Post-hoc Study of Climate Microtargeting on Social Media Ads with LLMs: Thematic Insights and Fairness Evaluation

Tunazzina Islam
Department of Computer Science
Purdue University
West Lafayette, IN 47907
`islam32@purdue.edu`

Dan Goldwasser
Department of Computer Science
Purdue University
West Lafayette, IN 47907
`dgoldwas@purdue.edu`

Abstract

Climate change communication on social media increasingly employs microtargeting strategies to effectively reach and influence specific demographic groups. This study presents a *post-hoc* analysis of microtargeting practices within climate campaigns by leveraging large language models (LLMs) to examine Meta (previously known as Facebook) advertisements. Our analysis focuses on two key aspects: **demographic targeting** and **fairness**. We evaluate the ability of LLMs to accurately predict the intended demographic targets, such as gender and age group. Furthermore, we instruct the LLMs to generate explanations for their classifications, providing transparent reasoning behind each decision. These explanations reveal the specific thematic elements used to engage different demographic segments, highlighting distinct strategies tailored to various audiences. Our findings show that ***young adults*** are primarily targeted through messages emphasizing *activism and environmental consciousness*, while ***women*** are engaged through themes related to *caregiving roles and social advocacy*. Additionally, we conduct a comprehensive fairness analysis to uncover biases in model predictions. We assess disparities in accuracy and error rates across demographic groups using established fairness metrics such as Demographic Parity, Equal Opportunity, and Predictive Equality. Our findings indicate that while LLMs perform well overall, certain biases exist, particularly in the classification of ***male*** audiences. The analysis of thematic explanations uncovers recurring patterns in messaging strategies tailored to various demographic groups, while the fairness analysis underscores the need for more inclusive targeting methods. This study provides a valuable framework for future research aimed at enhancing transparency, accountability, and inclusivity in social media-driven climate campaigns.

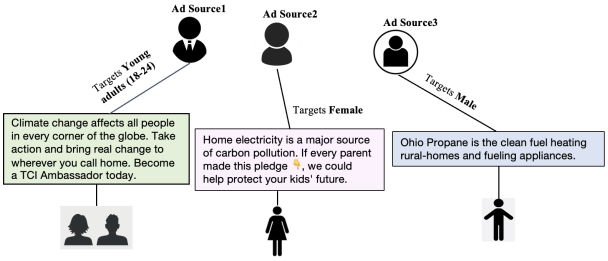
1 Introduction

Climate change represents one of the most pressing global challenges of the 21st century, necessitating widespread public awareness and engagement to drive meaningful environmental action (Moritz and Agudo, 2013; Dessler and Theater, 1995). As traditional media channels evolve, social media has emerged as a pivotal platform for climate communication, enabling organizations, activists, and policymakers to disseminate information, mobilize support, and influence public discourse on environmental issues (Nosek, 2020; Hestres and Hopke, 2017; Adger et al., 2003). The interactive and targeted nature of social media advertising allows for the customization of messages to resonate with specific audiences, thereby enhancing the effectiveness of communication strategies aimed at fostering climate awareness and behavioral change (Bloomfield and Tillery, 2019; Walter et al., 2018; Stoddart et al., 2016).

In recent years, the utilization of microtargeting strategies in social media campaigns has gained significant traction (Eldar and Hidir, 2025; Islam, 2025b; Nistor, 2024). Microtargeting involves the precise tailoring of messages to distinct demographic segments based on factors such as age, gender, location, and interests (Islam, 2025a; Prummer, 2020; Hersh, 2015; Barbu, 2014). This approach leverages vast amounts of user data to craft personalized content that is more likely to engage and persuade individual users. In the context of climate communication, microtargeting can enhance the relevance and impact of messages, potentially leading to greater public engagement and support for environmental initiatives. However, the sophistication of these strategies also raises critical questions about the transparency, ethical implications, and overall effectiveness of targeted climate messaging (Islam and Goldwasser, 2024a; Islam et al., 2023b).

Figure 1 illustrates the targeted climate adver-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15838–15859
November 4-9, 2025 ©2025 Association for Computational Linguistics



Climate change affects all people in every corner of the globe. Take action and bring real change to wherever you call home. Become a TCI Ambassador today.

Home electricity is a major source of carbon pollution. If every parent made this pledge 👇, we could help protect your kids' future.

Ohio Propane is the clean fuel heating rural-homes and fueling appliances.

Figure 1: Example of climate microtargeting.

tisements on social media, with a specific focus on demographic targeting. The first ad source targets **young adults aged 18-24**, with a message *encouraging action against climate change and inviting them to become The Climate Initiative (TCI)*[1] *) Ambassadors*. The second ad source is tailored to a **female** audience with a message emphasizing the *importance of reducing carbon pollution from home electricity and making a pledge for their children's future*. The third ad source targets a **male** audience, focusing on the *benefits of using clean fuel like Ohio Propane for heating rural homes and fueling appliances*.

Despite the growing prevalence of microtargeting in climate campaigns, there remains a limited understanding of the specific techniques and linguistic patterns employed to engage different demographic groups. Traditional method (Roberts et al., 2019; Vaismoradi et al., 2016, 2013; Braun and Clarke, 2012, 2006; Tuckett, 2005) often falls short in capturing the nuanced and context-dependent nature of targeted communication. This gap highlights the need for advanced analytical tools that can dissect and interpret the complex language and strategies used in microtargeted climate advertisements. Large language models (LLMs) (Brown et al., 2020), with their robust natural language processing (NLP) capabilities, offer a promising solution to this challenge. In this paper, we investigate whether the newly emerging paradigm in NLP- zero-shot prompting of LLMs (Brown et al., 2020) and the practice of providing explanations of answers are better equipped to address those challenges.

Explanations are fundamental to human learning (Ahn et al., 1992), as they underscore task principles that facilitate broad generalizations (Lombrozo and Carey, 2006; Lombrozo, 2006). Consider the example text (*“The turbines that provide clean wind energy also create new habitats for fish, keeping the ecosystem healthy."*) from Figure 2(a) for the gender identification task. An explanation can elaborate on a brief answer (e.g., `male`) by connecting it to the broader reasoning process necessary to solve the problem (e.g., “`Emphasizing fish habitats may appeal more to men interested in fishing and environmental conservation.`”). Thus, explanations enhance understanding by demonstrating how task principles connect questions to their answers.

While LLMs offer powerful capabilities for analyzing and generating text, their widespread use has also highlighted significant challenges related to fairness (Kumar et al., 2023; Li et al., 2023b,a; Sag, 2023) and bias (Lin et al., 2024; Kotek et al., 2023; Fang et al., 2024; Urman and Makhortykh, 2023; Esiobu et al., 2023). Research has shown that LLMs, like other machine learning (ML) models, can inherit and even amplify biases present in the data on which they are trained (Blodgett et al., 2020; Binns, 2018). These biases can manifest in various forms, such as differential accuracy across demographic groups (Dacon and Liu, 2021; Lu et al., 2020; Hitti et al., 2019; Díaz et al., 2018), harmful stereotypes (Bianchi et al., 2023; Matthews et al., 2022; Alemany et al., 2022), and discriminatory language patterns (Salguero and Espinilla, 2018). The implications are particularly concerning in high-stakes applications such as healthcare, finance, and social networks, where biased outcomes can perpetuate inequities and undermine trust in artificial intelligence (AI) systems (Mehrabi et al., 2021; Buolamwini and Gebru, 2018). Addressing these issues requires fairness evaluations using metrics like Demographic Parity, Equal Opportunity, and Predictive Equality to ensure that models perform equitably across all user groups (Žliobaitė, 2017; Hardt et al., 2016). In the context of this study, examining the fairness of LLMs' predictions in demographic targeting is crucial for understanding the microtargeted climate messaging.

In this study, we conduct a *post-hoc* analysis of climate microtargeting practices on social media by leveraging the power of LLMs (**OpenAI's o1-preview model**[2]). Post-hoc analysis, typically performed after the main experiment or event, allows us to retrospectively evaluate how effective these campaigns are in targeting specific demographics. Building upon data from previous research by Islam et al. (2023b); Islam and Goldwasser (2024a), we investigate the ability of LLMs to accurately

---

[1] https://theclimateinitiative.org/

[2] https://openai.com/index/learning-to-reason-with-llms/

Ad: **The turbines that provide clean wind energy also create new habitats for fish, keeping the ecosystem healthy.**

Report who can be targeted based on gender: male vs female by the provided ad. Choose the most probable option from the gender (either male or female). Don't generate neutral or both gender answers. Provide a brief explanation with the reasoning in 20 words.
Gender: ?
Explanation: ?

Gender: Male
Explanation: Emphasizing fish habitats may appeal more to men interested in fishing and environmental conservation.

(a) Prompt example for gender prediction (male).

Ad: **U.S. school facilities emit about 72 MILLION metric tons of carbon pollution every year. If every parent made this pledge 👇, we could help protect your kids' future.**

Report who can be targeted based on gender: male vs female by the provided ad. Choose the most probable option from the gender (either male or female). Don't generate neutral or both gender answers. Provide a brief explanation with the reasoning in 20 words.
Gender: ?
Explanation: ?

Gender: Female
Explanation: The ad appeals to mothers, focusing on children's future and environmental concerns, which often resonate more with women.

(b) Prompt example for gender prediction (female).

Figure 2: Prompt examples for gender prediction. (a) male, (b) female. Inputs are shown in blue, outputs in red.

detect targeted messaging based on specific demographic variables, including gender and age group. Additionally, LLMs provide explanations for their classification decisions, offering insights into the thematic and linguistic elements used to engage different audiences. Furthermore, we conduct a comprehensive fairness analysis to identify potential biases in model predictions. Our exploration leads to the following research questions (RQ), which are crucial for assessing the potential of LLMs to understand microtargeting patterns and provide deeper and more nuanced insights:

- **RQ1:** Given a text, can LLMs identify the targeted demographic of the corresponding text accurately and provide an explanation for the reasoning behind the prediction?
- **RQ2:** What are the recurring themes and aspects of **explanations** provided by LLMs?
- **RQ3:** How fair are the LLMs' predictions in terms of demographic targeting, and what are the disparities in prediction accuracy and error rates across different demographic groups?

The main idea of this paper is to utilize LLMs as a tool to analyze real-world targeting practices, measuring biases and fairness in actual campaigns. We explore how LLMs can be used post-hoc as third-party tools to analyze patterns in targeted communication, especially when **internal platform logic is not transparent**. While platforms have white-box access, external stakeholders (researchers, auditors, policymakers) **do not**. Our method offers an *explainable* approach to *reverse-engineer* targeting practices and uncover potential bias or messaging disparities.

The implications of this research are multifaceted. From an academic perspective, it contributes to the burgeoning field of computational social science (CSS) by showcasing the application of advanced language models in dissecting complex communication strategies. Practically, the findings offer valuable insights for policymakers, environmental organizations, and social media platforms seeking to enhance the transparency, accountability, and inclusivity of their climate communication efforts. By illuminating the specific methods used to tailor messages to different demographics and by highlighting the need for fairer and more inclusive targeting methods, this study lays the groundwork for future investigations into the role of AI in enhancing the efficacy and ethical standards of digital climate advocacy.

2 Related Work

The intersection of microtargeting, social media, and climate communication has drawn considerable interest in recent years, with studies examining how digital platforms influence public opinion and engagement (Islam and Goldwasser, 2024b,a; Holder et al., 2023; Islam et al., 2023b). The growing capabilities of social media platforms to deliver personalized messages based on user data have sparked a significant body of research into computational advertising (Abrams et al., 2022; Huh and Malthouse, 2020; Yang et al., 2017; Zhang et al., 2017), demographic targeting (Allen, 2022; Rummo et al., 2020; Ribeiro et al., 2019; Kaspar et al., 2019; Jansen et al., 2013; Jansen and Solomon, 2010), and the broader implications of these practices for societal discourse.

Microtargeting on social media platforms such as Meta has been widely studied in the context of political campaigns (Islam et al., 2023a; Capozzi et al., 2021; Serrano et al., 2020; Silva et al., 2020; Capozzi et al., 2020) and public health messaging (Islam and Goldwasser, 2022; Silva and Benevenuto, 2021; Mejova and Kalimeri, 2020), revealing both the potential benefits and ethical concerns associated with this practice. In the realm of climate communication, microtargeting can be a powerful tool for enhancing message relevance and impact by tailoring content to the specific values and interests of diverse demographic groups (Islam et al., 2023b). However, the effectiveness and ethical implications of such targeted messaging remain underexplored, particularly in terms of how different demographic groups engage and whether biases are present in targeting strategies.

Recent advances in LLMs have demonstrated their capability for in-context learning (ICL), significantly enhancing their ability to perform tasks traditionally handled by humans (Chowdhery et al., 2023; Kojima et al., 2022; Le Scao et al., 2022; Brown et al., 2020; Gilardi et al., 2023; De Paoli, 2023; Dai et al., 2023; Chiang and Lee, 2023; Ziems et al., 2024). This progress suggests a strong potential for effectively applying LLMs to our specific task. In the realms of qualitative research (QR) and NLP, innovative methods are being explored to integrate LLMs into Thematic Analysis (TA) (Braun and Clarke, 2012). Researchers have proposed various frameworks, including an LLMs-in-the-loop approach (Islam and Goldwasser, 2024b,a; Dai et al., 2023), integrating GPT-3 with expert-designed codebooks (Xiao et al., 2023), and developing collaborative interfaces that utilize LLMs for code generation and support in decision-making processes (Gao et al., 2023). Other recent work has shown that LLMs can benefit from examples that decompose the reasoning process (can be seen as an explanation), leading to an answer (Wei et al., 2022). Despite the impressive capabilities of LLMs, there are concerns about fairness, accountability, and transparency in LLMs' predictions, which have been highlighted in recent literature (Anthis et al., 2024; Dai et al., 2024; Kumar et al., 2023; Li et al., 2023b; Wu et al., 2023; Bender et al., 2021), emphasizing the need for a rigorous evaluation of biases and disparities in model performance across different groups. In this paper, we leverage LLMs to identify targeted demographics and provide explanations of demographic targeting regarding climate-related advertisements on Meta. Besides, we develop a new set of themes and aspects based on those explanations, specifically tailored for analyzing messaging. Additionally, we extend previous research by not only focusing on the accuracy of these predictions but also conducting a comprehensive fairness analysis to identify and address potential biases in the model's performance.

3 Dataset

We investigate the climate campaigns case study for this work. We work on the corpus of 21372 ads (Islam and Goldwasser, 2024a) originally released by Islam et al. (2023b). This dataset includes climate-related English ads on Meta from the US, spanning from January 2021 to January 2022. Each ad includes the following attributes: ad ID, title, ad description, ad body, funding entity, spend, impressions, and distribution of impressions broken down by gender (male, female, unknown), age (seven groups), and location down to the state level in the USA. Additional details about the dataset can be found in the original publications.

For this work, we consider two demographic indicators, i.e., **gender** and **age group**. We consider two gender categories, i.e., *male* vs *female*. Regarding age group, we consider four age group categories, i.e., *young adults* whose age range is 18-24, *early working* age group (25-44), *late working* age group (45-64), and *senior citizens* (65+). However, the Meta Ad Library API doesn't provide explicit targeting information beyond ad impressions (views). Note that this paper **infers targeting from exclusive impression distributions, which may differ from advertiser intent**. In this work, we focus only on ads viewed exclusively by either males or females; ads viewed by both genders were excluded. Therefore, if an ad is viewed solely by males or solely by females, we consider it a targeted ad. The same logic is applied to age groups. We do not include any ads that overlap across age categories, such as young adults, early working-age individuals, late working-age individuals, and senior citizens. By following this methodology, we narrow down the dataset to 227 ads. Among them, 59 ads target only females and 47 ads target only males. However, we find 25 ads target only young adults, 82 ads target only the early working age group, 8 ads target only the late working age group, and 6 ads target only senior citizens.

4 Experimental Setup

The identification of the targeted demographic (with explanation) in a text involves the following steps:

Gender prediction with Explanation: For a given text t, the task involves identifying the targeted gender and explaining the rationale behind its selection.

Age group prediction with Explanation: Subsequently, the task requires predicting the targeted age group and providing an explanation for the specific choice.

To identify the targeted demographics, we use the most recently[3] released by OpenAI, **o1-preview model**[4]. This is a new, large language model that uses reinforcement learning (RL) and chain of thought (COT) techniques for complex reasoning, allowing it to think through a detailed internal process before responding to users. We use the OpenAI Playground API to run **o1-preview** by keeping the default parameters. The prompt template for the demographic prediction task with an explanation using LLMs can be found in Fig. 3 in App. A.1. Fig. 2 shows the example prompts for gender prediction. Fig. 4 in App. A.1 shows the example prompts for age group prediction from the climate campaign dataset.

4.1 Results

Table 1 provides the overall accuracy of the targeted demographic prediction task by LLMs as well as a detailed breakdown of correct and incorrect predictions across each demographic category. LLMs can predict the targeted demographics with an accuracy of 88.55% answering **RQ1**. LLMs achieve high accuracy in predicting both females (94.92%) and males (85.10%). A small number of females are misclassified as males, and a few males are misclassified as females. Table 1 shows high accuracy for young (88.00%) and early working (91.46%) categories. Performance drops for the late working (75%) age group and significantly for the senior (33.33%) categories.

For baseline comparison, we use open-sourced LLMs Llama 3 (llama3-70b-8192[5]) (Touvron et al., 2023) and Mistral Large 2 (mistral-large-2407[6]) (Jiang et al., 2023). For llama3-70b-8192 and mistral-large-2407, we keep their default parameters. To run llama3-70b-8192, we use Groq API[7]. For the tokenizer, we use Meta-Llama-3-70B-Instruct from Hugging Face. For running Mistral Large 2, we use the Mistral AI API. Besides, we include a smaller pre-trained language model BERT (Devlin et al., 2019) (bert-base-uncased) for comparison. Furthermore, we compare with simple Logistic Regression (LR) (Cox, 1958) trained on term frequency–inverse document frequency (tf-idf) features baseline. However, OpenAI's o1-preview model outperforms the baselines both in gender (macro average F1 score 90.35%) and age group (macro average F1 score 71%) predictions (Table 2).

4.2 Statistical Tests

We perform a binomial test assessing whether the model's gender predictions are significantly better than random guessing (50%).

Null hypothesis, H_o : probability of correct prediction is 0.5 (random guessing), p-value: $\sim 4.37 \times 10^{-19}$, Conclusion: Reject H_o. This means the model's accuracy is significantly better than chance, with a very strong statistical significance. Table 4 shows the statistical test results from the gender subgroup.

Here are the results of the binomial test for age group predictions, assuming random guessing across 4 categories ($p = 0.25$):

Null hypothesis, H_o: probability of correct prediction is 0.25 (random guessing), p-value: $\sim 4.27 \times 10^{-45}$, Conclusion: Reject H_o. The model's performance is significantly better than random chance. This confirms that the predictions are statistically meaningful. Table 5 shows the statistical

---

[3] September 12, 2024

[4] https://openai.com/index/introducing-openai-o1-preview/

[5] `https://github.com/meta-llama/llama3`

[6] `https://mistral.ai/news/mistral-large-2407/`

[7] `https://wow.groq.com/`

Category	**Total Ads**	**Correct Pred.**	**Acc. (%)**	**Misclass.**
All	227	201	88.55	-
Female	59	56	94.92	3 (Male)
Male	47	40	85.10	7 (Female)
Young adults	25	22	88.00	2 (Early Working), 1 (Late Working)
Early Working	82	75	91.46	4 (Young), 4 (Late Working)
Late Working	8	6	75.00	2 (Early Working)
Senior	6	2	33.33	3 (Young), 1 (Late Working)

Table 1: Accuracy and Misclassifications for Demographics.

Model	**Demo.**	**Acc. (%)**	**F1 (%)**
$LR_{\text{tf-idf}}$	gender	69.00	65.00
	age	73.00	31.00
BERT	gender	72.00	70.00
	age	70.00	26.00
Llama 3	gender	80.19	79.67
	age	58.68	36.84
Mistral Large 2	gender	82.08	82.07
	age	74.38	48.68
o1-preview	**gender**	**90.57**	**90.35**
	age	**85.95**	**71.00**

Table 2: Baseline comparisons. Demo.: Demographics, Acc.: Accuracy.

test results from the age subgroup.

4.3 Error Analysis

Table 10 in App. A.2 presents an analysis of ad misclassifications based on gender and age group predictions. Each entry includes the actual demographic, the predicted demographic, and a brief explanation generated by LLMs. Explanations highlight how specific patterns and themes within an ad can lead to demographic misclassifications. In some cases, **traditional gender roles** and **age-related interests** played a significant role in these misclassifications. Understanding these nuances can help in refining predictive models and improving the accuracy of demographic targeting in future ad campaigns.

5 Thematic Insights of Explanations

As LLMs provide explanations to provide the reasoning behind their prediction, we use those explanations to understand thematic insights to answer **RQ2**. In this analysis, we **only include the correct predictions and their explanations**. Using only correct predictions is **not arbitrary**—it is supported by **methodological norms** (Vaismoradi et al., 2016, 2013; Braun and Clarke, 2012, 2006) to ensure pattern coherence.

5.1 Themes and Aspects of Gender Explanations

We prompt LLMs to provide the common theme and aspects under the specific theme of the explanations from 40 correct **male** predictions and 56 correct **female** predictions. The prompt template is shown in Figure 5 App. A.1. We detail the theme of gender explanation and the aspects of the explanation in Table 3. Two experts in NLP and CSS meticulously evaluate the generated themes and aspects derived from the explanations, ensuring their accuracy and alignment with the intended context. The inter-annotator agreement is 0.94 (almost perfect agreement) using Cohen's Kappa coefficient (Cohen, 1960). We detail the theme of the gender explanations and aspects of that explanation in App. A.3.

5.2 Themes and Aspects of Age Explanations

We prompt LLMs to provide the common theme and aspects under the specific theme of the explanations from 22 correct **young adult** predictions, 75 correct **early working age group** predictions, 6 correct **late working age group** predictions, and 2 correct **senior citizen** predictions. We detail the theme of the age group explanation and aspects in Table 6. Similarly to gender, the same two NLP and CSS experts review the generated themes and aspects from the explanations to ensure accuracy and contextual alignment. For the age group, the inter-annotator agreement is 0.91 (almost perfect agreement) using Cohen's Kappa coefficient. We detail the theme of the age group explanations and aspects of that explanation in App. A.4.

5.3 Baselines

We add Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and BERTopic (Grootendorst, 2022)

Gender	Theme of Explanation	Aspects of Explanation
Male	Perceived Interests and Roles	Interest in Technology and Innovation, Focus on Economic and Financial Issues, Property and Land Management, Traditional Male Activities, Engagement in Political and Infrastructure Topics, Conservative Views and Skepticism
Female	Roles as Caregivers, Environmental Advocates, and Socially Conscious Individuals	Parental and Caregiving Roles, Environmental Consciousness, Social Welfare and Community Involvement, Empathy and Emotional Appeal, Female Empowerment and Leadership, Health and Safety Concerns

Table 3: Gender based Themes and Aspects of Explanations.

Gender	p-value	Conclusion
Male	5.35×10^{-7}	Reject H_0
Female	5.95×10^{-14}	Reject H_0

Table 4: Significance tests by subgroup (gender).

Age Group	p-value	Conclusion
Senior	4.66×10^{-1}	Fail Reject H_0
EarlyWorking	1.04×10^{-35}	Reject H_0
Young	5.76×10^{-11}	Reject H_0
Late Working	4.23×10^{-3}	Reject H_0

Table 5: Significance tests by subgroup (age).

as unsupervised techniques for generating topics from LLMs-generated explanations. BERTopic[8] is a hierarchical topic modeling and an off-the-shelf neural baseline for clustering texts. BERTopic generates two topics from the prediction explanation of gender and age (Table 13 in App. A.5). We run LDA with four topics and show the top 10 words in App. A.5 in Table 13. We perform a human evaluation by asking annotators to assess whether the generated themes and topics match the ad's explanation. Table 12 in App. A.5 shows the human evaluation results on LLM (GPT-4o (OpenAI, 2024)) assignment, LDA, and BERTopic.

6 Fairness and Bias Analysis

In this section, we present a comprehensive fairness analysis of the model for gender and age group classifications to answer **RQ3**. We evaluate the models using established fairness metrics such as Demographic Parity, Equal Opportunity, and Predictive Equality to assess their performance across different groups. By analyzing confusion matrices and classification reports, we identify any disparities in prediction accuracy and error rates between groups. Our analysis aims to identify biases and investigate the underlying reasons for any observed biases. The insights gained from this analysis are critical for guiding future research in developing fair and inclusive algorithms.

6.1 Fairness Analysis on Gender Prediction

To assess the fairness of the model on gender prediction, we compute several fairness metrics, including Demographic Parity, Equal Opportunity, and Predictive Equality.

Demographic parity examines whether each gender group receives positive predictions at equal rates. Table 7 presents the Demographic Parity ratios for each gender. A ratio of 1 indicates perfect parity. The results show that **females have a slightly higher likelihood of receiving positive predictions compared to males**, suggesting a minor imbalance favoring the female class.

Equal Opportunity (True Positive Rate) focuses on the True Positive Rates (TPR) across gender groups, measuring the model's ability to correctly identify positive instances within each group. We have **Female TPR: 0.95** and **Male TPR: 0.85**. The TPR for females is higher by 0.10, indicating that the model may be more effective at correctly identifying females than males.

Predictive Equality (False Positive Rate) assesses the False Positive Rates (FPR) across gender groups, reflecting the rate at which negative instances are incorrectly labeled as positive. We

---

[8] `https://maartengr.github.io/BERTopic/#common`

Age group	Theme of Explanation	Aspects of Explanation
Young adults (18-24)	Activism and Environmental Consciousness	Passion for Climate Action, Support for Bold Environmental Leadership, Engagement with Activism, Desire for Immediate Change, Participation in Training and Advocacy
Early working (25-44)	Proactive and Responsible Mindset	Environmental Consciousness, Financial Stability and Disposable Income, Parental and Future Concerns, Career Engagement and Professional Roles, Interest in Innovation and Technology, Social and Political Engagement
Late working (45-64)	Responsibilities and Concerns	Economic and Environmental Responsibility, Homeownership and Financial Stability, Voter and Policy Engagement, Economic Concerns
Senior (65+)	Health and Safety Concerns	Health and Wellness Programs, Vulnerability and Safety

Table 6: Age group based Themes and Aspects of Explanations.

achieve **Female FPR: 0.07** and **Male FPR: 0.05**. The slightly higher FPR for females suggests that **females are more likely to be incorrectly predicted as positive compared to males**.

Gender	Demographic Parity Ratio
Female	1.0678
Male	0.9149

Table 7: Demographic Parity for Predicted Gender.

Age Group	Demographic Parity Ratio
Early Working	0.95
Late Working	1.50
Senior	0.33
Young	1.16

Table 8: Demographic Parity for Predicted Age Group.

Age Group	Equal Opportunity (TPR)	Predictive Equality (FPR)
Early Working	0.90	0.10
Late Working	0.75	0.05
Senior	0.33	0.00
Young	0.88	0.07

Table 9: TPR and FPR for Age Groups.

6.2 Fairness Analysis on Age Prediction

Results in Table 8 show that the **late working group has a higher likelihood of receiving positive predictions**, while the **senior group has a significantly lower rate,** indicating **potential bias against seniors**. The senior group has a Demographic Parity ratio of 0.33, significantly lower than the ideal value of 1, suggesting **under representation in positive predictions**.

Table 9 shows that the TPR for the senior group is notably lower, suggesting the model is less effective at correctly identifying individuals in this age group. The senior group has an FPR of 0, indicating no false positives, while the early working group has the highest FPR among the groups.

To understand the observed bias in the misclassified senior age group, we conduct an analysis of the misclassified instances (Table 11 in App. A.2.1). The goal is to identify patterns and underlying reasons why the model predicts seniors as belonging to the young and late working age group. The misclassification of the Senior age group as Young or Late Working can be attributed to several factors, such as: lack of age-specific cues, reliance on stereotypical associations, feature representation limitations etc. We detail in App. A.2.1.

6.3 Statistical Significance Tests

While 0.10 difference in TPR may seem modest, prior fairness literature (Hardt et al., 2016) recognizes it as meaningful in sensitive domains. We conduct two-proportion Z-tests to assess whether the observed differences in TPRs across demographic groups are statistically significant. For gender prediction, the difference in TPR between females (0.95) and males (0.85) resulted in a p-value of 0.086 with a 95% confidence interval of $[-0.018, 0.230]$. This result indicates that the gap is not statistically significant at the conventional 0.05 level. However, in the age group analysis, the TPR difference between young adults (0.88) and seniors (0.33) yields a p-value of 0.004 with a 95% confidence interval of $[0.138, 0.796]$, indicating a statistically significant disparity. This validates our concern about fairness in age-related predictions, particularly for underrepresented senior groups.

6.4 Sensitivity Analysis

Our study is descriptive and **does not claim causal effects**. We have added sensitivity analysis (Fig. 7 in App. A.6) by rerunning the fairness test after stratifying by *ad-spend* and by *topic cluster* to confirm the reported biases persist under two plausible confounders.

Fig 7a in App. A.6 shows that bias persists after conditioning on ad spend. Low-budget ads: $\Delta \approx 0.82$ (95% CI 0.68 – 0.96), High-budget ads: $\Delta \approx 0.76$ (95% CI 0.55 – 0.93). Overlapping CIs indicate similar magnitude, and **ad-budget is not driving the disparity**. Topic clusters tell the same story (Fig. 7b in App. A.6). Four of five clusters show sizable positive gaps (0.30 – 1.00). The one cluster with only female-authored ads (topic 2) cannot yield a gap, but its omission does not overturn the pattern. Accuracy gaps are modest and hover around zero. Where CIs cross zero, classification accuracy is comparable between genders; the **selection-rate bias remains the dominant concern**.

Fig 7c in App. A.6 illustrates that the gap is not a low-budget artifact: whether an ad costs \$5 or \$500, the model is still far more likely to assign it to some age groups than others. Four of five topics show large positive parity gaps—even where sample sizes differ. In Fig. 7d in App. A.6, Topic 1 does not contradict the bias claim as it simply contains ads that all target (and therefore are predicted for) one single age group. Stratified sensitivity checks confirm that age-group disparities persist across both advertising-budget tiers ($\Delta \approx 0.47$ – 1.0) and four of five latent topic clusters, indicating that **neither spend nor topic accounts for the bias**.

6.5 Robustness Analysis

To provide a more in-depth evaluation of the quality and robustness of the data, we plot a learning curve containing macro-F1 and demographic-parity against sample size (20–227 ads). Learning-curve plot in Fig. 8 in App. A.7 shows that macro-F1 and fairness gaps plateau well before 227 examples—evidence that the **analysis is not variance-limited**.

Both curves (Fig. 8a and Fig. 8b App. A.7) plateau by ≈ 60 examples, confirming that our results are robust to subsampling and that additional data in the current distribution yield diminishing returns. Both performance and bias saturate after $\sim 1/4$ of the data, indicating the analysis is not variance-limited.

Figures 8c and 8d in App. A.7 show the macro-F1 and age-parity gap at 20-ad increments. Macro-F1 rises to 0.70 by $n \approx 80$ and then plateaus; the 95% confidence band narrows to 0.8 pp. The parity gap likewise converges to ≈ 0.57 with error < 0.02. Thus, our age-bias findings are not driven by sample variance—adding more data in the current distribution neither boosts accuracy nor mitigates disparity.

7 Conclusion

This study shows that LLMs can function as practical, third-party auditors of microtargeted climate ads—identifying *whom messages are meant to reach*, *explaining why*, and surfacing *equity concerns* that are otherwise opaque. On our climate-ad corpus, the LLMs accurately infer demographic targets overall (88.55% accuracy), with especially strong performance for gender (94.92% for female-targeted ads; 85.10% for male-targeted ads) and more variable results across age groups. Beyond prediction, model-generated explanations enable a compact, reusable taxonomy of themes used to engage different audiences, providing interpretable signals that researchers and practitioners can inspect and act on. Besides, the fairness analysis conducted in our study underscores the importance of evaluating and addressing the biases. Disparities in prediction accuracy and error rates highlight the need for more inclusive and equitable targeting methods. Although we show our analysis on the climate campaigns dataset, our approach can easily be adapted to any dataset. It is designed to be scalable without any modifications. Code and dataset are available[9].

---

[9] `https://github.com/tunazislam/llms-posthoc-climate`

8 Limitations

Though we show the climate microtargeting case study, it can be applicable to political or health-related campaigns. However, the main idea of this work is to utilize LLMs as a tool to analyze real-world targeting practices, measuring biases and fairness in actual campaigns.

Our analysis relies on the OpenAI o1-preview model. We chose o1-preview instead of the open-source counterparts due to computational resource constraints. We only used pre-trained LLMs and did not consider fine-tuning due to the resource constraints.

The original dataset does not specify intersectional targeting, such as young females only. Our paper acknowledges that our approach infers targeting from exclusive impression data (view), since the Meta Ad Library API does not disclose actual targeting parameters. Meta's own transparency studies treat ads whose impressions are $\geq 95\%$ in one group as "effectively exclusive". Such ads deliver the same real-world exposure as explicit targeting.

We have a small sample size due to restrictive filtering criteria, which might result in underrepresented groups, particularly late working and senior demographics. The rationale for using this specific dataset is that it is publicly available and, to the best of our knowledge, it is the only publicly available dataset focused on targeted climate campaigns.

9 Ethics Statement

To the best of our knowledge, we did not violate any ethical code while conducting the research work described in this paper. We report the technical details for the reproducibility of the results. In this paper, we did not introduce any new dataset; instead, we experimented using an existing dataset that is adequately cited. The data do not contain personally identifiable information and report engagement patterns at an aggregate level. The author's personal views are not represented in any qualitative results we report, as it is solely an outcome derived from machine learning or AI models.

10 Acknowledgments

We are grateful to the anonymous reviewers for their insightful comments. This work was partially funded by the Purdue Graduate School Summer Research Grant (to TI) and NSF CAREER award IIS-2048001.

References

Adam Abrams, Melanie Beckerleg, and Maryia Shpak. 2022. Targeting advertisements and inferring demographics in the hospitality industry.

W Neil Adger, Saleemul Huq, Katrina Brown, Declan Conway, and Mike Hulme. 2003. Adaptation to climate change in the developing world. *Progress in development studies*, 3(3):179–195.

Woo-kyoung Ahn, William F Brewer, and Raymond J Mooney. 1992. Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2):391.

Laura Alonso Alemany, Luciana Benotti, Hernán Maina, Lucía González, Mariela Rajngewerc, Lautaro Martínez, Jorge Sánchez, Mauro Schilman, Guido Ivetta, Alexia Halvorsen, and 1 others. 2022. A methodology to characterize bias and harmful stereotypes in natural language processing in latin america. *arXiv preprint arXiv:2207.06591*.

Gabrielle Marie Allen. 2022. Targeted: How relevant parties position the ethics of online demographic-based targeted advertising.

Jacy Anthis, Kristian Lum, Michael Ekstrand, Avi Feller, Alexander D'Amour, and Chenhao Tan. 2024. The impossibility of fair llms. *arXiv preprint arXiv:2406.03198*.

Oana Barbu. 2014. Advertising, microtargeting and social media. *Procedia-Social and Behavioral Sciences*, 163:44–49.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504.

Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*, pages 149–159. PMLR.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Emma Frances Bloomfield and Denise Tillery. 2019. The circulation of climate change denial online: Rhetorical and networking strategies on facebook. *Environmental Communication*, 13(1):23–34.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Arthur Capozzi, Gianmarco De Francisci Morales, Yelena Mejova, Corrado Monti, André Panisson, and Daniela Paolotti. 2021. Clandestino or rifugiato? anti-immigration facebook ad targeting in italy. In *CHI*.

Arthur Capozzi, Gianmarco De Francisci Morales, Yelena Mejova, Corrado Monti, André Panisson, and Daniela Paolotti. 2020. Facebook ads: Politics of migration in italy. In *ICSI*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Jamell Dacon and Haochen Liu. 2021. Does gender matter in the news? detecting and examining gender bias in news articles. In *Companion Proceedings of the Web Conference 2021*, pages 385–392.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. Llm-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:2310.15100*.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6437–6447.

Stefano De Paoli. 2023. Can large language models emulate an inductive thematic analysis of semi-structured interviews? an exploration and provocation on the limits of the approach and the model. *arXiv preprint arXiv:2305.13014*.

Andrew Dessler and Lowman Student Center Theater. 1995. The science of climate change.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14.

Michael Eldar and Sinem Hidir. 2025. Political influence through microtargeting.

David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. 2023. Robbie: Robust bias evaluation of large generative language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3764–3814.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):5224.

Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhan, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2023. Collabcoder: A gpt-powered workflow for collaborative qualitative analysis. *arXiv preprint arXiv:2304.07366*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Eitan D Hersh. 2015. *Hacking the electorate: How campaigns perceive voters*. Cambridge University Press.

Luis E Hestres and Jill E Hopke. 2017. Internet-enabled activism and climate change. In *Oxford Research Encyclopedia of Climate Science*.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17.

Faye Holder, Sanober Mirza, Jake Carbone, and Ruth E McKie. 2023. Climate obstruction and facebook advertising: how a sample of climate obstruction organizations use social media to disseminate discourses of delay. *Climatic Change*, 176(2):16.

Jisu Huh and Edward C Malthouse. 2020. Advancing computational advertising: Conceptualization of the field and future directions. *Journal of Advertising*, 49(4):367–376.

Tunazzina Islam. 2025a. *UNDERSTANDING AND ANALYZING MICROTARGETING PATTERN ON SOCIAL MEDIA*. Ph.D. thesis, Purdue University Graduate School.

Tunazzina Islam. 2025b. Understanding microtargeting pattern on social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29269–29270.

Tunazzina Islam and Dan Goldwasser. 2022. Understanding covid-19 vaccine campaign on facebook using minimal supervision. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 585–595. IEEE.

Tunazzina Islam and Dan Goldwasser. 2024a. Discovering latent themes in social media messaging: A machine-in-the-loop approach integrating llms. *arXiv preprint arXiv:2403.10707*.

Tunazzina Islam and Dan Goldwasser. 2024b. Uncovering latent arguments in social media messaging by employing llms-in-the-loop strategy. *arXiv preprint arXiv:2404.10259*.

Tunazzina Islam, Shamik Roy, and Dan Goldwasser. 2023a. Weakly supervised learning for analyzing political campaigns on facebook. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 411–422.

Tunazzina Islam, Ruqi Zhang, and Dan Goldwasser. 2023b. Analysis of climate campaigns on social media using bayesian model averaging. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 15–25, New York, NY, USA. Association for Computing Machinery.

Bernard J Jansen, Kathleen Moore, and Stephen Carman. 2013. Evaluating the performance of demographic targeting using gender in sponsored search. *Information Processing & Management*, 49(1):286–302.

Bernard J Jansen and Lauren Solomon. 2010. Gender demographic targeting in sponsored search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 831–840.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kai Kaspar, Sarah Lucia Weber, and Anne-Kathrin Wilbers. 2019. Personally relevant online advertisements: Effects of demographic targeting on visual attention and brand evaluation. *PloS one*, 14(2):e0212419.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language generation models can cause harm: So what can we do about it? an actionable survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3299–3321.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023a. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Yunqi Li, Lanjing Zhang, and Yongfeng Zhang. 2023b. Fairness of chatgpt. *arXiv preprint arXiv:2305.18569*.

Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2024. Investigating bias in llm-based bias detection: Disparities between llms and human perception. *arXiv preprint arXiv:2403.14896*.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470.

Tania Lombrozo and Susan Carey. 2006. Functional explanation and the function of explanation. *Cognition*, 99(2):167–204.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday*, pages 189–202.

Sean Matthews, John Hudzina, and Dawn Sepehr. 2022. Gender and racial stereotype detection in legal opinion word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12026–12033.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Yelena Mejova and Kyriaki Kalimeri. 2020. Covid-19 on facebook ads: competing agendas around a public health crisis. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 22–31.

Craig Moritz and Rosa Agudo. 2013. The future of species under climate change: resilience or decline? *Science*, 341(6145):504–508.

Daniel Ionel Andrei Nistor. 2024. Thinking through targeting: Social media an effective tool for influencing people and society. In *A Conference Hosted By*.

Grace Nosek. 2020. The fossil fuel industry's push to target climate protesters in the us. *Pace Envtl. L. Rev.*, 38:53.

OpenAI. 2024. Hello gpt-4o. *https://openai.com/index/hello-gpt-4o/, 2024.*

Anja Prummer. 2020. Micro-targeting and polarization. *Journal of Public Economics*, 188:104210.

Filipe Nunes Ribeiro and 1 others. 2019. Inference of demographic data from digital advertising platforms based on social media.

Kate Roberts, Anthony Dowell, and Jing-Bao Nie. 2019. Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. *BMC medical research methodology*, 19:1–8.

Pasquale E Rummo, Omni Cassidy, Ingrid Wells, Jaime A Coffino, and Marie A Bragg. 2020. Examining the relationship between youth-targeted food marketing expenditures and the demographics of social media followers. *International journal of environmental research and public health*, 17(5):1631.

Matthew Sag. 2023. Fairness and fair use in generative ai. *Fordham Law Review, Forthcoming*.

Alberto Salguero and Macarena Espinilla. 2018. A flexible text analyzer based on ontologies: an application for detecting discriminatory language. *Language Resources and Evaluation*, 52:185–215.

Juan Carlos Medina Serrano and 1 others. 2020. The political dashboard: A tool for online political transparency. In *ICWSM*.

Márcio Silva and Fabrício Benevenuto. 2021. Covid-19 ads as political weapon. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pages 1705–1710.

Márcio Silva and 1 others. 2020. Facebook ads monitor: An independent auditing system for political ads on facebook. In *WWW*.

Mark CJ Stoddart, Randolph Haluza-DeLay, and David B Tindall. 2016. Canadian news media coverage of climate change: historical trajectories, dominant frames, and international comparisons. *Society & Natural Resources*, 29(2):218–232.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Anthony G Tuckett. 2005. Applying thematic analysis theory to practice: A researcher's experience. *Contemporary nurse*, 19(1-2):75–87.

Aleksandra Urman and Mykola Makhortykh. 2023. The silence of the llms: Cross-lingual analysis of political bias and false information prevalence in chatgpt, google bard, and bing chat.

Mojtaba Vaismoradi, Jacqueline Jones, Hannele Turunen, and Sherrill Snelgrove. 2016. Theme development in qualitative content analysis and thematic analysis.

Mojtaba Vaismoradi, Hannele Turunen, and Terese Bondas. 2013. Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing & health sciences*, 15(3):398–405.

Stefanie Walter, Michael Brüggemann, and Sven Engesser. 2018. Echo chambers of denial: Explaining user comments on climate change. *Environmental Communication*, 12(2):204–217.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models:

Combining codebook with gpt-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 75–78.

Yanwu Yang, Yinghui Catherine Yang, Bernard J Jansen, and Mounia Lalmas. 2017. Computational advertising: A paradigm shift for advertising and marketing? *IEEE Intelligent Systems*, 32(3):3–6.

Yong Zhang, Hongming Zhou, Nganmeng Tan, Saeed Bagheri, and Meng Joo Er. 2017. Targeted advertising based on browsing history. *arXiv preprint arXiv:1711.04498*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, pages 1–55.

Indrė Žliobaitė. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089.

A Appendix

A.1 Prompting

Fig. 3 illustrates the prompt template. Fig. 4 shows the example prompts for age group prediction from the climate campaign dataset. Prompt template for generating theme and aspects from predictions and explanations is shown in Fig. 5.

A.2 Demographic Misclassifications

Table 10 presents an analysis of ad misclassifications based on gender and age group predictions.

A.2.1 Misclassification - Senior Instances

Table 11 shows the details of misclassified senior instances. The misclassification of the Senior age group as Young or Late Working can be attributed to several factors:

Thematic Content and Topic Association: The first three misclassified ads (Table 11) focus on climate change, environmental activism, and sustainability. These topics are often associated with younger demographics, particularly young adults (18-24), who are perceived to be more engaged in activism and environmental causes. The model appears to have learned an association between these topics and the Young age group, leading to misclassification when seniors engage with similar content.

In Table 11, the fourth instance shows a misclassification where the model predicts it as a late working (45-64 years) age group. The ad mentions ‘working families’, which is a term commonly associated with individuals in the Late Working age group who are actively engaged in the workforce and supporting families. The content revolves around a political campaign emphasizing the need for change and active participation, themes often associated with the $45 - 64$ age demographic who are typically more politically active and influential.

Lack of Age-Specific Cues: The misclassified ads do not contain explicit references to seniors or age-specific language that would signal the content is intended for the senior demographic. The language is broad and does not mention age-related concerns, such as retirement, health issues prevalent among seniors, or senior-specific programs.

Reliance on Stereotypical Associations: The explanations generated by LLMs indicate that the model relies on stereotypes, associating certain topics exclusively with specific age groups. By assuming that environmental activism is primarily of interest to young adults, the model overgeneralizes by overlooking the possibility that seniors are also engaged in these issues.

Feature Representation Limitations: The model may lack features that capture subtle cues indicating the ad’s target age group when explicit age markers are absent. The model may not effectively utilize contextual information that could hint at the intended audience beyond topic associations. For example, the interests and concerns of the late working and senior age groups can overlap, especially in areas like politics and social change.

A.3 Thematic Analysis of Gender Explanations

From 1st row of Table 3, we notice that the overall theme revolves around targeting ***men*** by aligning ads with their **perceived interests and roles** in technology, finance, property, traditional male activities, and political or economic discourse. The explanations consistently emphasize the following aspects:

Interest in Technology and Innovation: Men are often depicted as being more engaged with technology, engineering, and renewable energy solutions. Ads related to technical aspects of engines, energy efficiency, and infrastructure are considered more likely to appeal to men.

Focus on Economic and Financial Issues: Many explanations suggest that men are more concerned with economic benefits, investment opportunities, and financial savings, making them the likely target

Ad: <Text >

Report who can be targeted based on gender: male vs female by the provided ad. Choose the most probable option from the gender (either male or female). Don't generate neutral or both gender answers. Provide a brief explanation with the reasoning in 20 words.
Gender: ?
Explanation: ?

Gender: <Predicted gender>
Explanation: <Prediction reasoning explanation>

(a) Prompt template for gender prediction.

Ad: <Text >

Report who can be targeted based on the following four age groups: Young adults (18-24), Early working age (25-44), Late working age (45-64), and Senior citizens (65+) by the provided ad. Choose the most probable option from the age groups. Don't generate neutral or multiple age group answers. Provide a brief explanation with the reasoning in 20 words.
Age group: ?
Explanation: ?

Age group: <Predicted age group>
Explanation: <Prediction reasoning explanation>

Figure 3: Prompt template for targeted demographic prediction. (a) gender, (b) age group. Inputs are shown in blue, and outputs are shown in red.

Ad: **Apply for a Summer Internship with 350NH! Take your organizing skills to the next level in the fight to stop the climate crisis.**

Report who can be targeted based on the following four age groups: Young adults (18-24), Early working age (25-44), Late working age (45-64), and Senior citizens (65+) by the provided ad. Choose the most probable option from the age groups. Don't generate neutral or multiple age group answers. Provide a brief explanation with the reasoning in 20 words.
Age group: ?
Explanation: ?

Age group: Young adults (18-24)
Explanation: Targeting students seeking internships to develop skills and actively participate in climate activism initiatives.

(a) Prompt example for age group prediction (young adults).

Ad: **Formula 1 recently announced efforts to become carbon-neutral in 2021 and net-zero by 2030. We applaud F1 for making the switch to 100% sustainable biofuel for all teams starting in 2026! Learn more:**

Report who can be targeted based on the following four age groups: Young adults (18-24), Early working age (25-44), Late working age (45-64), and Senior citizens (65+) by the provided ad. Choose the most probable option from the age groups. Don't generate neutral or multiple age group answers. Provide a brief explanation with the reasoning in 20 words.
Age group: ?
Explanation: ?

Age group: Early working age (25-44)
Explanation: This demographic likely has disposable income and a strong interest in sustainability and motorsports, aligning with the ad.

(b) Prompt example for age group prediction (early working).

Ad: **Investing in state parks, wildlife, forests, and public lands keeps Colorado thriving. Thank you for investing in these irreplaceable public health and economic resources.**

Report who can be targeted based on the following four age groups: Young adults (18-24), Early working age (25-44), Late working age (45-64), and Senior citizens (65+) by the provided ad. Choose the most probable option from the age groups. Don't generate neutral or multiple age group answers. Provide a brief explanation with the reasoning in 20 words.
Age group: ?
Explanation: ?

Age group: Late working age (45-64)
Explanation: The ad appeals to 45-64-year-olds who invest in public resources and value economic and environmental sustainability.

(c) Prompt example for age group prediction (late working).

Figure 4: Prompt examples for age group prediction. (a) young adults, (b) early working age, (c) late working age. Inputs are shown in blue, and outputs are shown in red.

Ad Text	**Actual**	**Prediction (Misclassification)**	**Explanations by LLMs**
Meet the first carbon negative fragrance. Air Eau de Parfum converts CO2 into something beautiful: sustainable ethanol blended with notes of orange peel, fig leaf, powdery musk and tobacco invigorating scent that redefines what is possible. Available exclusively at aircompany.com	Male	Female	The fragrance notes like fig leaf, powdery musk, and sustainable ethanol typically resonate more with female audiences in fragrance marketing.
Over generations, Illinois farm families have learned how to work with our climate to produce healthy crops and livestock.	Female	Male	The ad emphasizes farming and agriculture, traditionally male-dominated fields, making men the more likely target audience for this message.
Renewable energy is projected to surpass fossil gas as the dominant source of energy production by 2030. Texas, a leader in wind output, is in a great position as the nation transitions.	Young	Early working	This group is likely investing in sustainable energy options, driven by long-term economic and environmental considerations.
`Take Action Now to Fight Climate Change! Plant a tree in California treestoglobe.org/plantingchallengecalifornia`	Early working	Young	This age group is often more environmentally active and responds well to social media campaigns promoting climate action initiatives.
The worst impacts of climate change could be irreversible by 2030. The time to switch to a renewable energy plan is now. Not tomorrow. Not next week.	Late working	Early working	They can switch energy plans and are motivated to act before irreversible impacts by 2030.
Tell Rep. Schrader: Now is the time to go big on climate. VOTE YES on the Build Back Better Act.	Senior	Young	Targets young adults passionate about climate action and eager to influence political decisions for their future.

Table 10: Error Analysis.

```
What is the common theme and aspects of following <Number> predictions and their explanation:

Gender1:  <predicted gender1 >
Explanation1: <explanation text1>

Gender2:  <predicted gender2 >
Explanation2: <explanation text2>
:
:
:
Gender_n:  <predicted gender_n>
Explanation_n: <explanation text_n>

<Generated theme and aspects>
```

Figure 5: Prompt template for generating theme and aspects from predictions and explanations. Inputs are shown in blue, and outputs are shown in red.

Ad Text	**Prediction**	**Explanations by LLMs**
The Hebrew University Center for Climate Science was established in Israel to fight climate change worldwide.	Young	Targets university-age students interested in climate science to study and combat climate change.
Tell Rep. Schrader: Now is the time to go big on climate. VOTE YES on the Build Back Better Act.	Young	Targets young adults passionate about climate action and eager to influence political decisions for their future.
Get clean water and pollution-free electricity to all of America.	Young	Targets environmentally conscious young adults concerned about sustainability and the future impact of clean water and energy.
Let's make one thing clear: Pennsylvania will be the single most competitive Senate race of 2022... It's one of the reasons I am running for the U.S. Senate seat in Pennsylvania.	Late working	Targets 45-64-year-olds by focusing on working families, anti-celebrity politics, and referencing the Trump Era.

Table 11: Misclassified Senior Instances.

for ads that emphasize these aspects.

Property and Land Management: The theme of land ownership, property value improvement, and land management is frequently mentioned, with the assumption that men are more interested in these areas.

Traditional Male Activities: Ads that involve traditionally male-oriented activities, such as beer consumption, vehicle-related savings, home maintenance, and physical strength, are seen as more likely to target men.

Engagement in Political and Infrastructure Topics: Men are often portrayed as more engaged in political discourse, infrastructure initiatives, and discussions around energy and policy, making them the primary audience for ads focused on these themes.

Conservative Views and Skepticism: Some explanations suggest that men are more likely to resonate with conservative views, skepticism about environmental claims, and anti-establishment sentiments.

From 2nd row of Table 3, we observe that the overall theme revolves around targeting ***women*** by aligning ads with their **roles as caregivers, environmental advocates, and socially conscious individuals** who prioritize the well-being of their families, communities, and the environment. The explanations consistently emphasize the following aspects:

Parental and Caregiving Roles: Many explanations highlight that women, particularly mothers, are more likely to resonate with messages about protecting children's futures, parental responsibilities, and family well-being. These ads often appeal to maternal instincts and the role of women as primary caregivers.

Environmental Consciousness: Women are frequently depicted as being more engaged with environmental issues, sustainability, and community health. The explanations suggest that women are more proactive and vocal about climate change, conservation, and eco-friendly initiatives.

Social Welfare and Community Involvement: The explanations note that women are more likely to be concerned with social issues such as paid leave, affordable childcare, healthcare, and community well-being. Ads that emphasize these themes are seen as more likely to appeal to women.

Empathy and Emotional Appeal: The explanations often mention that women are more responsive to ads that evoke empathy, emotional concerns, and collective action. This includes ads that focus on protecting the environment for future generations and supporting social safety nets.

Female Empowerment and Leadership: Some explanations specifically mention themes of women's empowerment, leadership, and support for female scientists or leaders. These themes are likely to resonate more with female audiences who identify with or support gender equality and empowerment.

Health and Safety Concerns: Women are portrayed as being more attentive to issues related to health, safety, and the well-being of their families and communities. This includes a strong focus on environmental health and sustainability.

A.4 Thematic Analysis of Age Explanations

From 1st row of Table 6, we observe that the overall theme revolves around the **activism and the environmental consciousness** of ***young adults***, positioning them as a key demographic for campaigns and initiatives focused on climate change and sustainability. The explanations highlight the following aspects:

Passion for Climate Action: Young adults are described as being particularly passionate about addressing climate change, often leading or participating in environmental activism and campaigns.

Support for Bold Environmental Leadership: This age group is likely to support bold and urgent actions related to environmental protection and sustainability.

Engagement with Activism: The explanations emphasize that young adults are more likely to be engaged in climate-related activism and are motivated to take meaningful actions.

Desire for Immediate Change: There is a recurring mention of the desire for immediate and meaningful change, reflecting the urgency with which young adults approach environmental issues.

Participation in Training and Advocacy: The group is also characterized as eager to participate in training programs and initiatives that allow them to contribute actively to environmental causes.

From 2nd row of Table 6, we can see that the overall theme revolves around the **proactive and responsible mindset** of ***early working-age*** adults, who are not only financially capable but also motivated by a strong sense of social and environmental responsibility. They are seen as key targets for initiatives that combine sustainability with practical, career-oriented, and family-focused benefits. The explanations consistently emphasize the following aspects:

Environmental Consciousness: This age group is described as being highly engaged with environmental issues, such as climate change, sustainability, and clean energy. They are likely to support initiatives and products that align with eco-friendly values.

Financial Stability and Disposable Income: Many explanations note that individuals in this group have disposable income, making them financially capable of supporting and investing in sustainable products, services, and causes.

Parental and Future Concerns: This demographic is often portrayed as parents or future-focused individuals who are concerned about the impact of environmental issues on their children and future generations.

Career Engagement and Professional Roles: The explanations frequently mention that this age group is active in their careers, often holding decision-making roles that influence corporate and household sustainability practices.

Interest in Innovation and Technology: Individuals in this age group are also depicted as being interested in innovative industries, clean energy solutions, and sustainability technologies, which align with their professional and personal goals.

Social and Political Engagement: The group is characterized as being engaged in socio-political issues, particularly those related to corporate accountability, sustainability, and environmental advocacy.

From 3rd row of Table 6, we can notice that the overall theme revolves around the **responsibilities and concerns** of individuals in the ***late working (45-64) age*** group, focusing on their roles as homeowners, voters, and economically engaged citizens who are likely to be influenced by environmental, economic, and policy-related messaging. The explanations specifically emphasize the following aspects:

Economic and Environmental Responsibility: Many of the explanations mention that individuals in this age group are concerned with sustainability, home energy efficiency, and environmental impact. They are likely to invest in public resources and adopt changes that contribute to economic and environmental sustainability.

Homeownership and Financial Stability: This demographic is characterized as established homeowners who are financially secure. They are seen as key targets for changes related to home energy efficiency, such as adopting solar power, due to their financial means and homeownership status.

Voter and Policy Engagement: The explanations suggest that this age group is politically active, particularly concerned with public safety, and likely to support policy changes by voting on local measures.

Economic Concerns: There is an emphasis on economic factors such as unemployment, inflation, and gas prices, with concerns about current economic policies affecting their businesses and financial stability.

From 4th row of Table 6, we notice that the overall theme centers on **health and safety concerns** that are particularly important to ***senior citizens***, with a focus on programs that cater to their specific needs and the heightened risks they face in certain situations. The key aspects highlighted in the explanations are:

Health and Wellness Programs: The first explanation mentions programs like SilverSneakers and Silver&Fit, which are specifically designed for senior citizens to support their physical health and well-being.

Vulnerability and Safety: The second explanation focuses on the increased vulnerability of seniors to COVID-19 and emphasizes the risks they face, particularly in the context of political decisions or public health issues.

A.5 Unsupervised Baselines

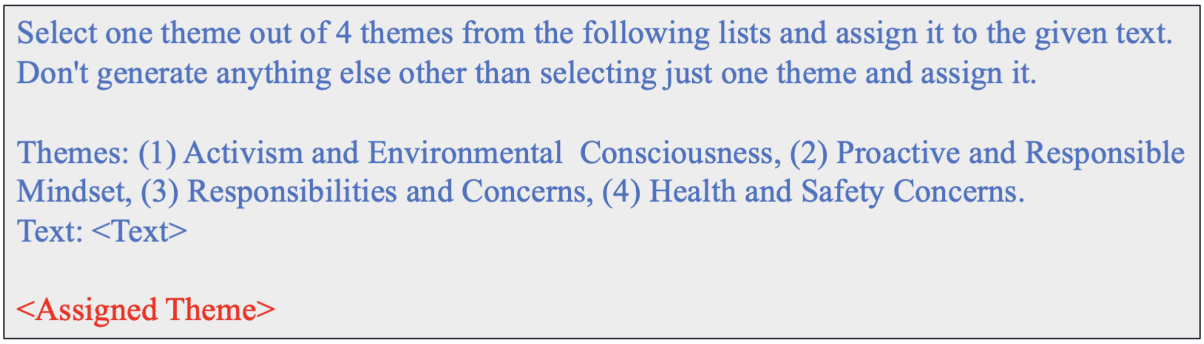
Generated topics from the prediction explanation of gender and age using LDA and BERTopic is shown in Table 13. Table 12 shows the human evaluation results on GPT-4o assignment, LDA, and BERTopic. Prompt template for LLM assignment is shown in Fig. 6.

Model	**Demo.**	**Acc. (%)**
LDA	gender	61.32
	age	33.06
BERTopic	gender	56.60
	age	45.45
GPT-4o	gender	**88.67**
	age	**68.56**

Table 12: Assignment Comparison w.r.t. Human Judgments. Demo.: Demographics, Acc.: Accuracy.

Dimension	Method	Topics / Top Words
Gender	BERTopic	**Topic 1**: Practical and Economic Appeals to Men (Technology, Property, Costs) **Topic 2**: Emotional and Social Appeals to Women (Spirituality, Motherhood, Environment)
	LDA	**Topic 1**: like, resonate, men, ads, males, focus, property, care, leave, paid **Topic 2**: men, ad, male, energy, traditionally, engagement, likely, emphasizes, climate, higher **Topic 3**: ad, female, environmental, emphasizes, resonate, typically, males, audiences, target, ads **Topic 4**: women, ad, environmental, making, issues, future, likely, mothers, children, concerned
Age	BERTopic	**Topic 1**: Career-Oriented Adults Focused on Sustainability and Economic Impact **Topic 2**: Young, Socially Active, and Environmentally Conscious Professionals
	LDA	**Topic 1**: climate, action, ad, likely, change, young, passionate, targets, eager, adults **Topic 2**: likely, engaged, environmental, group, energy, climate, issues, adults, support, typically **Topic 3**: group, likely, age, sustainability, energy, aligning, engaged, clean, career, corporate **Topic 4**: group, conscious, environmentally, sustainable, making, disposable, income, likely, like, products

Table 13: Comparison of BERTopic and LDA-generated topics for demographic dimensions of Gender and Age. BERTopic provides interpretable topic labels, while LDA results include top-10 words per topic.

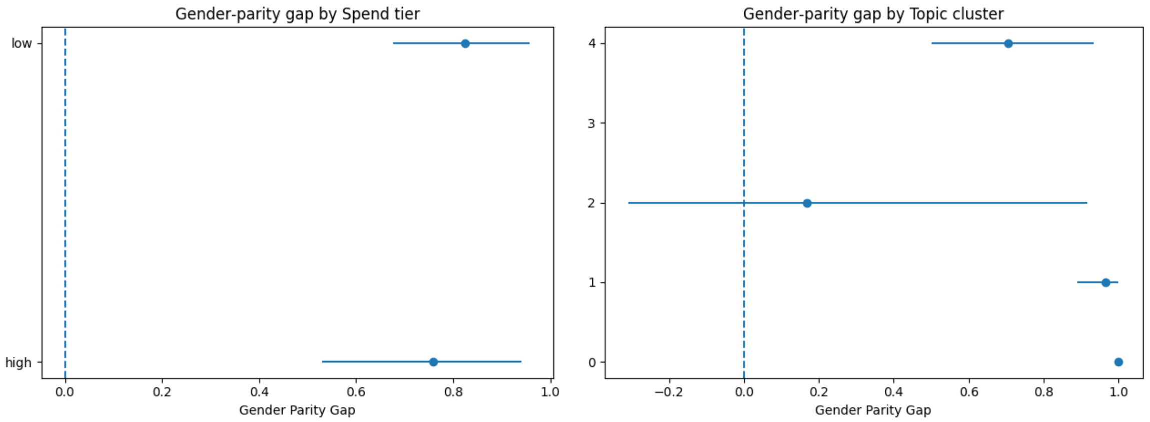


Select one theme out of 4 themes from the following lists and assign it to the given text. Don't generate anything else other than selecting just one theme and assign it.

Themes: (1) Activism and Environmental Consciousness, (2) Proactive and Responsible Mindset, (3) Responsibilities and Concerns, (4) Health and Safety Concerns.
Text: <Text>

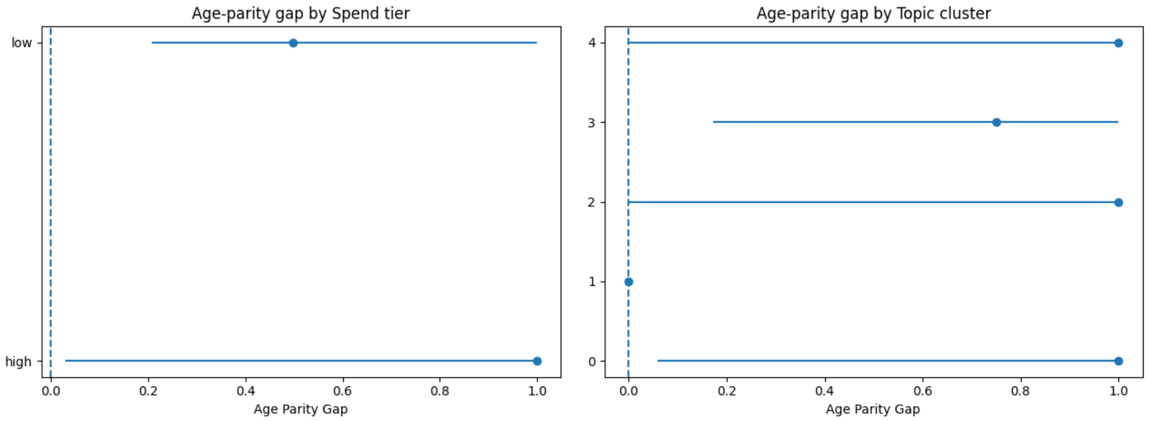
<Assigned Theme>

Figure 6: Prompt template for assigning theme using LLM. Inputs are shown in blue, and outputs are shown in red.



(a) Fairness by spend tier (Gender).

(b) Fairness by topic cluster (Gender).



(c) Fairness by spend tier (Age)

(d) Fairness by topic cluster (Age).

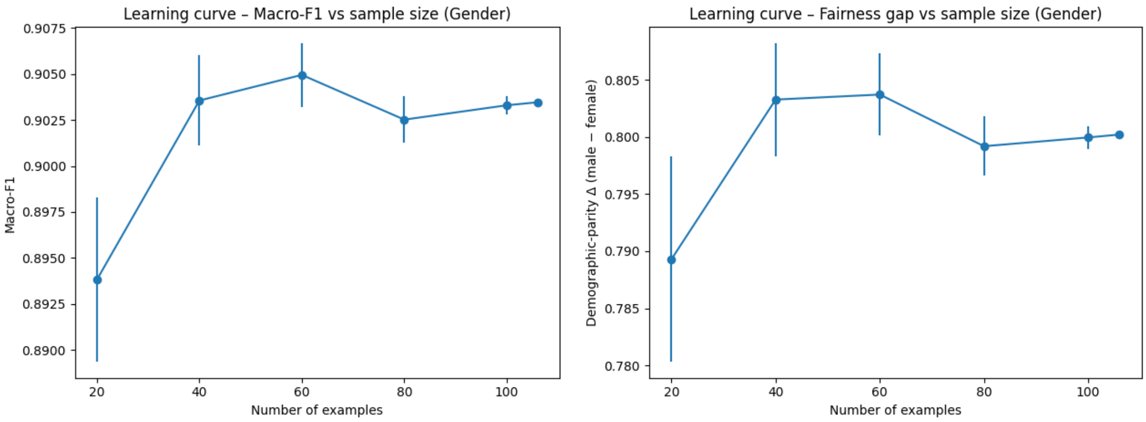
Figure 7: Sensitivity Analysis: Rerunning the fairness test after stratifying by ad-spend and by topic cluster.

A.6 Sensitivity Analysis on Fairness

We explicitly frame our work as a **post-hoc audit** and the manuscript avoids causal verbs such as “impact” or “cause” when reporting accuracy or fairness metrics. Fig. 7 shows the sensitivity analysis by rerunning the fairness test after stratifying by ad-spend and by topic cluster.

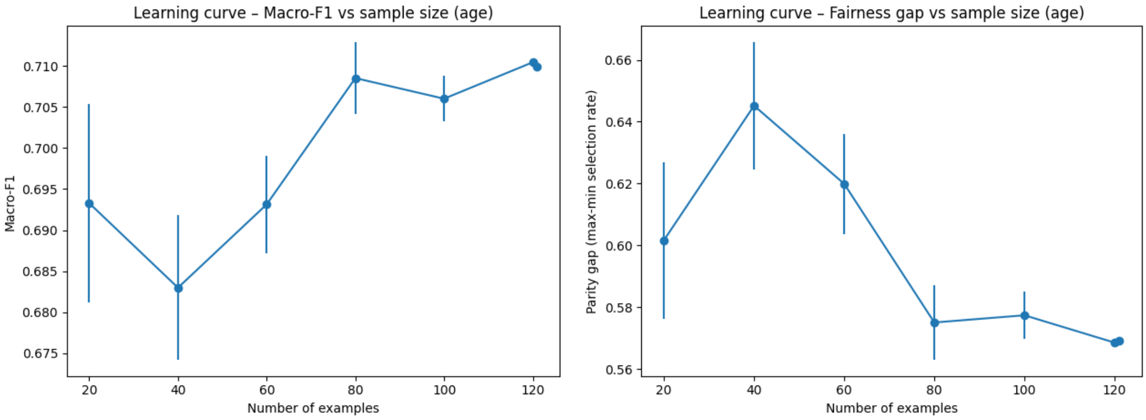
A.7 Learning-Curve Robustness Analysis

Fig. 8 shows the learning curve having macro average F1 score vs. sample size and fairness gaps vs. sample size.



(a) F1 vs sample size (Gender).

(b) Parity gap vs sample size (Gender).



(c) F1 vs sample size (Age).

(d) Parity gap vs sample size (Age).

Figure 8: Learning-curve robustness analysis.