infrastructure, and other physical interventions. Higher scores indicate content that addresses concrete transitional mechanisms rather than abstract climate concepts or general environmental concerns.

**(v) Specificity:** The Specificity metric utilizes the "*climatebert/distilroberta-base-climate-specificity*" model to assess whether given response by the model provides specific, actionable information versus general statements or vague recommendations. Higher specificity scores indicate content containing concrete actions, measurable targets, defined timelines, or detailed examples. This dimension is crucial for distinguishing between aspirational climate rhetoric and communications that provide implementable guidance or precise information that can drive meaningful action in real life.

In summary, CAQ serves as a robust metric for evaluating the effectiveness of climate communications in LLMs. Higher CAQ scores indicate communications that strongly align with climate objectives, characterized by coherent articulation, high resonance, strong evidence, concrete transitional strategies, and actionable specificity; hence, aligning communication strategies of LLMs with climate objectives.

### 4.3 Evaluation

We benchmarked five state-of-the-art LLMs on CliME and evaluated them using our CAQ metric. The models include GPT-4o (Hurst et al., 2024), LLaMA 3.3 70B (Dubey et al., 2024), Gemini 2.0 Flash (Team, 2024), Qwen QwQ (Qwen with Question) 32B (Yang et al., 2024), and Claude 3.7 Sonnet (Anthropic, 2024). All models were configured to generate outputs at a temperature of 0.1 to ensure deterministic responses. The scores are reported in Table 1; a comprehensive analysis

Table 1: Climate Alignment Quotient (CAQ) Comparison Across Different LLMs. Color intensity indicates performance level, with deeper colors representing higher scores. Column headers are color-coded by lens (yellow for Actionability, red for Criticality, green for Justice), while component scores use blue shades proportional to their values.

	Actionability						Criticality						Justice					
Model	R	A	E	T	Sp	CAQ	R	A	E	T	Sp	CAQ	R	A	E	T	Sp	CAQ
GPT-4o	1.0	.86	.68	.19	.18	<b>.69</b>	1.0	.81	.85	.15	.18	<b>.70</b>	1.0	.81	.98	.18	.15	<b>.73</b>
Claude 3.7 Sonnet	1.0	.82	.95	.18	.20	<b>.73</b>	1.0	.84	.97	.17	.21	<b>.74</b>	1.0	.83	1.0	.11	.18	<b>.73</b>
Gemini 2.0 Flash	1.0	.74	.84	.18	.15	<b>.68</b>	1.0	.69	.78	.17	.20	<b>.66</b>	1.0	.82	.99	.14	.15	<b>.73</b>
LLaMA 3.3 70b	1.0	.72	.74	.21	.15	<b>.66</b>	1.0	.78	.86	.14	.19	<b>.70</b>	1.0	.84	1.0	.15	.19	<b>.74</b>
Qwen QwQ 32b	.99	.81	.80	.20	.20	<b>.70</b>	0.99	.77	.86	.17	.22	<b>.70</b>	1.0	.77	1.0	.19	.18	<b>.73</b>

Legend: R = Resonance, A = Articulation, E = Evidence, T = Transition, Sp = Specificity  
Weights: Resonance (0.25), Articulation (0.30), Evidence (0.20), Transition (0.15), Specificity (0.1)

of the CAQ scores across the models along our proposed lenses is shown in Figure 3. We find from our experiments that resonance (R) scores are consistently high (near 1.0) across all models, indicating a strong alignment with climate-related context in the generated descriptors. In terms of *Actionability*, Claude 3.7 Sonnet achieves an average CAQ of 0.73, while Gemini 2.0 Flash and LLaMA 3.3 70B record slightly lower scores of 0.68 and 0.66, respectively; GPT-4o and Qwen both score around 0.70. The *Criticality* lens shows Claude and LLaMA performing comparably (0.74 and 0.70, respectively), with Gemini 2.0 Flash trailing at 0.66. For the *Justice* dimension, all models converge between 0.73 and 0.74. This uniformity in Justice scores suggests that fairness and equity considerations are consistently addressed, while the variability in Actionability and Criticality highlights differences in the models’ abilities to generate concrete, actionable guidance and critically evaluative discourse. The articulation (A) measure further reveals that GPT-4o (0.86) and Claude 3.7 Sonnet (0.82) offer more coherent communication compared to Gemini 2.0 Flash (0.74) and LLaMA 3.3 70B (0.72). Moreover, the evidence (E) scores are particularly high for Claude 3.7 Sonnet and LLaMA 3.3 70B, indicating that these models more effectively ground their outputs with verifiable data. In addition to this, transition scores are consistently low (ranging from 0.11 to 0.21) across all models, suggesting a common challenge in referencing physical climate transition processes. Overall, while all models demonstrate a balanced under-

standing of climate discourse with overall CAQ scores hovering between 0.70 and 0.74, Claude 3.7 Sonnet and GPT-4o tend to produce more articulate and evidence-supported discourse, whereas Gemini 2.0 Flash shows slightly lower performance overall (cf. Appendix A.5 for qualitative analysis).

#### 4.4 Analysis

In this section, we present a comprehensive analysis of the CAQ scores across the proposed lenses of **Actionability**, **Criticality**, and **Justice**. This analysis encompasses both the distribution of individual CAQ scores and the evaluation of gaps between these dimensions, providing a detailed understanding of the multimodal climate discourse generated by LLMs.

##### 4.4.1 Distribution of CAQ Scores

In Figure 4, where the x-axis is *Criticality*, the y-axis is *Justice*, and the z-axis is *Actionability*, for Claude 3.7 Sonnet (cf. Appendix A.4 for other models), we observe that most points cluster around the mid-range of each axis (approximately 0.70-0.80), indicating that the LLM-generated content tends to balance urgency (Criticality), fairness (Justice), and practicality (Actionability). When *Actionability* increases (both in the z-coordinate and in the color scale), there is often a slight upward score in both *Criticality* and *Justice*, suggesting that more action-oriented content tends to incorporate at least moderate levels of urgency and equity considerations. Conversely, at lower *Actionability* scores (cooler colors), there’s a wider spread in *Criticality* and *Justice*, suggesting that content

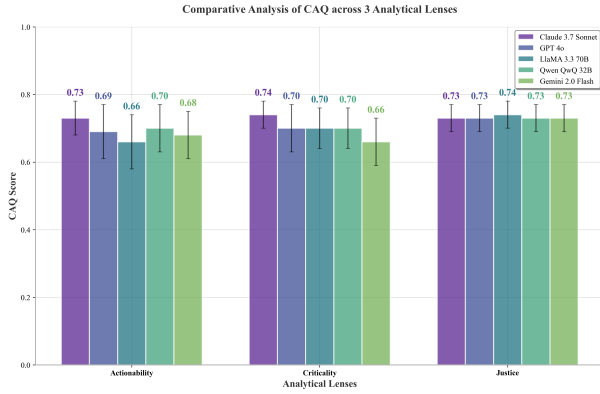


Figure 3: Comparative CAQ scores across the *Actionability*, *Criticality*, and *Justice* lenses for five LLMs on *CliME*. Each bar shows the mean CAQ score with error bars denoting standard deviation. Claude 3.7 Sonnet consistently outperforms others, scoring above 0.70 across lenses with low variability.

with fewer calls to action is perceived with more varied urgency and equity.

#### 4.4.2 Gap Analysis

In addition to this, we perform a gap analysis to evaluate the differences between CAQ scores across our proposed axes (cf. Appendix A.3 for box plots and heatmaps for all the models). Effective climate communication necessitates a balanced integration of these aspects. Overemphasis on one dimension, such as *Actionability*, at the expense of others like *Justice*, can lead to skewed narratives that overlook systemic inequities or fail to motivate comprehensive action. In the case of Claude 3.7 Sonnet, one of the best scoring models on CAQ, we empirically find that the average gap between dimensions varies across the 2,579 evaluated descriptors. The gap between *Criticality* and *Justice* is the smallest at 0.0313, showing these dimensions are generally well-aligned. However, the gap between *Actionability* and *Justice* tends to be larger, averaging around 0.0344, suggesting a slight underrepresentation of justice considerations in certain outputs. Similarly, the *Actionability-Criticality* gap averages about 0.0324. In case of Claude 3.7 Sonnet, our dimensional frequency analysis reveals that the most common largest gap type is between *Actionability* and *Criticality*, occurring in 37.1% of descriptors. These variations across different posts provide valuable insights for developing more holistic approaches to climate communication that effectively balance considerations across these three dimensions.

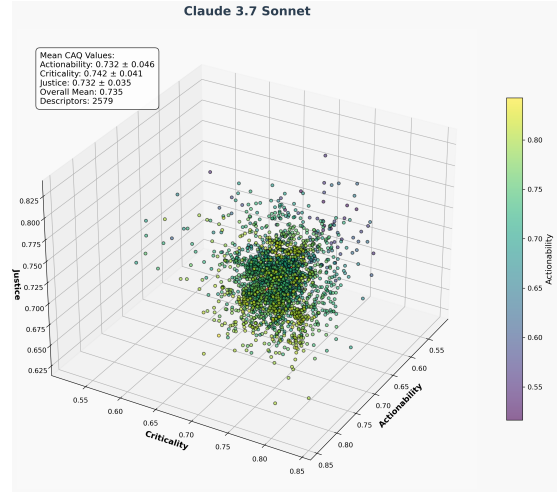


Figure 4: 3D scatter plot of CAQ scores for the Claude 3.7 Sonnet model on the *CliME* dataset. Each point represents a description’s CAQ values across *Actionability* (x), *Criticality* (y), and *Justice* (z). Color denotes *Actionability* score. Points near the center reflect balanced discourse; deviations indicate over- or under-emphasis on specific lenses.

## 5 Conclusion

In this work, we present *CliME*, a novel multimodal dataset curated from social media, primarily featuring memes and skeptics, and introduce the *Climate Alignment Quotient (CAQ)* as a comprehensive metric to evaluate climate discourse generated by LLMs. By benchmarking five state-of-the-art models, we analyzed their outputs across three lenses: *Actionability*, *Criticality*, and *Justice*. Our experiments reveal that while all models consistently capture climate-related context, significant variations exist in generating actionable and critically evaluative messages. The gap analysis uncovers subtle imbalances among these dimensions, indicating areas for improvement in the models’ outputs. In today’s landscape, misinformation and polarized narratives on climate change contribute to social harm and undermine effective action; our CAQ framework offers a promising approach to understanding the strengths and weaknesses of LLMs. By fostering balanced climate communication, our work aims to prevent the spread of misleading information and support informed public dialogue, ultimately guiding policy-making for social good. Future efforts will focus on refining the CAQ metric and expanding the *CliME* dataset to include a broader spectrum of multimodal content, thereby empowering both LLMs and VLMs to contribute more effectively to a sustainable future.



## Limitations

Though the proposed CAQ framework provides a structured and comprehensive way to evaluate climate discourse, it currently relies on existing pre-trained models for its assessment. Since these models may not be extensively trained on the latest social media data, particularly memes and other highly contextual content, there is a risk of missing nuanced climate change signals. Additionally, the nature of social media platforms, where language evolves rapidly and memes can quickly become outdated or repurposed, presents challenges in ensuring that all relevant domain-specific shifts are captured.

## Ethics Statement

All data in our CLiME dataset originate from publicly accessible Reddit and Twitter posts. We strictly followed platform guidelines during data collection, focusing on content explicitly marked for public sharing and ensuring that no personally identifiable information (PII) was retained. Although memes, infographics, and other materials often exhibit strong emotional or political underpinnings, our goal is to assess climate-related discourse rather than endorse any particular viewpoint. The proprietary models employed in our study were accessed strictly via valid subscriptions, in accordance with the terms of service provided by the respective providers.

## Acknowledgments

We thank the reviewers for their insightful comments and suggestions, which have significantly improved the quality of this paper. We gratefully acknowledge support from the Texas A&M Institute of Data Science via an internal grant. Portions of this work were conducted using the advanced computing resources provided by Texas A&M High-Performance Research Computing (HPRC). We also extend our gratitude to Soham Wasmakar and Aarhi Borah for their assistance with the dataset annotation process.

## References

Manar Mohamed Abdallah and Mostafa Youssef. 2023. *Role of Social Media in Enhancing Climate Change Literacy Among Egyptian Users*. 2023(25):1–25.

Anthropic. 2024. *Claude 3.7 Sonnet and Claude Code*.

Nan Bai, Ricardo da Silva Torres, Anna Fensel, Tamara Metze, and Art Dewulf. 2024. Inferring climate change stances from multimodal tweets. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2467–2471.

Ivar A Baste and 1 others. 2021. Making peace with nature: a scientific blueprint to tackle the climate, biodiversity and pollution emergencies.

Lea Berrang-Ford, AR Siders, Alexandra Lesnikowski, Alexandra Paige Fischer, Max W Callaghan, Neal R Haddaway, Katharine J Mach, Malcolm Araos, Mohammad Aminur Rahman Shah, Mia Wannewitz, and 1 others. 2021. A systematic global stocktake of evidence on human adaptation to climate change. *Nature climate change*, 11(11):989–1000.

Johannes Buchner. 2013. *Imagehash: A python perceptual image hashing module*.

Jannis Bulian, Mike S Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Huebscher, Christian Buck, Niels G Mede, Markus Leippold, and 1 others. 2023. Assessing large language models on climate information. *arXiv preprint arXiv:2310.02932*.

Projected Climate Change. 2018. Global warming of 1.5 c. *World Meteorological Organization: Geneva, Switzerland*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Barbara Dancygier. 2023. Multimodal media: Framing climate change. *Discourse Studies*, 25(2):220–236.

T. et al. Diggelmann. 2021. Climate-fever: A dataset for verification of real-world climate claims. In *Proceedings of NeurIPS Datasets and Benchmarks*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. *The llama 3 herd of models*. *arXiv preprint arXiv:2407.21783*.

Muhammad Waqas Farooq, Abdul Rauf, Raja Irfan Sabir, Faiza Nawaz, and 1 others. 2024. How do social media platforms shape the public perception and support of policy issues and initiatives in climate change? *Bulletin of Business and Economics (BBE)*, 13(2):1018–1025.

Amalia Gómez-Casillas and Victoria Gómez Márquez. 2023. The effect of social network sites usage in climate change awareness in latin america. *Population and environment*, 45(2):7.

- Reiner Grundmann and Ramesh Krishnamurthy. 2010. The discourse of climate change: A corpus-based approach. *Critical approaches to discourse analysis across disciplines*, 4(2):125–146.
- Mustafa Can Gursesli, Pittawat Taveekitworachai, Febri Abdullah, Mury F Dewantoro, Antonio Lanata, Andrea Guazzini, Van Khôi Lê, Adrien Villars, and Ruck Thawonmas. 2023. The chronicles of chatgpt: Generating and evaluating visual novel narratives on climate change through chatgpt. In *International Conference on Interactive Digital Storytelling*, pages 181–194. Springer.
- Mona Hamed Mussa. 2023. [The Impact of Social Media Influence on Attitudes toward Sustainable Consumption "A Mediation Analysis of Climate Change Awareness" An applied Study on Higher Education Students in Egypt](#). 14(3):975–1024.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. [Gpt-4o system card](#).
- USA TODAY Jessica Guynn. 2024. [Science vs. social media: Why climate change denial still thrives online](#).
- Elphin Tom Joe, Sai Dileep Koneru, and Christine J Kirchoff. 2024. Assessing the effectiveness of gpt-4o in climate change evidence synthesis and systematic assessments: Preliminary insights. *arXiv preprint arXiv:2407.12826*.
- A. et al. Kovacheva. 2022. Climate change memes on social media: Engagement, moderation, and skepticism. *Environmental Communication*, 16:345–362.
- Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, and 1 others. 2024. Automated fact-checking of climate change claims with large language models. *arXiv preprint arXiv:2401.12566*.
- Bienvenido León, Samuel Negredo, and María Carmen Erviti. 2022. Social engagement with climate change: principles for effective visual representation on social media. *Climate Policy*, 22(8):976–992.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- Sankaralingam Mohan and Akash Sinha. 2023. Multimodal climate change prediction in a monsoon climate. *Journal of Water and Climate Change*, 14(9):2919–2934.
- Gaku Morio and Christopher D Manning. 2023. [An nlp benchmark dataset for assessing corporate climate policy engagement](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 39678–39702. Curran Associates, Inc.
- Vincent Nguyen, Sarvnaz Karimi, Willow Hallgren, Ashley Harkin, and Mahesh Prakash. 2024. My climate advisor: An application of nlp in climate adaptation for agriculture. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 27–45.
- Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, and 1 others. 2023. Chatreport: Democratizing sustainability disclosure analysis through llm-based tools. *arXiv preprint arXiv:2307.15770*.
- OpenAI. 2022. [GPT 3.5 Turbo](#).
- Saffron O’Neill. 2020. More than meets the eye: A longitudinal analysis of climate change imagery in the print media. *Climatic Change*, 163(1):9–26.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.
- Harpreet Sahota. 2025. [Memes are the vlm benchmark we deserve](#).
- Axios Sara Fischer. 2023. [Exclusive: Gpt-4 readily spouts misinformation, study finds](#).
- Ujjwal Sharma, Stevan Rudinac, Joris Demmers, Willemijn van Dolen, and Marcel Worring. 2024. Greenscreen: A multimodal dataset for detecting corporate greenwashing in the wild. In *International Conference on Multimedia Modeling*, pages 96–109. Springer.
- Nakatani Shuyo. 2014. [Langdetect: Port of google’s language-detection library to python](#).
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MpNet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.
- DeSmog Stella Levantesi. 2024. [The abcs of ai and environmental misinformation](#).
- G Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024). URL <https://go.gle/GeminiV1-5>.

David Thulke, Yingbo Gao, Petrus Pelsler, Rein Brune, Richa Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, and 1 others. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*.

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

Twikit. 2025. [Twikit: A simple twitter api scraper](#).

Jiawen Wang, Longfei Zuo, Siyao Peng, and Barbara Plank. 2024a. Multiclimat: Multimodal stance detection on climate change videos. *arXiv preprint arXiv:2409.18346*.

Jiawen Wang, Longfei Zuo, Siyao Peng, and Barbara Plank. 2024b. [MultiClimate: Multimodal stance detection on climate change videos](#). In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 315–326, Miami, Florida, USA. Association for Computational Linguistics.

N. et al. Webersinke. 2021. Climatebert: A pretrained language model for climate-related nlp tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4369–4379.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

YARS. 2024. [Yars: Yet another reddit scraper](#).

Kathryn Yusoff and Jennifer Gabrys. 2011. Climate change and the imagination. *Wiley Interdisciplinary Reviews: Climate Change*, 2(4):516–534.

Haiqi Zhou, David Hobson, Derek Ruths, and Andrew Piper. 2024. Large scale narrative messaging around climate change: A cross-cultural comparison. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 143–155.

## A Appendix

### A.1 Analytical Lenses

We present the prompts used for our three proposed lenses: *Actionability*, *Criticality*, and *Justice*.

#### Actionability

**Description:** {description}

**Instruction:** Analyze the climate-related message in the above description through an actionability lens. Respond in one unified paragraph that summarizes the key climate issues, identifies actionable solutions, evaluates their feasibility (high/medium/low), assesses explicit commitments (who, what, when, how), and highlights risks or unaddressed challenges. Do not output any extra information other than this analysis in your response.

#### Criticality

**Description:** {description}

**Instruction:** Critically evaluate the climate-related message in the above description. Respond in one unified paragraph that identifies core claims, assesses evidence quality, highlights unsubstantiated claims or oversimplified arguments, evaluates engagement with competing viewpoints, and analyzes its impact on climate discourse. Do not output any extra information other than this analysis in your response.

#### Justice

**Description:** {description}

**Instruction:** Analyze the climate-related message in the above description through a justice lens. Respond in one unified paragraph that identifies centered/absent communities, assesses distribution of responsibility, evaluates acknowledgment of historical power imbalances, examines impacts on marginalized groups, and considers inclusion of cultural contexts. Do not output any extra information other than this analysis in your response.

### A.2 CLiME Examples

Table 2 presents various Twitter and Reddit posts, alongside their corresponding images, textual content, and the automatically generated descriptions produced by Janus Pro filtered after human annotations.

### A.3 Gap Analysis

Figures 5, 6, 7, 8, and 9 collectively present a gap analysis of the CAQ scores. These figures visualize the pairwise differences between the *Actionability*, *Criticality*, and *Justice* dimensions using box plots and heatmaps across five of our experimented models.

### A.4 Distribution of CAQ Scores for other models

We visualize the distribution of language model responses across the three CAQ dimensions, using 3D scatter plots. Table 3 presents these distributions for four state-of-the-art language models: GPT-4o, LLaMA 3.3 70B, Qwen QwQ 32B, and Gemini 2.0 Flash. Each point represents a single response, with its position determined by the three CAQ dimensions and color intensity corresponding to Actionability scores. This visualization allows us to identify patterns in how these models balance critical discourse, justice orientation, and actionable content when discussing climate change issues.

### A.5 Qualitative Analysis

To complement our quantitative evaluation, we conducted a small human fluency study to assess LLM response quality across our proposed dimensions (cf. Table 4). We analyzed model outputs using a color-coded annotation system highlighting effective elements (shown in green) and areas requiring improvement (shown in red).

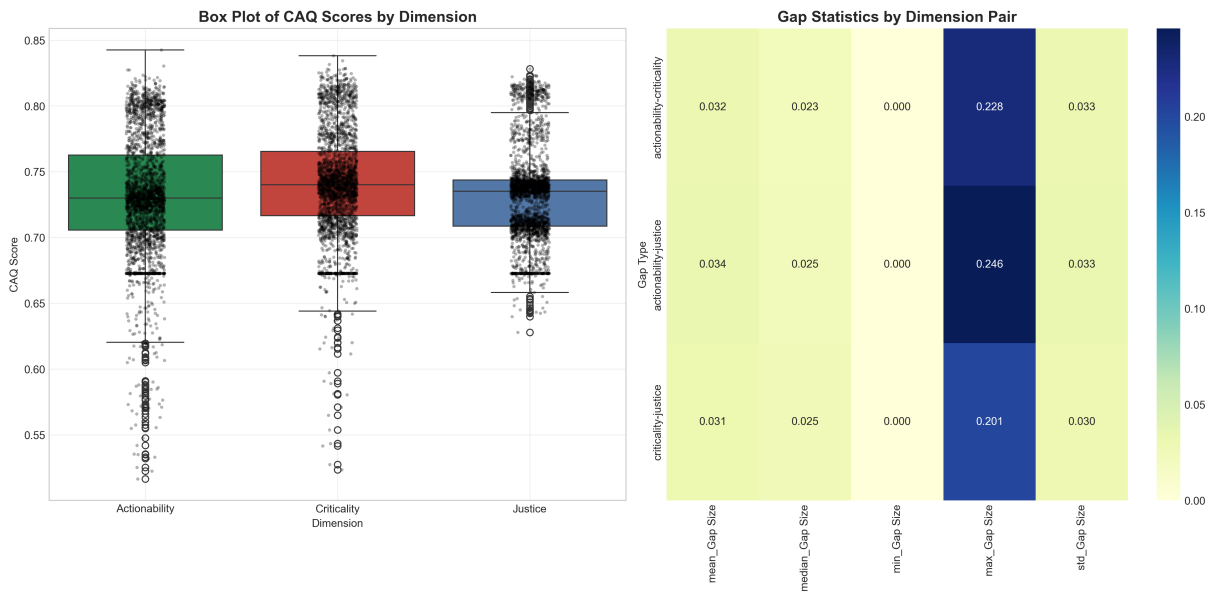








Figure 5: Gap Analysis for Claude 3.7 Sonnet’s CAQ Score Performance. The left panel shows box plots of scores across three dimensions: Actionability (mean: 0.7321), Criticality (mean: 0.7416), and Justice (mean: 0.7321). The right panel displays a heatmap of gap statistics between dimension pairs, with the Actionability-Justice gap (0.0344) being the most significant, followed by the Actionability-Criticality gap (0.0324), while the Criticality-Justice gap (0.0313) shows the best balance. The analysis reveals more balanced dimensional scores compared to other models, with fewer large gaps across all dimension pairs.



Figure 6: Gap Analysis for GPT-4o’s CAQ scores based on 2579 descriptors. The left panel shows box plots of scores across three dimensions: Actionability (mean: 0.6895), Criticality (mean: 0.7033), and Justice (mean: 0.7312). The right panel displays a heatmap of gap statistics between dimension pairs, with the Actionability-Justice gap (0.0704) being the most prominent, followed by the Actionability-Criticality gap (0.0608), while the Criticality-Justice gap (0.0514) shows the best balance. Darker colors in the heatmap indicate larger dimensional gaps.

Table 2: More data samples from CliME.

No.	Reddit/Twitter Post	Description
1	<p data-bbox="405 465 703 510">With extreme climate situations all around the world, this is all of us</p>  <p data-bbox="480 730 630 752">So true it hurts :'_</p>	<p data-bbox="842 465 1386 636">The image and caption highlight the severe impact of climate change on communities, particularly vulnerable populations. The burning house in the background symbolizes the increasing frequency and intensity of natural disasters, such as wildfires, which are exacerbated by climate change. The caption emphasizes the shared responsibility of addressing these challenges, urging collective action to mitigate the risks and protect communities. ...</p>
2	 <p data-bbox="424 1032 687 1055">And 2026 and 2027 and so on ...</p>	<p data-bbox="842 768 1386 1014">The image and caption highlight the alarming reality of Earth's climate, specifically the record-breaking heat experienced during the summer of 2025. The caption emphasizes the urgency of addressing climate change, stating that the second year in a row has been the hottest on record. It underscores the risks associated with climate change, such as extreme heat events, and calls for actionable solutions to mitigate these risks. The image of two individuals, one looking distressed and the other looming ominously, symbolizes the dire consequences of inaction and the need for immediate and collective action to protect the planet.</p>
3	<p data-bbox="363 1070 544 1115"><b>TALKING ABOUT CLIMATE CHANGE</b></p>  <p data-bbox="384 1267 528 1312">« OH NO WE MUST DO SOMETHING! »</p> <p data-bbox="571 1070 746 1115"><b>MENTIONING THE THINGS TO DO:</b></p>  <p data-bbox="592 1267 735 1312">« WELL NOW I'M NOT DOING IT »</p> <p data-bbox="424 1312 687 1330">Meat/car/air travel are so taboo</p>	<p data-bbox="842 1064 1386 1288">The image and caption highlight the urgent need for action in addressing climate change. The left side of the image shows a character expressing concern and urgency, with the text "OH NO WE MUST DO SOMETHING!!" emphasizing the need for immediate action. The right side of the image shows a character sitting with crossed arms, indicating a lack of action, with the text "WELL NOW I'M NOT DOING IT." This contrast underscores the importance of taking concrete steps to mitigate climate change. ...</p>
4	 <p data-bbox="320 1536 794 1554">What a gas! How serious climate action can be funny, too</p>	<p data-bbox="842 1346 1386 1563">The image depicts a cartoon of the Earth, anthropomorphized and wearing a hat, sitting in a room engulfed in flames. The Earth character appears calm and unfazed, sipping from a mug. The caption reads, "This is fine." The climate-related message in this post is a satirical commentary on the lack of urgency and action in addressing climate change. The phrase "This is fine" is often used to dismiss serious issues, and in this context, it highlights the complacency and denial that can hinder meaningful climate action...</p>
5	 <p data-bbox="300 1839 815 1881">What 30 million tonnes of ice that Greenland loses every hour looks like</p>	<p data-bbox="842 1581 1386 1742">The image depicts a massive iceberg, approximately 30 million tonnes in size, floating in the Arctic Ocean. This visual representation highlights the alarming rate at which Greenland is losing ice, with an estimated 100 billion tonnes of ice melting every hour. The caption emphasizes the urgent need for climate action, as the loss of ice threatens global sea levels, ecosystems, and human societies ...</p> <p data-bbox="842 1753 1386 1821">In conclusion, the post serves as a powerful reminder of the urgent need for climate action and emphasizes the importance of taking immediate and collective steps to address the climate crisis.</p>

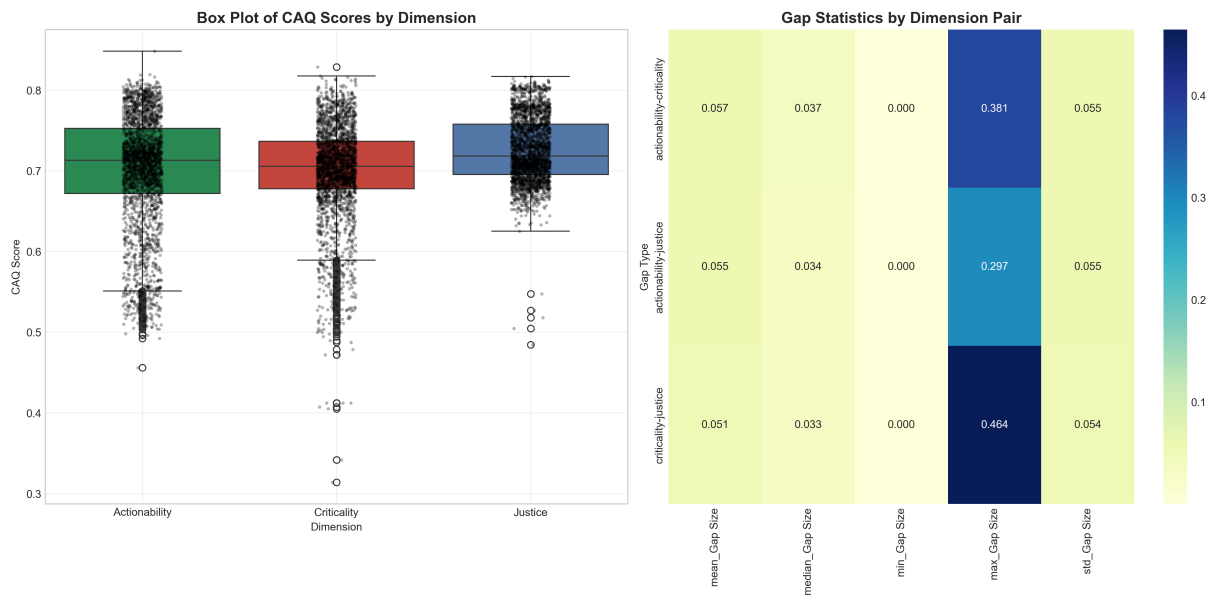


Figure 7: Gap Analysis for Qwen QwQ 32B's CAQ Score Performance. The left panel shows box plots of scores across three dimensions: Actionability (mean: 0.7003), Criticality (mean: 0.6990), and Justice (mean: 0.7257). The right panel displays a heatmap of gap statistics between dimension pairs, with the Actionability-Criticality gap (0.0572) being the most significant, followed by the Actionability-Justice gap (0.0548), while the Criticality-Justice gap (0.0510) shows the best balance. The analysis reveals a more pronounced variance in dimensional scores compared to other models, with a high number of messages falling into the large gap category across all dimension pairs.

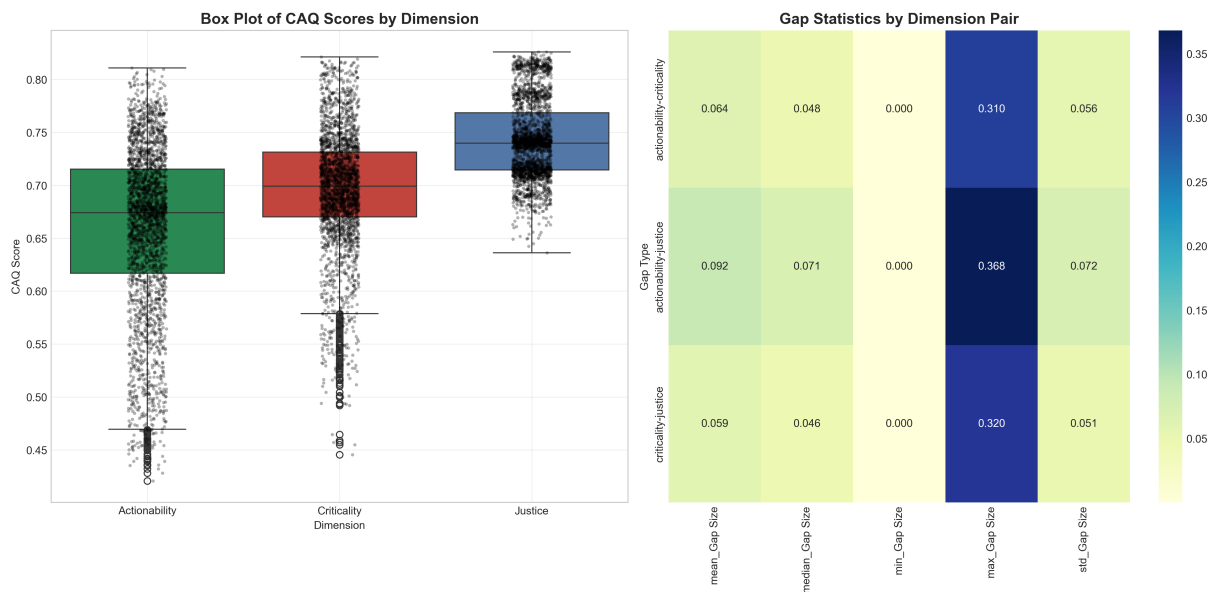


Figure 8: Gap Analysis for LLaMA 3.3 70B's CAQ Score Performance. The left panel shows box plots of scores across three dimensions: Actionability (mean: 0.6591), Criticality (mean: 0.6958), and Justice (mean: 0.7432). The right panel displays a heatmap of gap statistics between dimension pairs, with the Actionability-Justice gap (0.0918) being the most substantial, followed by the Actionability-Criticality gap (0.0639), while the Criticality-Justice gap (0.0590) shows the best balance. The analysis reveals a significant discrepancy between Actionability and Justice dimensions, with 520 posts (20% of the dataset) falling into the large gap category for this dimension pair.

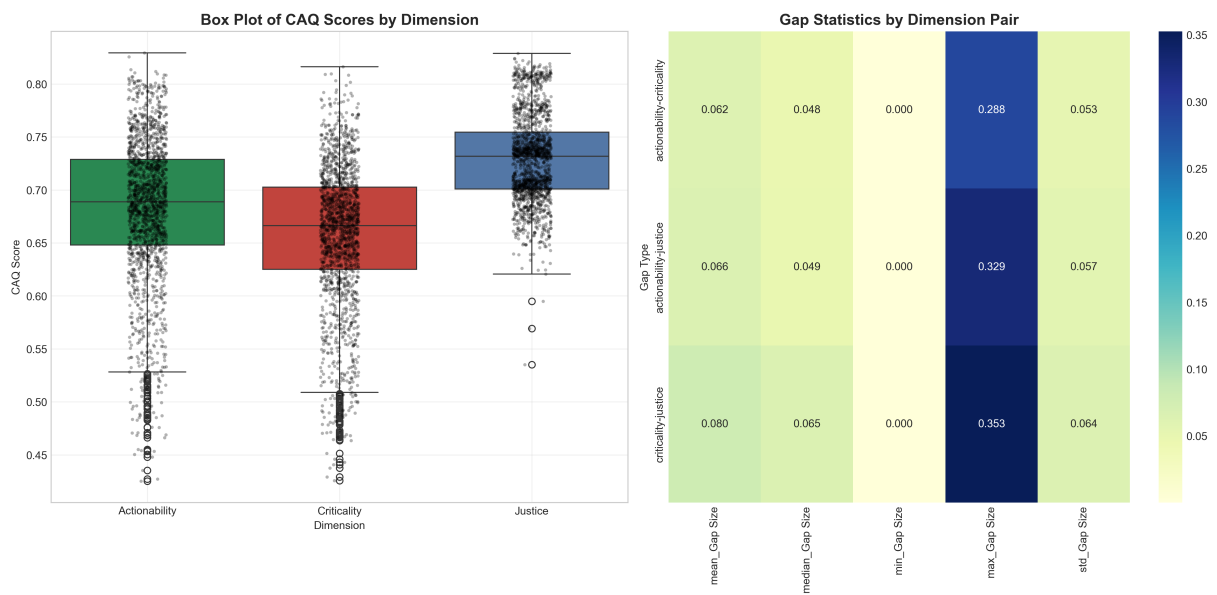


Figure 9: Gap Analysis for Gemini 2.0 Flash's CAQ Score Performance. The left panel shows box plots of scores across three dimensions: Actionability (mean: 0.6823), Criticality (mean: 0.6578), and Justice (mean: 0.7303). The right panel displays a heatmap of gap statistics between dimension pairs, with the Criticality-Justice gap (0.0803) being the most pronounced, followed by the Actionability-Justice gap (0.0655), while the Actionability-Criticality gap (0.0622) shows the best balance. The analysis reveals a significant discrepancy between Justice and Criticality dimensions, with 247 posts falling into the large gap category for this dimension pair, indicating a potential area for improvement in balanced climate communication.



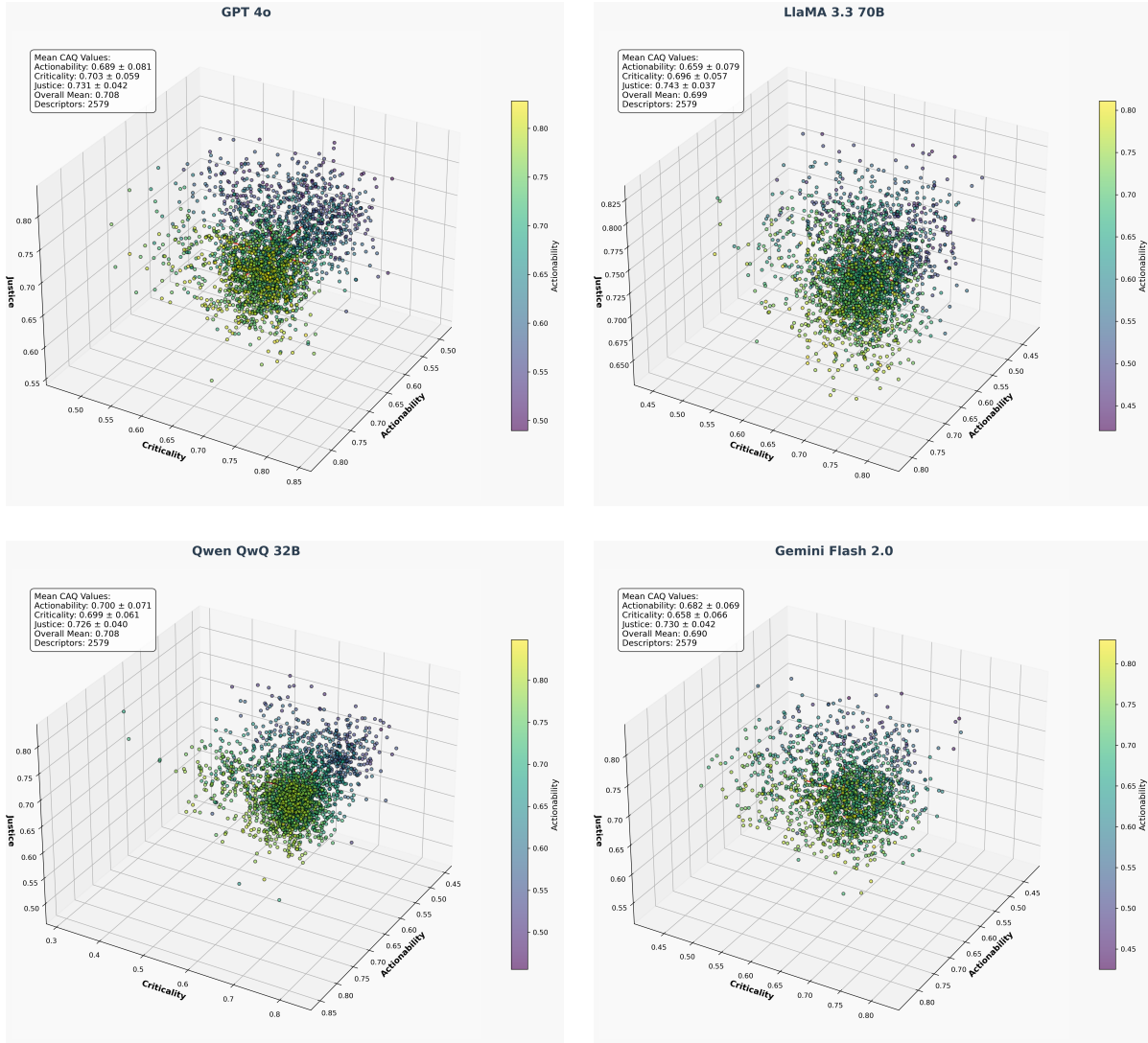


Table 3: 3D scatter plots of CAQ scores for (top-left) GPT-4o, (top-right) LLaMA 3.3 70B, (bottom-left) Qwen QwQ 32B, and (bottom-right) Gemini 2.0 Flash. The x-, y-, and z-axes represent the CAQ dimensions of *Criticality*, *Justice*, and *Actionability*, respectively, while the color scale on the right encodes the *Actionability* CAQ value, with lighter colors indicating higher values. Points near the center in each plot indicate relatively balanced discourse across all three dimensions, whereas outliers suggest potential gaps or imbalances. We observe that GPT-4o and LLaMA 3.3 70B exhibit more points clustered near the center, indicating a higher consistency of balanced messages, while Qwen QwQ 32B and Gemini 2.0 Flash show a wider spread, reflecting greater variability in how they handle critical and justice-oriented aspects of climate communication.

Table 4: Comparison of LLM Responses across Actionability, Criticality, and Justice Dimensions


Image	Image Tweet
	<p style="text-align: center;"> <i>#climate</i>  <i>#ClimateCrisis</i>  <i>#ClimateEmergency</i>  <i>#climatememe</i>  <i>#ClimateBrawl</i>  <i>#babysteps Thanks</i>  <i>@MemesChange</i> </p>
<b>Description</b>	
<p>The image and caption convey a satirical message about the ineffectiveness of government climate policies. The visual depicts a person in a building window, seemingly indifferent to a large fire raging outside, while holding a green hat. The text above the image reads, "Government's Climate Policies be like:" and the text below the image says, "Yep, that should do it." The message suggests that current climate policies are insufficient or ineffective in addressing the severity of climate change. The person in the window symbolizes a lack of urgency or action from governments to combat climate change. The fire outside represents the escalating climate crisis, while the green hat implies a lack of meaningful or effective solutions. The overall tone of the meme is critical and sarcastic, highlighting the need for more ambitious and immediate climate action. It emphasizes the risks of inaction and the importance of implementing more effective and comprehensive policies to address the climate crisis.</p>	
<b>Actionability</b>	
<p><b>Claude 3.7 Sonnet Response:</b> The meme criticizes government climate policies as performative and inadequate relative to the urgency of the climate crisis, highlighting the disconnect between symbolic gestures and the scale of action needed. Actionable solutions include implementing carbon pricing mechanisms (high feasibility), accelerating renewable energy transitions (medium feasibility), strengthening international climate agreements with binding targets (medium feasibility), and investing in climate adaptation infrastructure (medium feasibility). The meme lacks explicit commitments, failing to specify which actors should implement these solutions, on what timeline, or through which specific mechanisms—a significant weakness that mirrors its critique of vague government approaches. Key risks include political resistance to meaningful regulation, economic transition challenges for fossil fuel-dependent communities, and the possibility that incremental approaches may be insufficient given the accelerating pace of climate change, requiring more transformative systemic changes than current policy frameworks contemplate.</p>	
<p><b>Human Analysis:</b> The LLM's response <b>effectively captures the meme's satirical critique</b> of performative climate policy, aligning with its depiction of indifference (the green-hatted figure) amidst crisis (the raging fire). Its <b>suggestion of solutions</b> like carbon pricing and renewable energy aligns with the problem identification aspect of actionability. <b>However, the feasibility ratings are unsubstantiated</b>, diminishing analytical depth. While the LLM <b>rightly notes the meme's lack of concrete solutions, it fails to offer specific 'who, what, when, how' details</b> for its own proposals, mirroring the very vagueness it critiques. Still, its <b>acknowledgment of key risks</b>, such as political resistance and incrementalism, adds valuable pragmatic insight.</p>	

Table 5: Comparison of LLM Responses (Continued) - Criticality and Justice Dimensions

<b>Continued from Table 4</b>
<b>Criticality</b>
<p><b>Claude 3.7 Sonnet Response:</b> The described meme presents a simplistic critique of government climate policies, portraying them as wholly inadequate through the metaphor of a person with a green hat (symbolizing superficial environmentalism) ignoring a raging fire (representing climate change). While it effectively captures frustration about the pace and scale of climate action, the image lacks nuance by implying universal government inaction without acknowledging the significant variation in climate policies across different countries and jurisdictions. The satirical format offers no specific evidence or examples of policy failures, nor does it engage with the complex political, economic, and social challenges that complicate climate policy implementation. The meme also fails to recognize incremental progress or acknowledge competing viewpoints about appropriate climate response strategies. While potentially effective at generating emotional resonance and raising awareness about the urgency of climate action, this type of oversimplified messaging may ultimately contribute to polarization in climate discourse by reducing a multifaceted global challenge to a binary narrative of complete success versus complete failure.</p>
<p><b>Human Analysis:</b> The LLM provides a largely <b>effective criticality analysis</b> of the meme, <b>accurately identifying its core message as a simplistic critique</b> that portrays government policies as inadequate (symbolized by the indifferent green-hatted figure). It <b>rightly critiques the meme’s lack of nuance</b>, noting its failure to consider policy variation, implementation challenges, opposing views, and incremental progress. The LLM’s point that such messaging <b>may fuel polarization in climate discourse</b> is especially insightful. <b>However, to deepen its critical analysis, the response should go beyond surface-level critique</b> to explore the systemic roots of policy failure and the appeal of reductive narratives in climate communication.</p>
<b>Justice</b>
<p><b>Claude 3.7 Sonnet Response:</b> This meme critiques government climate inaction but fails to center marginalized communities who bear disproportionate climate impacts despite contributing least to the crisis. The satire places responsibility primarily on governments without acknowledging how corporate interests and wealthy nations have historically shaped inadequate policies while benefiting from carbon-intensive development. The imagery of a distant observer watching destruction unfolds without recognizing how climate disasters exacerbate existing inequalities for Indigenous peoples, communities of color, and economically disadvantaged populations. By presenting climate change as a universal threat without differentiation, the meme overlooks how historical power imbalances determine who has decision-making authority in climate policy and whose knowledge systems are valued. The satirical framing, while effective at highlighting policy inadequacy, misses the opportunity to incorporate diverse cultural contexts and traditional ecological knowledge that could inform more just and effective climate solutions.</p>
<p><b>Human Analysis:</b> The LLM offers a <b>strong justice-oriented critique</b> of the meme, <b>rightly highlighting its failure to center marginalized communities</b> most affected by climate change despite contributing least to it. It <b>effectively critiques the narrow focus on government inaction</b>, noting the overlooked roles of corporate power and wealthy nations, and <b>astutely observes the absence of attention to how climate disasters exacerbate existing inequalities</b> for groups like Indigenous peoples and communities of color. The LLM also <b>correctly critiques the portrayal of climate change as a universal threat</b>, ignoring historical power imbalances and erasing diverse knowledge systems. <b>However, to enhance its impact, the analysis could go beyond diagnosis to suggest how such satire might better incorporate justice concerns</b>, for example, by referencing specific inequities or inclusive solutions.</p>