Discovering Latent Themes in Social Media Messaging: A Machine-in-the-Loop Approach Integrating LLMs

Tunazzina Islam, Dan Goldwasser

Department of Computer Science

Purdue University, West Lafayette, IN 47907, USA

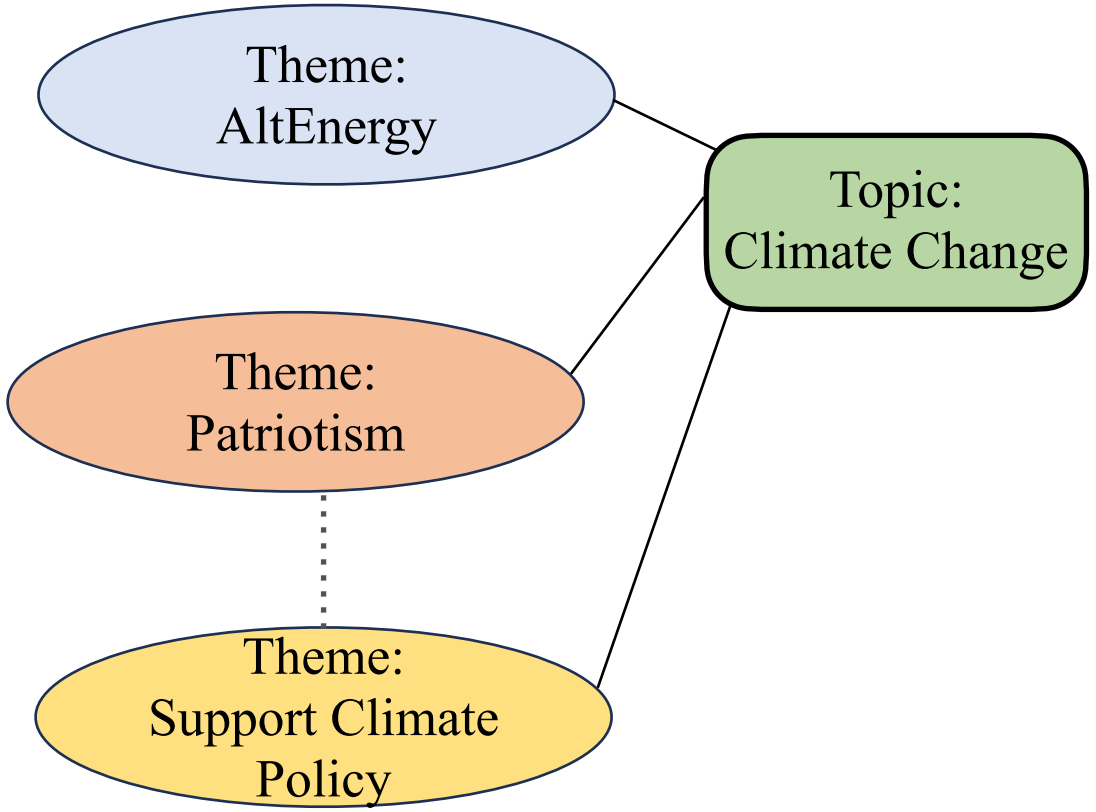
International AAAI Conference on Web and Social Media

June 23-26, 2025, Copenhagen, Denmark

PURDUE UNIVERSITY

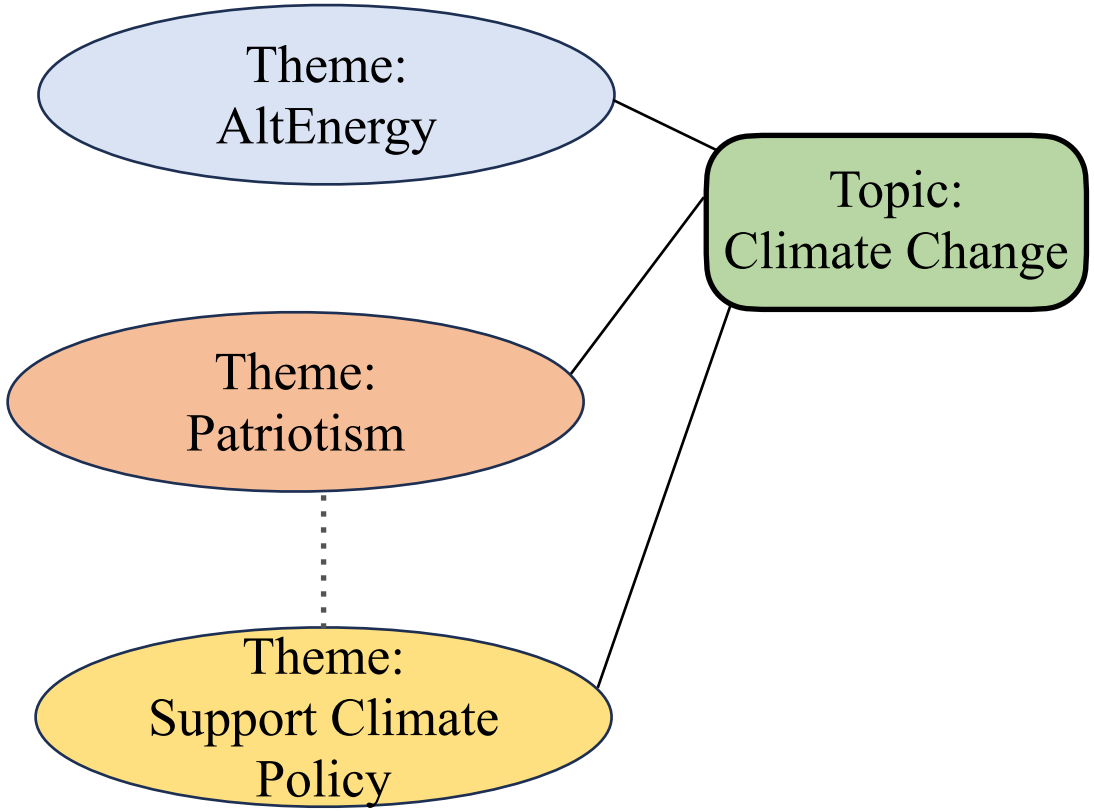
Unsupervised Text Analysis

- Topic Modeling.
  - Shallow Themes.



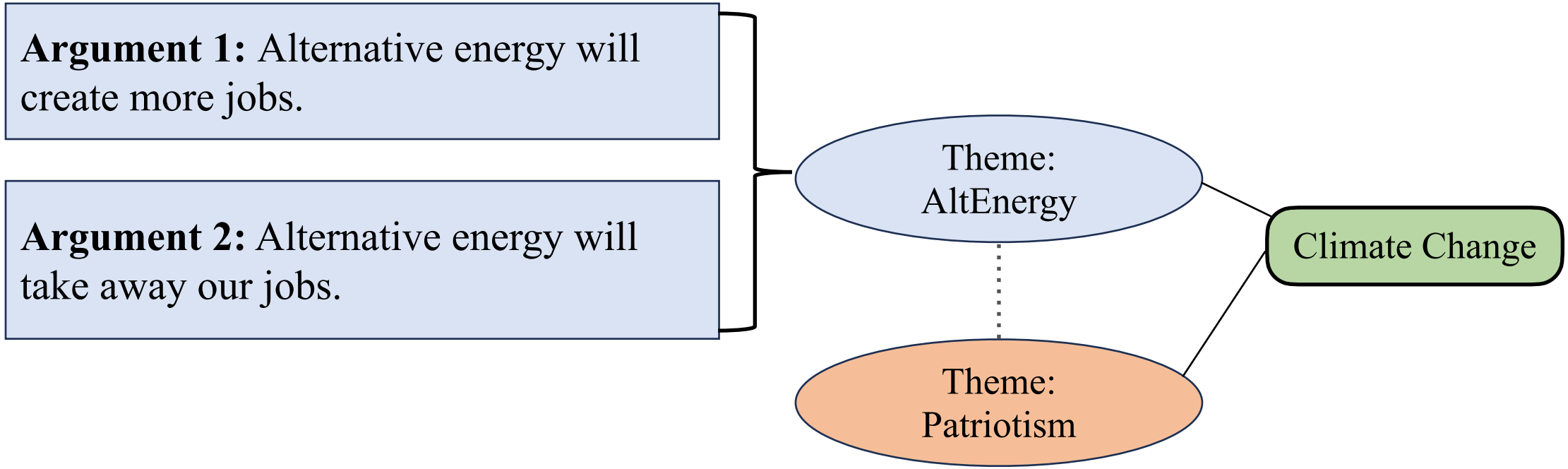
Theme Discovery - Previous Works

- Topic Modeling.
  - Shallow Themes.
- Manual and qualitative coding *(Hagen et al., 2022; Nguyen et al., 2021; Del Valle et al., 2020)*.



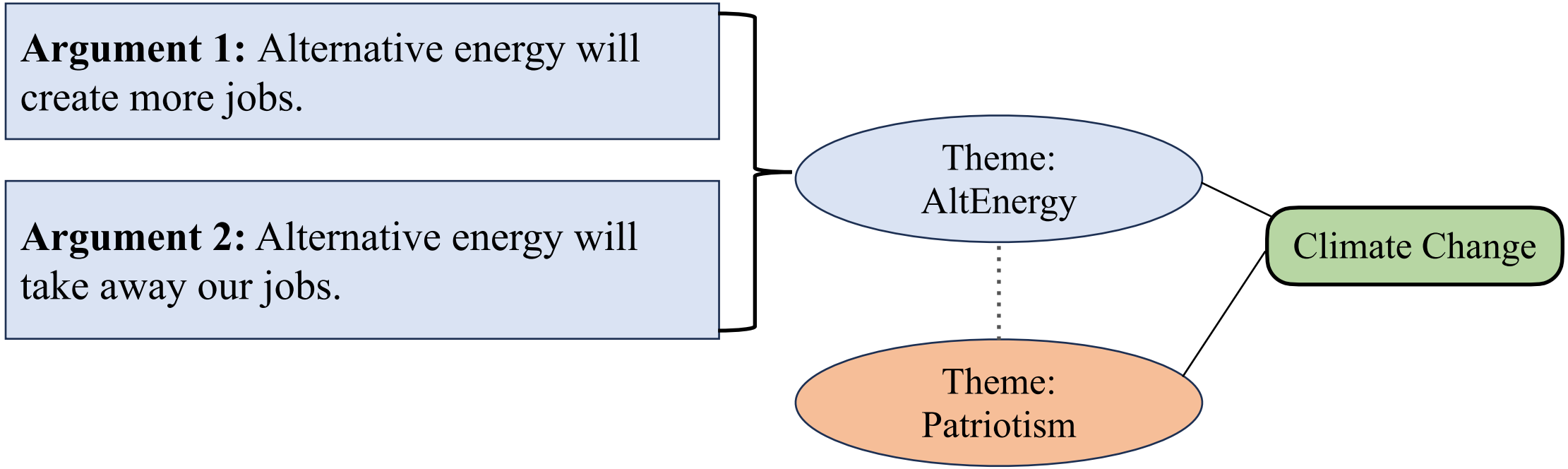
Theme Discovery - Previous Works

- **Predefined set** of labels, themes, and arguments *(Islam et al.,2023; Islam & Goldwasser, 2022)*.
  - **Fixed and established based on existing topics or theoretical frameworks**, such as Moral Foundations Theory (MFT) *(Haidt and Graham 2007)*.
  - Often **fails to capture the nuances** of messaging choices.



Theme Discovery - Previous Works

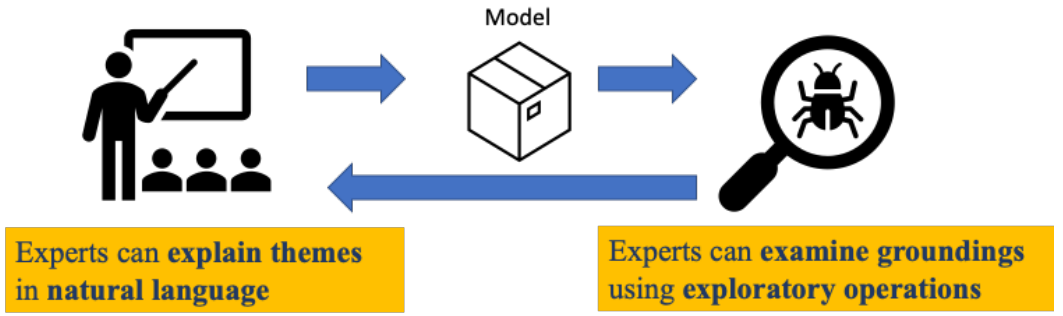
- **Predefined set** of labels, themes, and arguments *(Islam et al.,2023; Islam & Goldwasser, 2022)*.
  - **Fixed and established based on existing topics or theoretical frameworks**, such as Moral Foundations Theory (MFT) *(Haidt and Graham 2007)*.
  - Often **fails to capture the nuances** of messaging choices.



- Latent Theme Discovery *(Pacheco et al., 2023; Pacheco et al., 2022b;a)*.

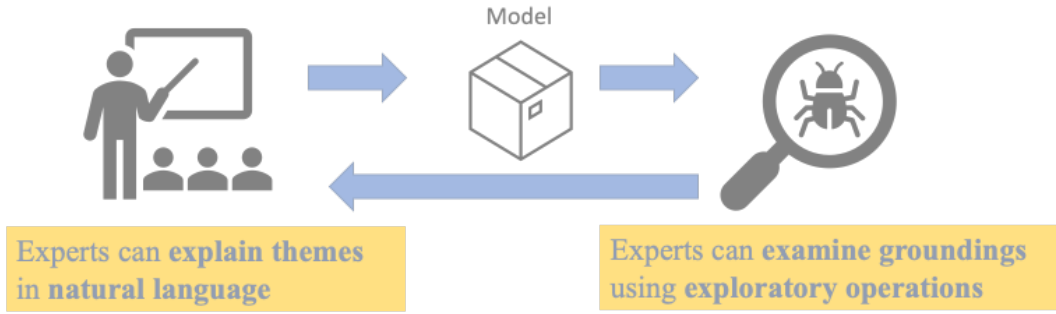
Theme Discovery - Previous Works

- Human-in-loop *(Pacheco et al. 2022b;a)*.
  - Costly scalability.
  - Time consuming.

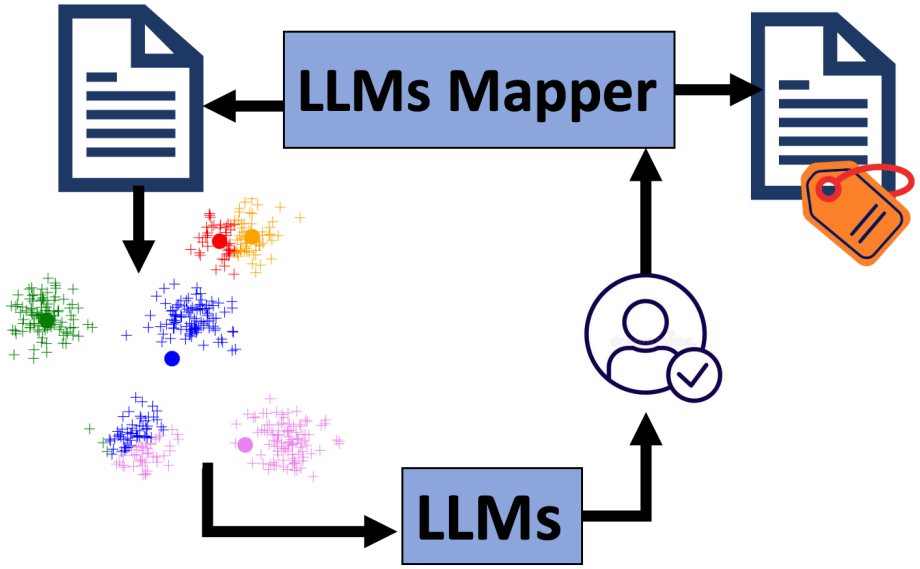


Machine-in-the-Loop Approach

- Human-in-loop *(Pacheco et al. 2022b;a)*.
  - Costly scalability.
  - Time consuming.



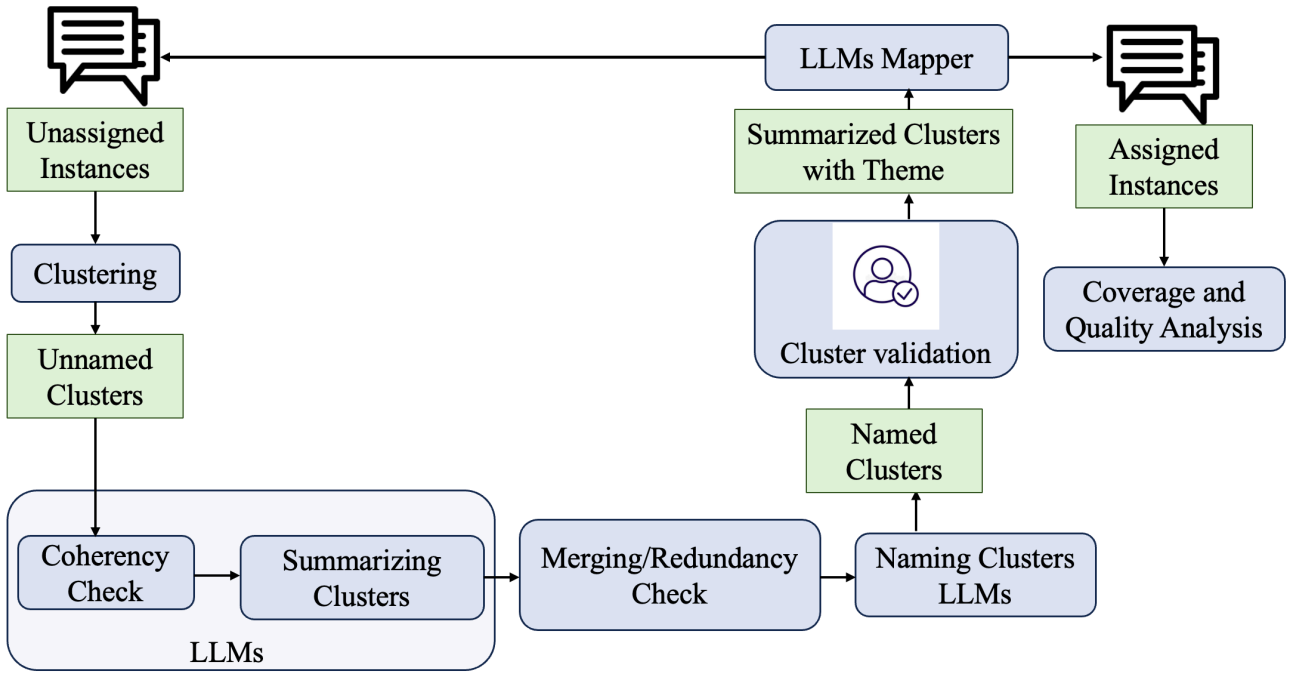
- **Machine-in-the-Loop: LLMs-in-the-Loop.**
  - LLMs possess **extensive domain insights**.
  - **Reasoning** capabilities.
  - **Accelerate** the process of refinement.



Sketch of *Machine-in-the-Loop* Approach

3 steps Process:

- **Candidate generation** –
  - (a) Clustering.
  - (b) Cluster coherency check: LLMs
  - (c) Cluster summery: LLMs
  - (d) Merging/redundancy check.

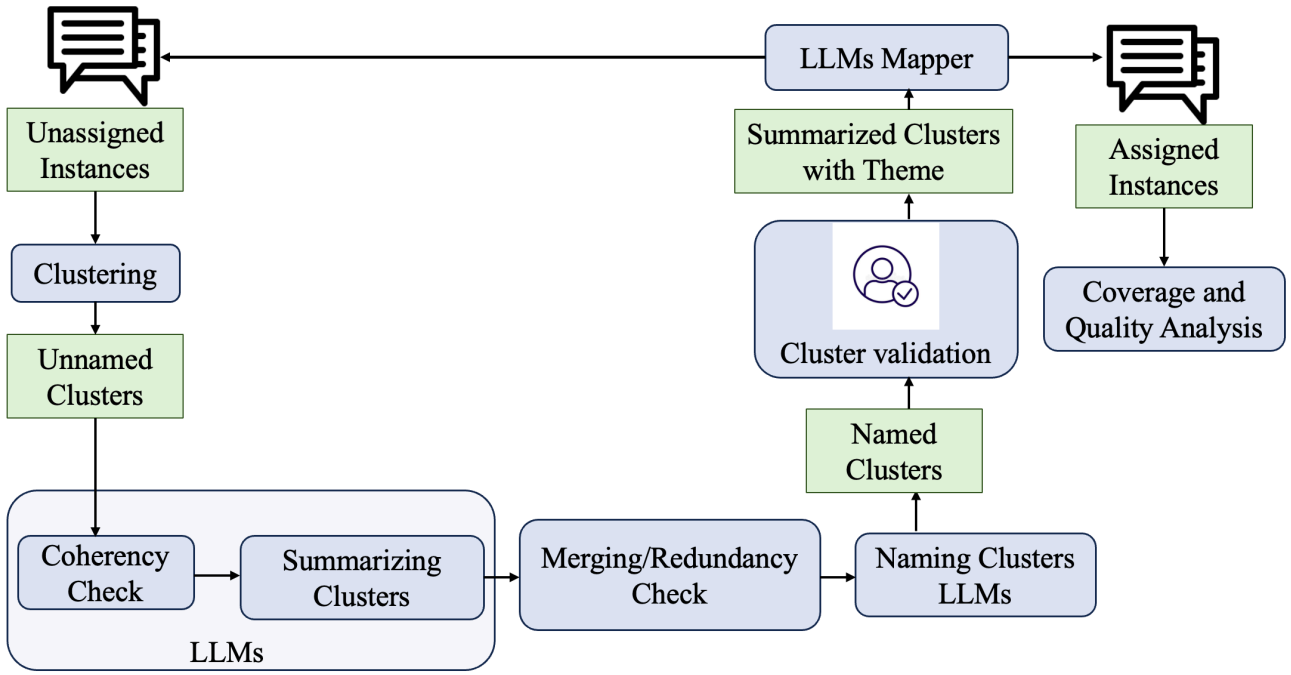


Sketch of *Machine-in-the-Loop* Approach

3 steps Process:

- **Candidate generation** –
  - (a) Clustering.
  - (b) Cluster coherency check: LLMs
  - (c) Cluster summery: LLMs
  - (d) Merging/redundancy check.

- **Human validation** - given the new set of clusters decide if - (a) Should we merge? (b) Does the cluster summary seem ok? (c) Coherent?



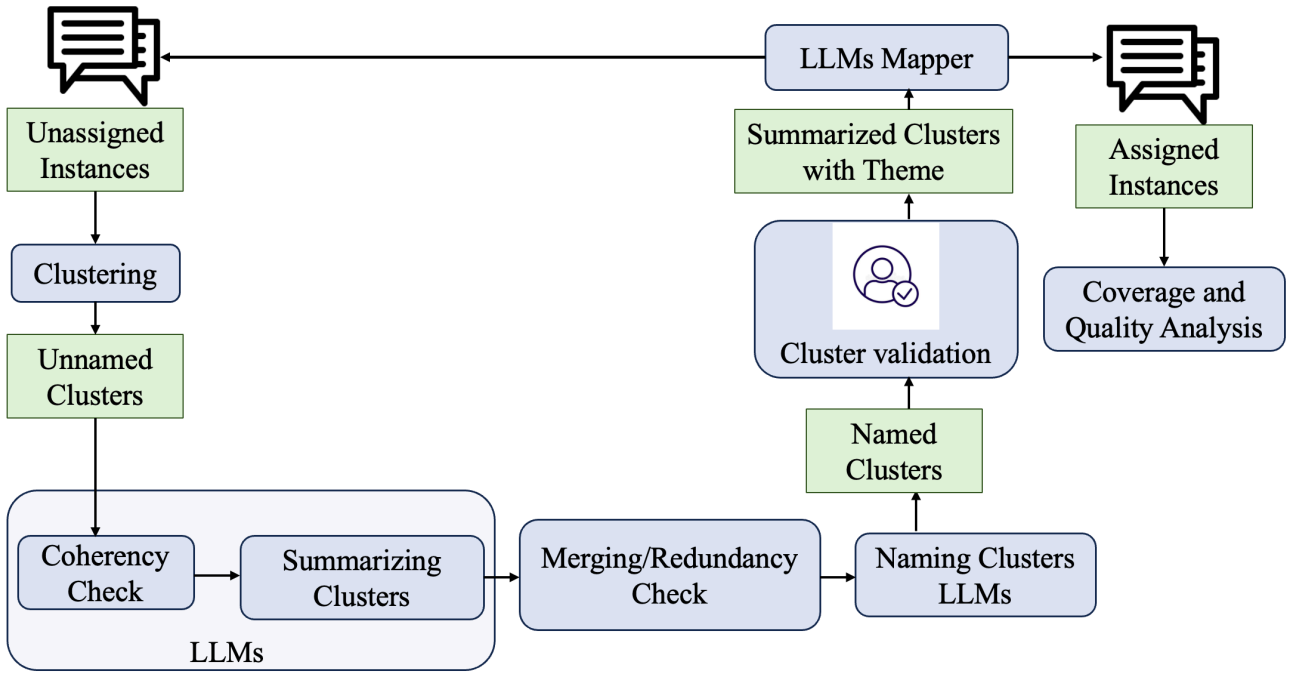
Sketch of *Machine-in-the-Loop* Approach

3 steps Process:

- **Candidate generation** –
  - (a) Clustering.
  - (b) Cluster coherency check: LLMs
  - (c) Cluster summery: LLMs
  - (d) Merging/redundancy check.

- **Human validation** - given the new set of clusters decide if - (a) Should we merge? (b) Does the cluster summary seem ok? (c) Coherent?

- **Assignment** - check if a new text belongs to the cluster summary by few-shot prompting LLMs.



Sketch of *Machine-in-the-Loop* Approach

3 steps Process:

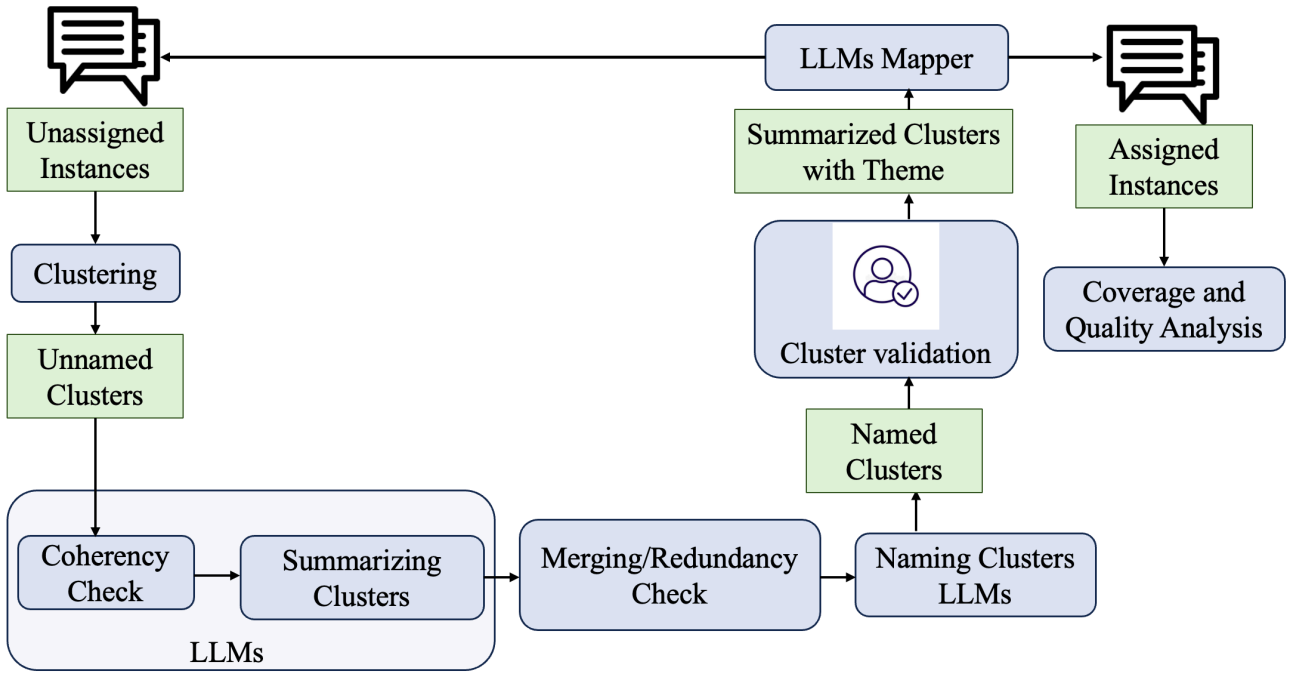
- **Candidate generation** –
  (a) Clustering.
  (b) Cluster coherency check: LLMs
  (c) Cluster summery: LLMs
  (d) Merging/redundancy check.

GPT-4

- **Human validation** - given the new set of clusters decide if - (a) Should we merge? (b) Does the cluster summary seem ok? (c) Coherent?

GPT-4

- **Assignment** - check if a new text belongs to the cluster summary by few-shot prompting LLMs.

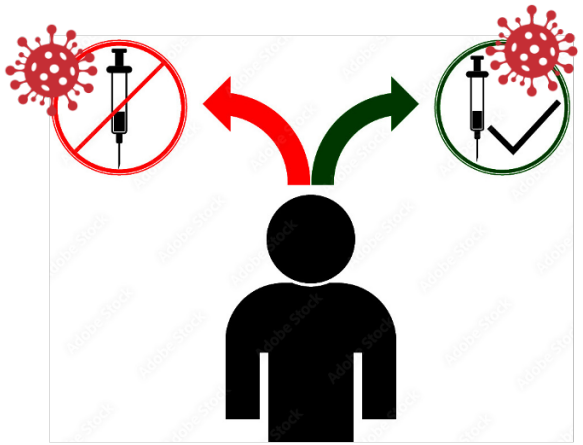


Case Studies

- **Climate** campaigns.
  - **21k ads** *(Islam et al. 2023)*, January 2021 to January 2022.
  - **Stance** (e.g., *pro-energy, clean-energy*) and **seed theme** (e.g., *support climate policy*).

Case Studies

- **Climate** campaigns.
  - **21k ads** *(Islam et al. 2023)*, January 2021 to January 2022.
  - **Stance** (e.g., *pro-energy, clean-energy*) and **seed theme** (e.g., *support climate policy*).
- **COVID-19** vaccine campaigns.
  - **9k ads** *(Islam and Goldwasser 2022)*, December 2020 to January 2022.
  - **Moral foundation** (e.g., *care/harm*) *(Haidt and Graham, 2007)* and **seed theme** (e.g., *vaccine equity*).



Results: Coverage

- SBERT embedding for theme assignment for each ad.

Case Study	Method	Num. Themes	Num. Covered Ads			
			THR < 0.6	THR < 0.5	THR < 0.4	THR < 0.3
Climate	Pre-existing	13	14652	9725	3731	558
	+After Iter1	20	18702	14583	8646	2944
	+After Iter2	**25**	**18988**	**15052**	**9079**	**3180**
COVID-19	Pre-existing	15	7889	6426	3480	771
	+After Iter1	20	8852	7627	4737	1302
	+After Iter2	**23**	**9092**	**7898**	**5038**	**1590**

Results: Coverage

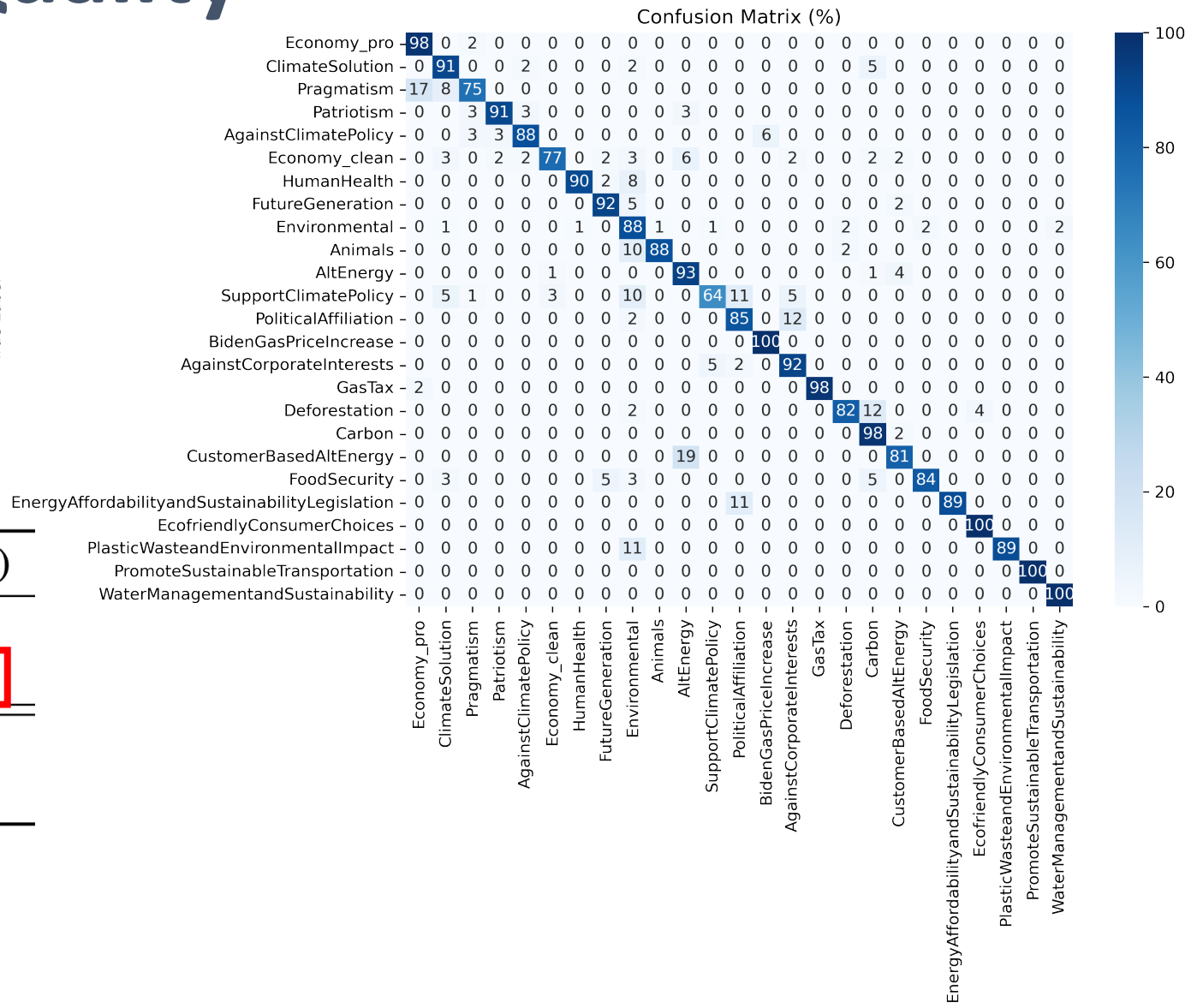
- SBERT embedding for theme assignment for each ad.
- **Better coverage** after **two iterations** of *machine-in-the-loop* approach.

Case Study	Method	Num. Themes	Num. Covered Ads			
			THR < 0.6	THR < 0.5	THR < 0.4	THR < 0.3
Climate	Pre-existing	13	14652	9725	3731	558
	+After Iter1	20	18702	14583	8646	2944
	+After Iter2	**25**	**18988**	**15052**	**9079**	**3180**
COVID-19	Pre-existing	15	7889	6426	3480	771
	+After Iter1	20	8852	7627	4737	1302
	+After Iter2	**23**	**9092**	**7898**	**5038**	**1590**

Results: Mapping Quality

- Ground truth:
  - 1072 climate ads.
  - 565 COVID-19 ads.
- Mapping Quality w.r.t Human Judgements.

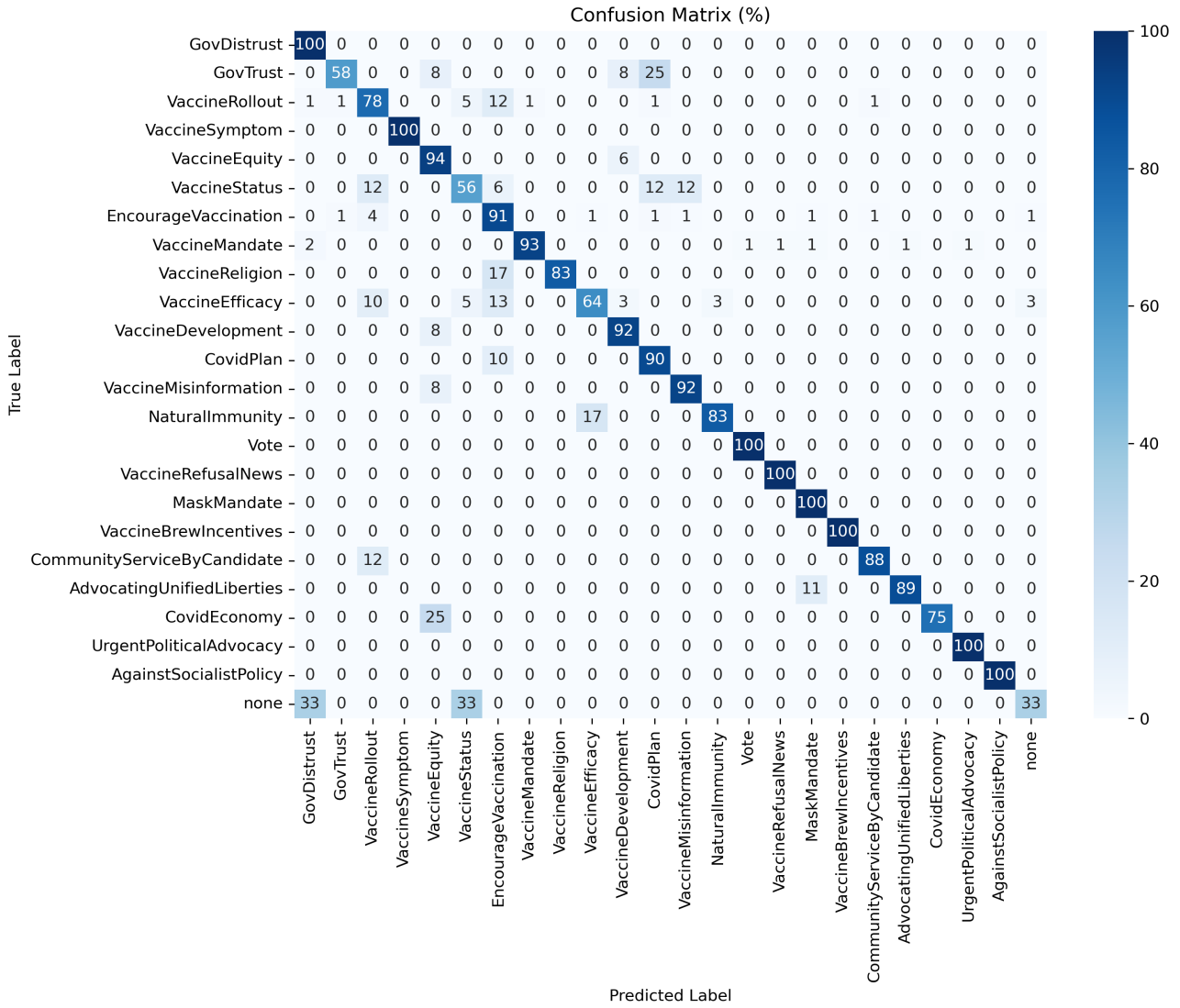
Case Study	Method	Acc. (%)	F1 (%)
Climate	SBERT Assign.	84.05	79.32
	LLMs Mapper	**88.15**	**89.24**
COVID-19	SBERT Assign.	41.42	44.83
	LLMs Mapper	**85.49**	**81.74**



Results: Mapping Quality

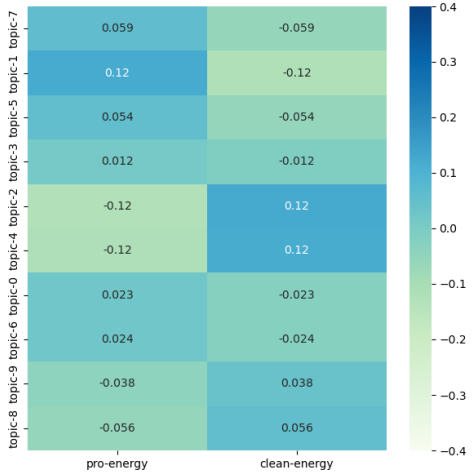
- Ground truth:
  - 1072 climate ads.
  - 565 COVID-19 ads.
- Mapping Quality w.r.t Human Judgements.

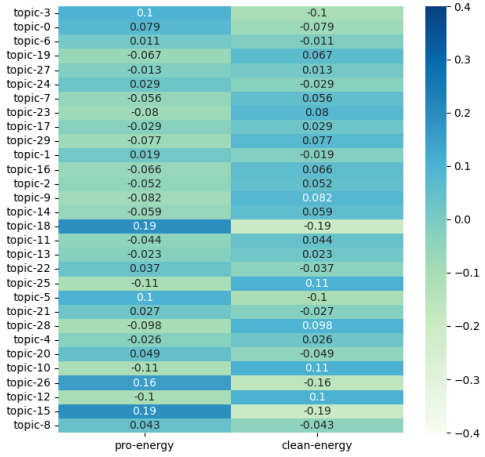
Case Study	Method	Acc. (%)	F1 (%)
Climate	SBERT Assign.	84.05	79.32
	LLMs Mapper	**88.15**	**89.24**
COVID-19	SBERT Assign.	41.42	44.83
	LLMs Mapper	**85.49**	**81.74**

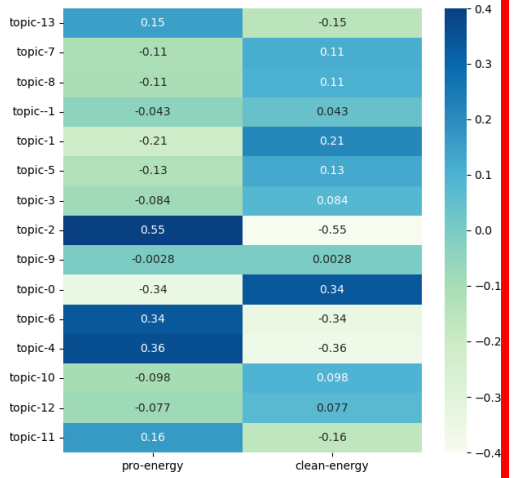


Qualitative Analysis: Climate

- Correlation heatmap between identified **themes** and **stances** expressed in the ads (i.e., pro-energy or clean-energy).
- **Stronger correlations with stances** than the derived LDA and BERTopic topics.

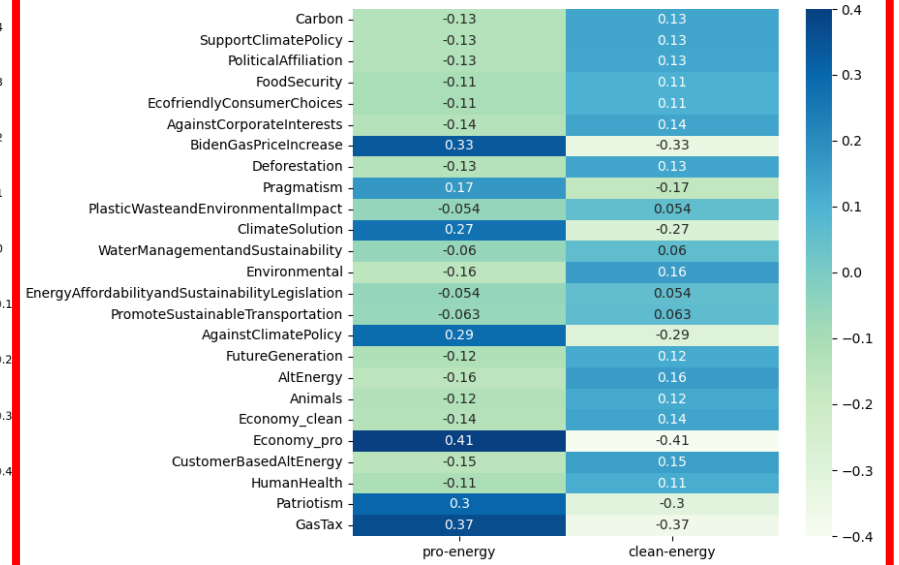






Pro-Energy → 'BidenGasPriceIncrease' & 'GasTax'

Clean-Energy → 'Environmental' & 'AltEnergy'



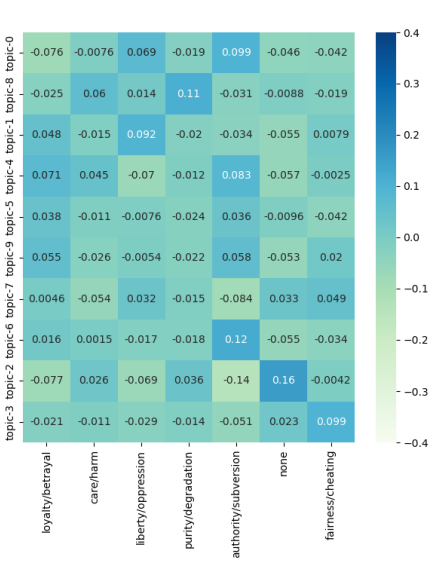
Baseline: 10 LDA Topics **Baseline:** 30 LDA Topics **Baseline:** 15 BERTopic Topics **Ours: After 2nd round of iteration**

Qualitative Analysis: COVID-19

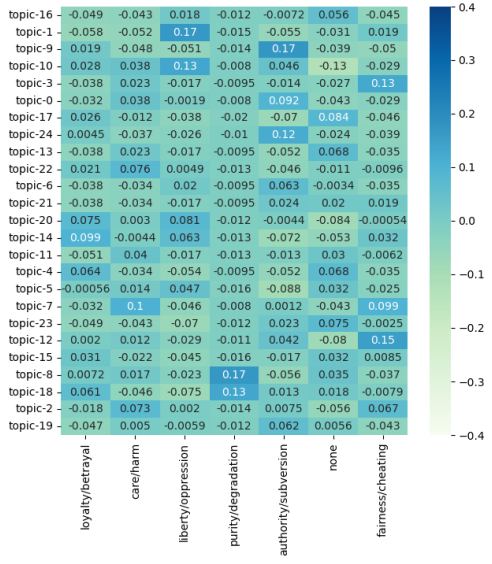
- Correlation heatmap between identified **themes** and **moral foundation** expressed in the ads (i.e., liberty/ oppression, fairness/cheating).
- **Stronger correlations with moral foundations** than the derived LDA and BERTopic topics.

⚖️ Fairness/Cheating → 'vaccine equity'

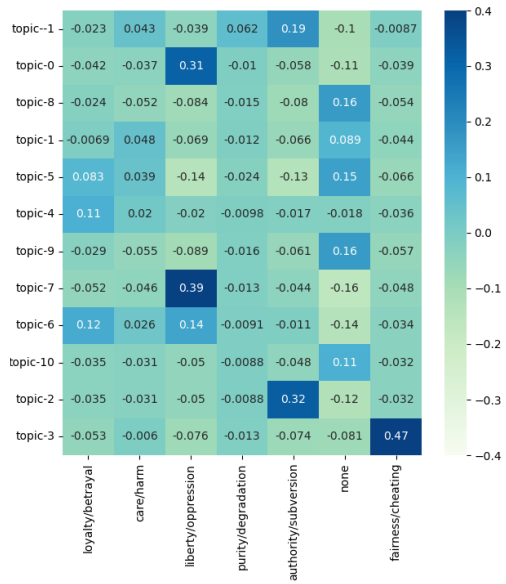
🗽 Liberty/Oppression → 'vaccine mandate' & 'advocating unified liberties'



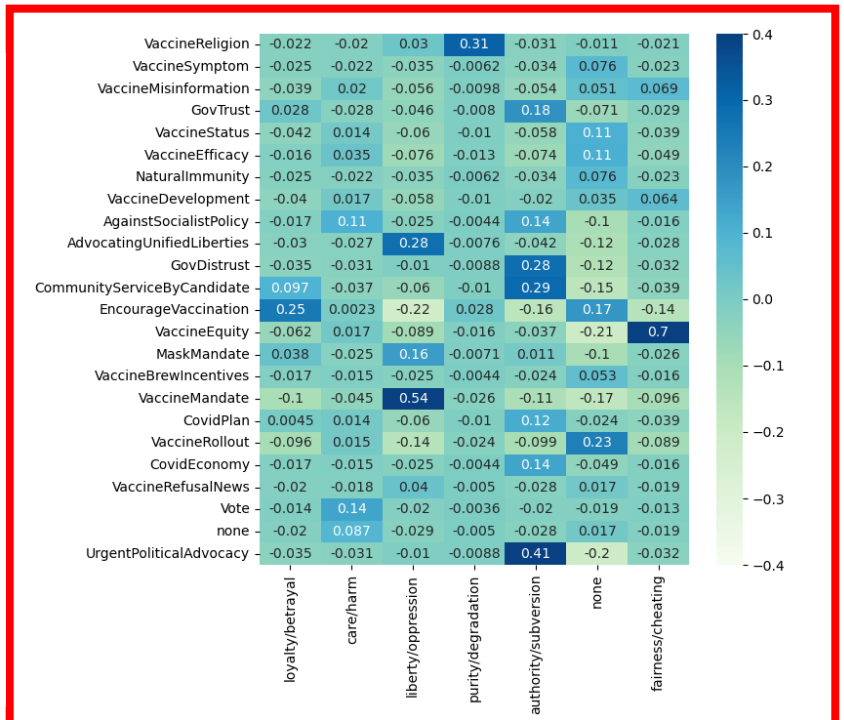
Baseline: 10 LDA Topics



Baseline: 25 LDA Topics



Baseline: 15 BERTopic Topics



Ours: After 2nd round of iteration

Demographic Targeting

- How Themes Differ based on **Gender**.
- Extract ads **targeted exclusively** at ***males*** and ***females***.

Demographic Targeting

- How Themes Differ based on **Gender**.
- Extract ads **targeted exclusively** at ***males*** and ***females***.

Theme: GasTax

Demographic Targeting

- How Themes Differ based on **Gender**.
- Extract ads **targeted exclusively** at ***males*** and ***females***.

Theme: GasTax

Theme: FutureGeneration

Demographic Targeting

- How Themes Differ based on **Red** vs. **Blue** States.
  - ***North Dakota (ND)*** vs. ***Vermont (VT)***
- Different emphasis placed on the *entity* '**Community**'.

Demographic Targeting

- How Themes Differ based on **Red** vs. **Blue** States.
  - ***North Dakota (ND)*** vs. ***Vermont (VT)***
- Different emphasis placed on the *entity* '**Community**'.

Theme: Economy

Claim: Oil & gas supports jobs and **community** growth.

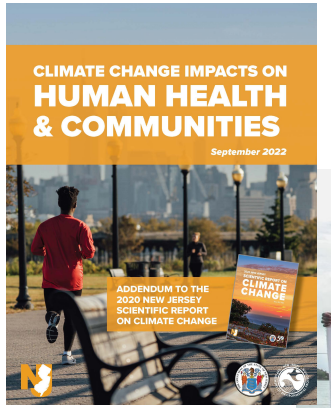
Demographic Targeting

- How Themes Differ based on **Red** vs. **Blue** States.
  - ***North Dakota (ND)*** vs. ***Vermont (VT)***
- Different emphasis placed on the *entity* '**Community**'.

Theme: Economy

Claim: Oil & gas supports jobs and **community** growth.

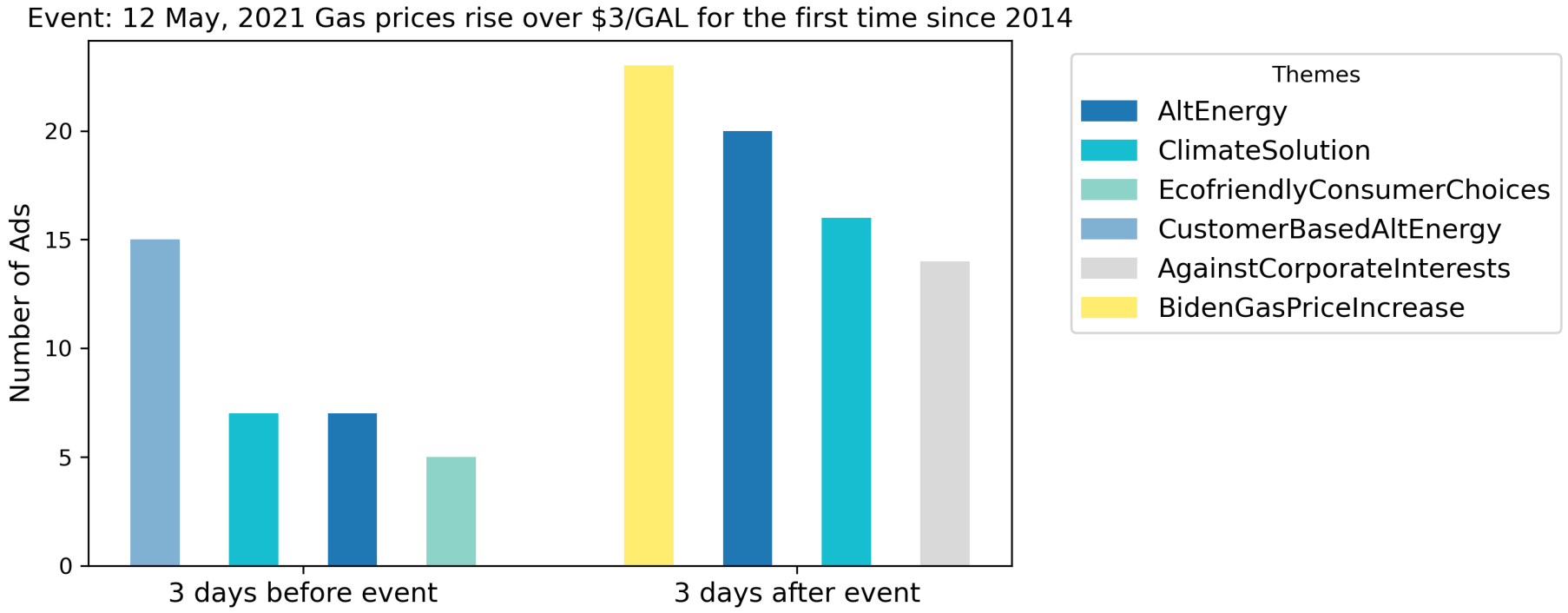
Theme: HumanHealth



Claim: Climate change impacts health, family & **community**.

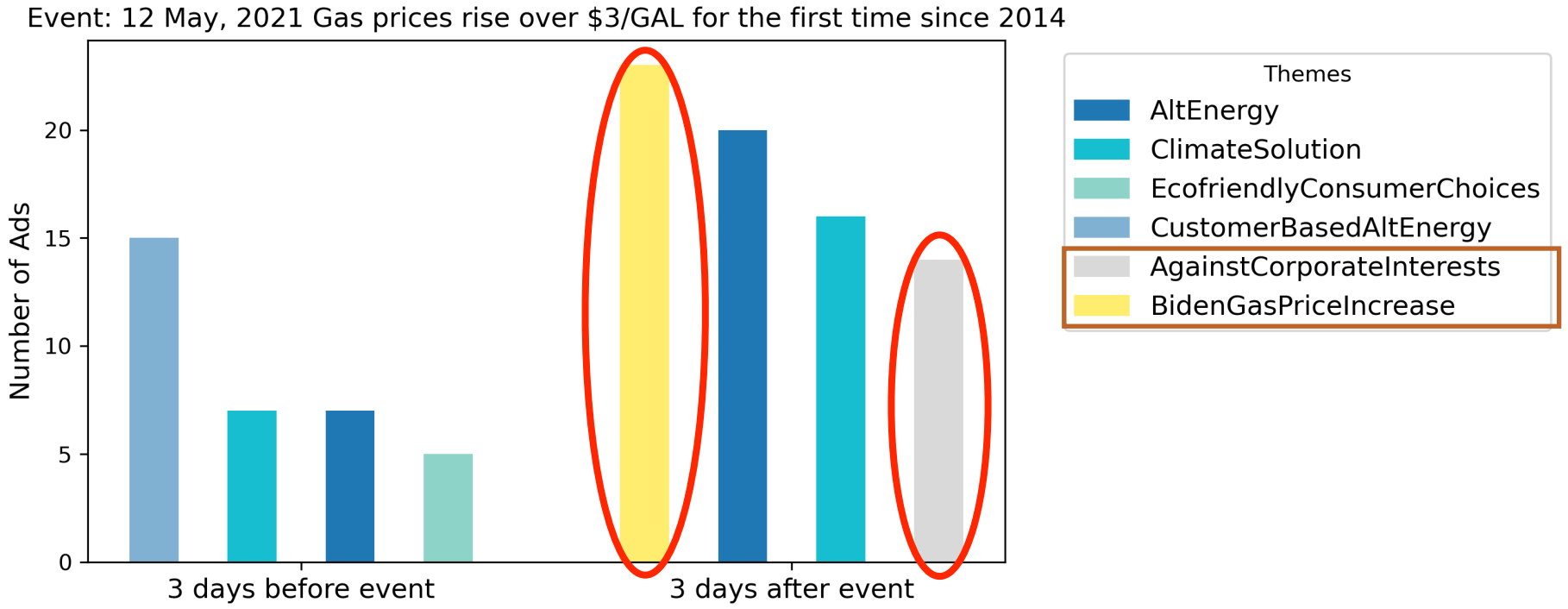
Theme Shifts Triggered by Key Events

- Event1: **Gas Price Increase**, Date: **May 12, 2021**.
- Theme Freq.: AltEnergy, ClimateSolution **increased**.



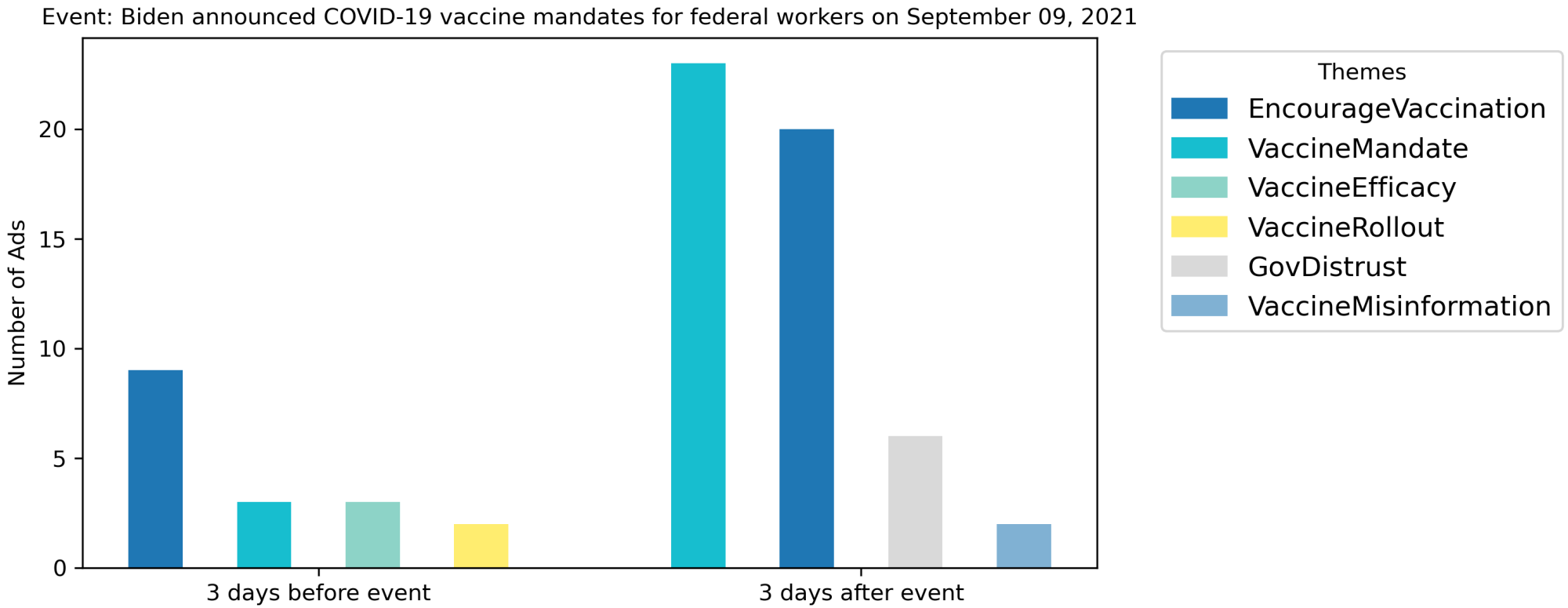
Theme Shifts Triggered by Key Events

- Event1: **Gas Price Increase**, Date: **May 12, 2021**.
- Theme Freq.: AltEnergy, ClimateSolution **increased**.
- New Themes: **AgainstCorporateInterests**, **BidenGasPriceIncrease**.



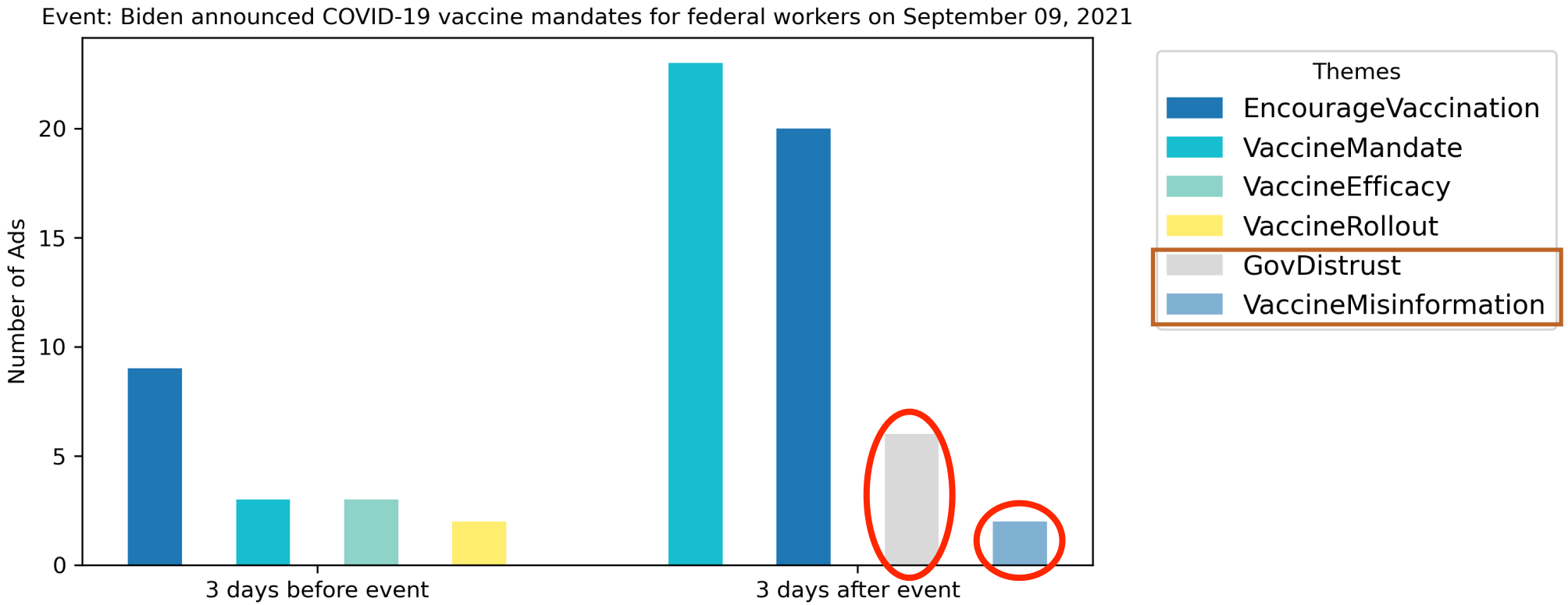
Theme Shifts Triggered by Key Events

- Event2: **Federal COVID-19 vaccine mandate**, Date: **September 09, 2021**.
- Theme Freq.: VaccineMandate **increased**.



Theme Shifts Triggered by Key Events

- Event2: **Federal COVID-19 vaccine mandate**, Date: **September 09, 2021**.
- Theme Freq.: VaccineMandate **increased**.
- New Themes: **GovDistrust**, **VaccineMisinformation**.



References

1. Carroll J Glynn and Michael E Huge. Public opinion. The international encyclopedia of communication, 2008.
2. Vincent Price. On the public aspects of opinion: Linking levels of analysis in public opinion research. Communication research, 15(6):659–679, 1988.
3. J. Haidt and J. Graham, “When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize,” Social Justice Research, vol. 20, no. 1, pp. 98–116, 2007.
4. T. Islam, R. Zhang, and D. Goldwasser, “Analysis of climate campaigns on social media using bayesian model averaging,” in Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, ser. AIES ’23, Montréal, QC, Canada: Association for Computing Machinery, 2023, pp. 15–25, isbn: 9798400702310.
5. Tunazzina Islam and Dan Goldwasser. Understanding covid-19 vaccine campaign on facebook using minimal supervision. In 2022 IEEE International Conference on Big Data (BigData), pp. 585–595. IEEE, 2022.
6. Tunazzina Islam and Dan Goldwasser. Discovering Latent Themes in Social Media Messaging: A Machine-in-the-Loop Approach Integrating LLMs. arXiv preprint arXiv:2403.10707, 2024.
7. Mar´ıa Leonor Pacheco*, Tunazzina Islam*, Monal Mahajan, Andrey Shor, Ming Yin, Lyle Ungar, and Dan Goldwasser. A holistic framework for analyzing the covid-19 vaccine debate. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5821–5839, 2022a. (* equal contribution)
8. Maria Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin, and Dan Goldwasser. Interactively uncovering latent arguments in social media platforms: A case study on the covid-19 vaccine debate. In Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances), pp. 94–111, 2022b.
9. Maria Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin, and Dan Goldwasser. Interactive concept learning for uncovering latent themes in large text collections. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), Findings of the Association for Computational Linguistics: ACL 2023, pp. 5059–5080, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.313. URL https://aclanthology.org/2023.findings-acl.313.

References

10. Loni Hagen, Ashley Fox, Heather O'Leary, DeAndre Dyson, Kimberly Walker, Cecile A Lengacher, and Raquel Hernandez. The role of influential actors in fostering the polarized covid-19 vaccine discourse on twitter: Mixed methods of machine learning and inductive coding. Jmir Infodemiology, 2(1):e34231, 2022.
11. Thu T. Nguyen, Shaniece Criss, Eli K. Michaels, Rebekah I. Cross, Jackson S. Michaels, Pallavi Dwivedi, Dina Huang, Erica Hsu, Krishay Mukhija, Leah H. Nguyen, Isha Yardi, Amani M. Allen, Quynh C. Nguyen, and Gilbert C. Gee. Progress and push-back: How the killings of ahmaud arbery, breonna taylor, and george floyd impacted public discourse on race and racism on twitter. SSM - Population Health, 15:100922, 2021. ISSN 2352-8273. doi: https://doi.org/10.1016/j.ssmph.2021.100922. URL https://www.sciencedirect. com/science/article/pii/S235282732100197X.
12. Marc Esteve Del Valle, Rimmert Sijtsma, Hanne Stegeman, and Rosa Borge. Online deliberation and the public sphere: Developing a coding manual to assess deliberation in twitter political networks. Javnost-The Public, 27(3):211–229, 2020.

Thank You!!!

Code and Dataset: https://github.com/tunazislam/latent-themes-llms/tree/main

Slide: https://tunazislam.github.io/files/LatentThemesLLM.pdf

Questions

Tunazzina Islam, Ph.D.

Department of Computer Science,

Purdue University, West Lafayette, IN.

Email: **islam32@purdue.edu**

https://tunazislam.github.io/

@Tunaz_Islam

PURDUE UNIVERSITY

Backup Slides

Research Questions (RQs)

- **RQ1**: Can LLMs determine if two given texts, without prior knowledge of existing codes, are discussing the same topic?

Research Questions (RQs)

- **RQ1**: Can LLMs determine if two given texts, without prior knowledge of existing codes, are discussing the same topic?
- **RQ2**: If provided with a definition of a theme, can LLMs successfully categorize other texts under that specific theme?

Coherency Checking

1. Here are some key things to know about used motor oil:

💧 200 million gallons of oil are dumped into U.S. waterways, sewers, and landfills by people changing their own motor oil.

☣ Used motor oil is considered hazardous waste in California.

♻️ Used motor oil is recyclable! Every 1 gallon of used oil recovered can produce 2.5 quarts of re-refined oil.

🚗 Oil never wears out—it just gets dirty. Most cars can go more than 5,000 miles between oil changes.

2. Check out more interesting facts about used motor oil and ways to prevent oil pollution at https://bit.ly/3wFZue1.
The oil and gas industry in Eddy County is the economic driver for our state. I fully support the oil and gas industry and am proud of the environmental standards these great companies impose on themselves.

3. The EPA will hold three virtual listening sessions to take public input on the Agency's upcoming regulations for the Oil and Natural Gas Industry. What you cannot see at an oil & gas well can hurt you. Local Front Range residents are visiting oil and gas sites to learn how invisible methane gas and other health-harming pollution are threatening their communities.

4. Biodiesel (also known as biofuel) is a gallon-for-gallon substitute for petroleum-based fuels, which have a higher carbon intensity. By 2030, it's estimated that biodiesel will displace 529 million gallons of heating oil!

5. Biodiesel is nontoxic, biodegradable renewable liquid fuel. The biofuel that's blended with heating oil to create Bioheat fuel is generally produced by agricultural byproducts, including used cooling oil, animal fats, inedible corn oil, soybean oil and canola oil. This puts excess oil and fats to good use!

Supports motor oil

Against oil & gas industry

Supports biofuel

Coherency Checking: *Incoherent* Cluster

1. Here are some key things to know about used motor oil:

💧 200 million gallons of oil are dumped into U.S. waterways, sewers, and landfills by people changing their own motor oil.

☣ Used motor oil is considered hazardous waste in California.

♻️ Used motor oil is recyclable! Every 1 gallon of used oil recovered can produce 2.5 quarts of re-refined oil.

🚗 Oil never wears out—it just gets dirty. Most cars can go more than 5,000 miles between oil changes.

2. Check out more interesting facts about used motor oil and ways to prevent oil pollution at https://bit.ly/3wFZue1.
The oil and gas industry in Eddy County is the economic driver for our state. I fully support the oil and gas industry and am proud of the environmental standards these great companies impose on themselves.

3. The EPA will hold three virtual listening sessions to take public input on the Agency's upcoming regulations for the Oil and Natural Gas Industry. What you cannot see at an oil & gas well can hurt you. Local Front Range residents are visiting oil and gas sites to learn how invisible methane gas and other health-harming pollution are threatening their communities.

4. Biodiesel (also known as biofuel) is a gallon-for-gallon substitute for petroleum-based fuels, which have a higher carbon intensity. By 2030, it's estimated that biodiesel will displace 529 million gallons of heating oil!

5. Biodiesel is nontoxic, biodegradable renewable liquid fuel. The biofuel that's blended with heating oil to create Bioheat fuel is generally produced by agricultural byproducts, including used cooling oil, animal fats, inedible corn oil, soybean oil and canola oil. This puts excess oil and fats to good use!

Supports motor oil

Against oil & gas industry

Supports biofuel

LLMs categorize this cluster as **incoherent** cluster.

Merging Checking

Summary of cluster 1: #PlantingChallenge aims to plant trees in California to combat climate change, while another initiative promotes the planting of "super trees" with superior carbon storage capabilities. Oxford is considering a Tree Planting and Urban Forestry program to maintain its aging tree canopy, emphasizing the multiple benefits of trees. Criticism is directed at the Obama Presidential Center for cutting down over 300 mature trees, underlining the importance of preserving older trees for carbon storage. Additionally, a service offers to plant trees worldwide for a small monthly fee to help mitigate climate change and replenish forests.

Summary of cluster 18: The Tongass National Forest in Alaska, a crucial carbon sink and habitat, is at the center of environmental advocacy efforts. Accusations against Central Maine Power (CMP) suggest misleading claims about renewable energy development, while campaigns emphasize the Tongass's role in storing 44% of carbon absorbed by U.S. national forests and its significance for clean air, water, and climate. Calls to action focus on reinstating the Roadless Rule to protect the forest from logging and development, with the U.S. Forest Service planning to reinforce these protections. This highlights the forest's importance for salmon populations, local communities, and climate resilience, urging public involvement in conservation efforts.

- **Embed** the summaries using **SBERT**.
- Compute the **cosine similarity** between the embedding summaries of cluster pairs.

Merging Checking: Merged

Summary of cluster 1: #PlantingChallenge aims to plant trees in California to combat climate change, while another initiative promotes the planting of "super trees" with superior carbon storage capabilities. Oxford is considering a Tree Planting and Urban Forestry program to maintain its aging tree canopy, emphasizing the multiple benefits of trees. Criticism is directed at the Obama Presidential Center for cutting down over 300 mature trees, underlining the importance of preserving older trees for carbon storage. Additionally, a service offers to plant trees worldwide for a small monthly fee to help mitigate climate change and replenish forests.

Summary of cluster 18: The Tongass National Forest in Alaska, a crucial carbon sink and habitat, is at the center of environmental advocacy efforts. Accusations against Central Maine Power (CMP) suggest misleading claims about renewable energy development, while campaigns emphasize the Tongass's role in storing 44% of carbon absorbed by U.S. national forests and its significance for clean air, water, and climate. Calls to action focus on reinstating the Roadless Rule to protect the forest from logging and development, with the U.S. Forest Service planning to reinforce these protections. This highlights the forest's importance for salmon populations, local communities, and climate resilience, urging public involvement in conservation efforts.

Argument: Take initiatives to plant new trees and protect existing forests to save our planet.

Clusters are **merged** (cosine similarity ≥ 0.6).

Results: Coverage

- **Climate**
  - Pre-existing themes:
    - Coverage: 17.5% ads.
  - **After Iter1**:
    - Coverage: 40.5% ads.
  - **After Iter2**:
    - Coverage: **42.5%** ads.

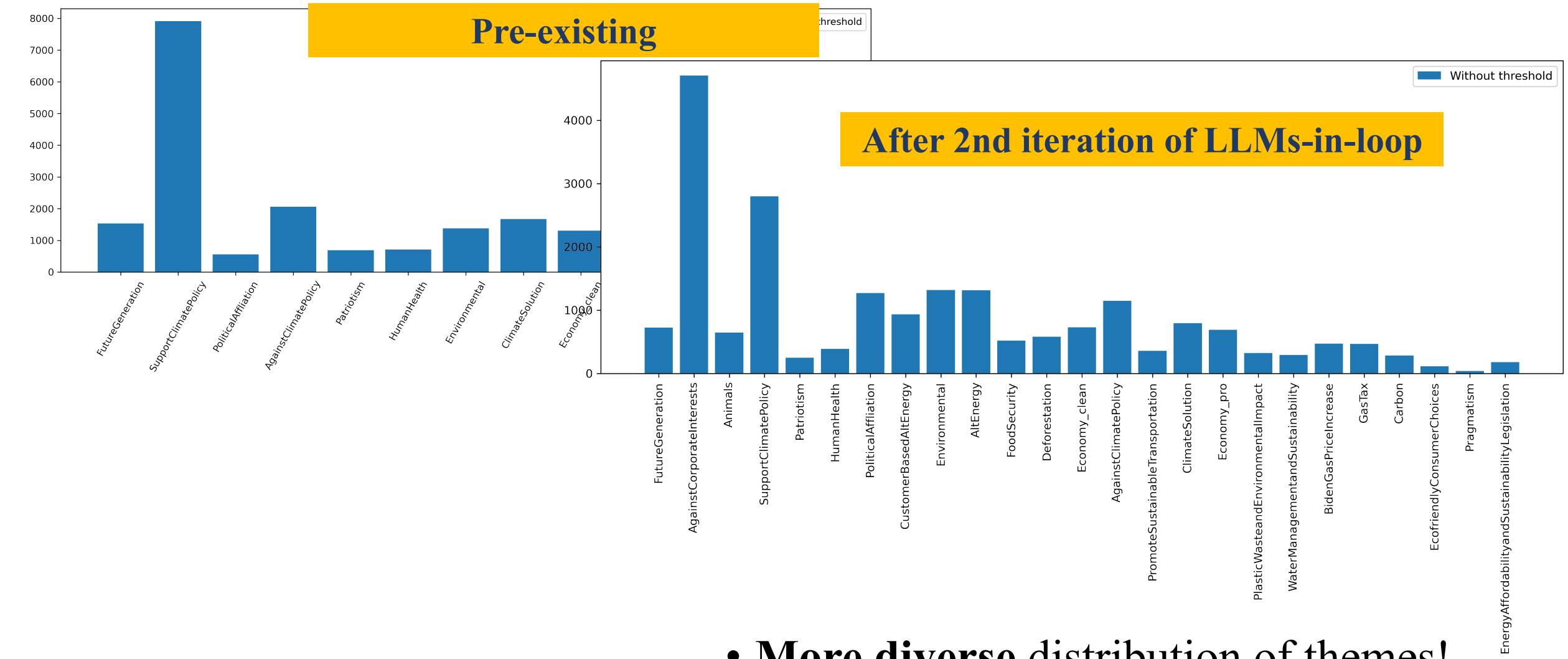
CASE STUDY	METHOD	NUM. THEMES	NUM. COVERED ADS			
			THR < 0.6	THR < 0.5	THR < 0.4	THR < 0.3
Climate	Pre-existing	13	14652	9725	3731	558
	+After Iter1	20	18702	14583	8646	2944
	+After Iter2	**25**	**18988**	**15052**	**9079**	**3180**
COVID-19	Pre-existing	15	7889	6426	3480	771
	+After Iter1	20	8852	7627	4737	1302
	+After Iter2	**23**	**9092**	**7898**	**5038**	**1590**

Results: Coverage

- **Climate**
  - Pre-existing themes:
    - Coverage: 17.5% ads.
  - **After Iter1**:
    - Coverage: 40.5% ads.
  - **After Iter2**:
    - Coverage: **42.5%** ads.
- **COVID-19**
  - Pre-existing themes:
    - Coverage: 35.08% ads.
  - **After Iter1**:
    - Coverage: 47.75% ads.
  - **After Iter2**:
    - Coverage: **50.79%** ads.

CASE STUDY	METHOD	NUM. THEMES	NUM. COVERED ADS			
			THR < 0.6	THR < 0.5	THR < 0.4	THR < 0.3
Climate	Pre-existing	13	14652	9725	3731	558
	+After Iter1	20	18702	14583	8646	2944
	+After Iter2	**25**	**18988**	**15052**	**9079**	**3180**
COVID-19	Pre-existing	15	7889	6426	3480	771
	+After Iter1	20	8852	7627	4737	1302
	+After Iter2	**23**	**9092**	**7898**	**5038**	**1590**

Is *Machine-in-the-Loop* Helpful?



- **More diverse** distribution of themes!