2025 Project HAYSTACK

Highlighting Action in Your State Through AI-Curated Knowledge

Abstract

Climate Cabinet Education has successfully completed a phase one trial for Project HAYSTACK (*Highlighting Action in Your State Through AI-Curated Knowledge*), which analyzed the ability of AI to identify, track, and target the highest-impact state legislation on climate. With hundreds of thousands of bills, determining how to identify what legislation is most important to their areas of interest is a key obstacle for voters, advocates and even elected officials. Our trial indicated that an open-sourced large language model (LLM) could effectively and accurately automate and scale the tracking of state legislation, when guided by specific prompts and knowledge from policy experts.

If funded, the development of these tools and practices would level the playing field — enabling lawmakers, advocates, and the public to understand, participate, and respond to policies being considered by lawmakers.

Introduction

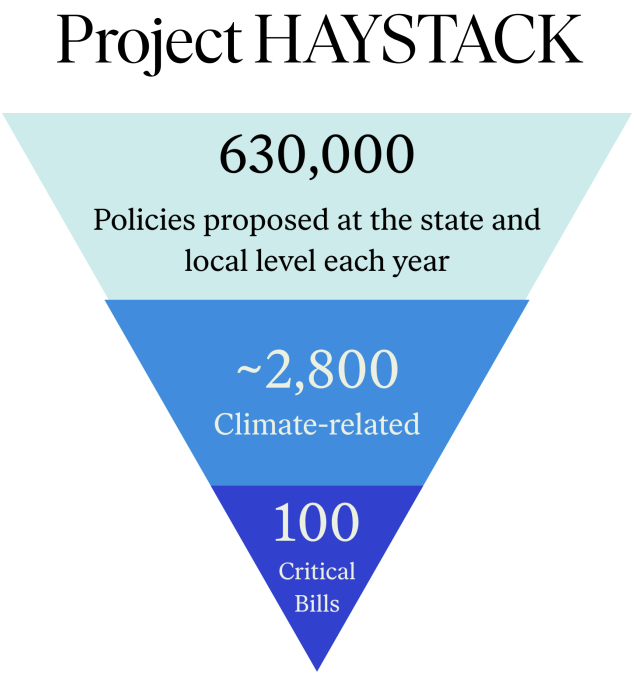
State and local governments play critical roles in U.S. policymaking, especially with current federal instability. The impacts of shifting sands in Congress and the White House have been felt acutely for pro-climate policies, as seen by the rollback of IRA funds. However, there is still a path forward. From housing and transportation policy to solar and wind energy, state and local governments have incredible tools at their disposal to create a climate-safe future and expand the pro-climate job market.

But given the high volume of legislation that gets introduced, it is difficult for anyone to comprehensively understand what is happening in state and local policymaking without intensively investing in a robust lobbying operation.

State and local governments introduce far more bills than Congress. Based on available studies, we estimate that over 630,000 bills and ordinances are introduced each year across the country.[1] Given our fractured media landscape, the decline of local news, and lack of transparency from many legislative bodies, there is often very low visibility about what is happening. Locally-based advocacy groups, voters, and even lawmakers themselves can struggle to know the status and

[1] This estimate was reached by considering Lexis Nexis' collection of 160,000 pieces of legislation collected from 1,000 municipalities nationwide. With a total of over 80,000 municipal and local-level bodies in the US, we feel confident expanding that to 500,000 pieces of municipal-level legislation nationwide, in addition to 130,000 bills introduced at the state level.

details of proposed legislation in their state. There have even been news stories about lawmakers passing legislation by accident.



This lack of visibility poses a significant barrier to civic engagement, smart policymaking, and effective advocacy. Historically, contending with this large volume of information requires resources, infrastructure, and time that most pro-climate organizations and lawmakers have only limited amounts of. While organizations representing polluting interests like the Koch funded Goldwater Institute in Arizona can have a revenue of over $30 million in 2023, this solution is not scalable for pro-climate organizations and does not encourage a healthy civic democracy.

AI is a perfect tool to overcome the overwhelming volume of bills without having to go toe-to-toe with anti-climate organizations in terms of funding in every state. An LLM can digest and parse massive amounts of information to identify key legislation and trends. This can then enable people inside and outside of state and local governments to follow the process accurately and make smarter, more strategic choices.

To unlock this future, in 2024, Climate Cabinet Education launched Project HAYSTACK. In our first phase of research, we ran an initial trial to help develop a first-in-class tool and approach that could unlock state policy through AI. This trial tested whether an existing, open-source LLM could effectively sift through legislation.

Our initial findings were promising. When paired with an expert and proprietary data, AI was able to effectively identify important bills on specific topics. However, more initial model refinement is required. Based on these results, we identified 5 further research opportunities to explore how these tools can be used for supercharging climate policy.

Methodology & Key Results

3 tests were run to analyze the capabilities of LLMs in legislative analysis and tracking - in other words, to assess a LLM's ability to find climate "needles" in the overwhelming haystack of the US policy landscape.

Test 1: Identify Topic-Specific Bills — The first test, presented this spring at San Francisco Climate Week, looked at whether we could use an LLM to improve on a traditional "keyword" search by using a more complicated sentence structure for more targeted results.

Test 2: Identify Climate Relevance – The second test assessed an LLM's ability to identify relevance. We gave an LLM 1,756 active bills in the Arizona State Legislature and asked the LLM to rate them on a scale from 1 (Not Relevant) to 4 (Highly Relevant). We then compared these results to a policy expert's assessment.

Test 3: Recommend a "support" or "oppose" stance. The final test looked at an LLM's ability to measure a bill's impact on climate. To do so, we asked the LLM to evaluate a list of bills that policy experts had already classified by climate impact – both positive and negative. Without knowing the policy experts' position, we asked the LLM to score the legislation and report whether it thought pro-climate advocates should support or oppose each bill and how impactful each bill would be.

For all tests, the following tools were used:

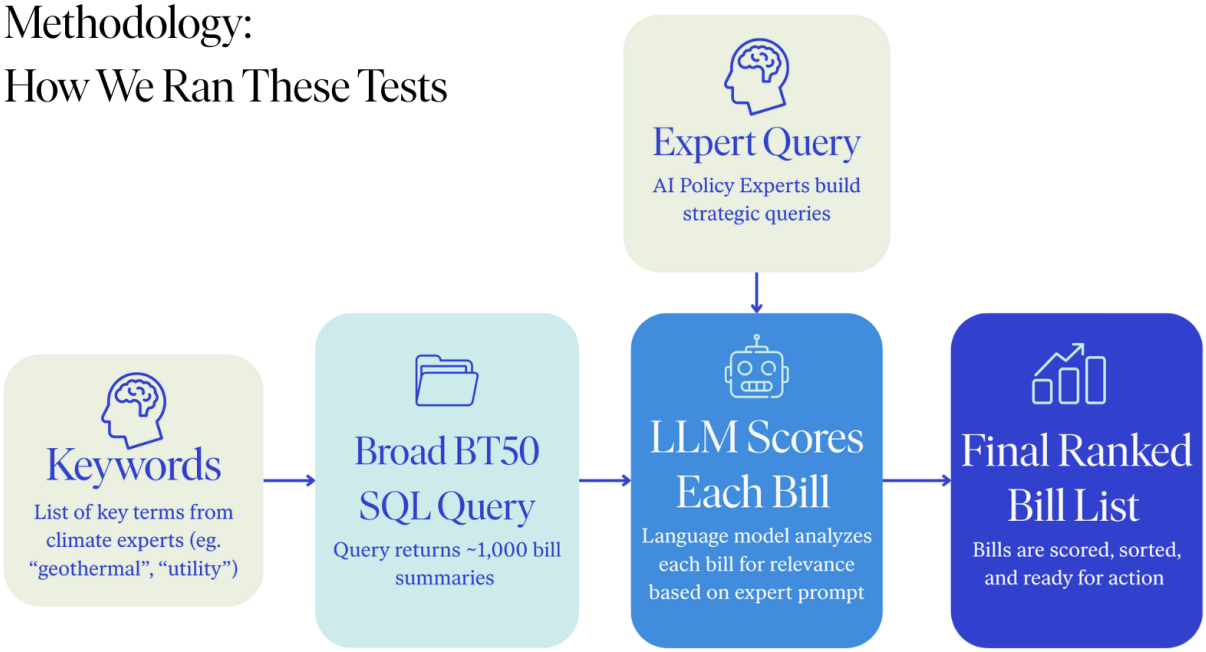
- **BillTrack50** for bill metadata retrieval and filtering. Bill metadata included bill name, activity history, summary, and AI summary provided by BillTrack 50.
- **LLaMA 3.1 405B**, an open-source large language model, for analysis and scoring. LLaMA 3.1 405B open-source has lower reasoning capacity than GPT-4, but fewer constraints and more flexibility for prompt experimentation—well suited for early-stage R&D on legislative classification.
- **Expert-curated control sets** to assess performance across tests.

Each test followed a four-step process:

1. **Download bill metadata** from BillTrack50 using keyword and state-based filters.
2. **Design an expert query** that clearly defines what we are looking for, including framing, scope, and key criteria.
3. **Prompt** the LLM to rate each bill and provide a short justification. For the first two tests, we directed the LLM to use a relevance scoring system of: "1-Not Applicable", "2-Tangential", "3-Directly Related", and "4-Highly Applicable". For the third test, when testing whether a pro-climate organization would support or oppose the bill, we asked the LLM to identify a bill as "Support" or "Oppose" on a scale of 1-3 for each category as a positional score.
4. **Compare LLM results to expert-established benchmarks** to identify alignment and

gaps.

Methodology: How We Ran These Tests



Test 1: Identify Topic-Specific Bills

Background: Traditionally, identifying legislation most relevant to a specific issue area has required substantial amounts of time and labor. The best way to narrow the pool of bills up to now has been keyword searches, where a user searches for something like "climate" or "solar." While these searches are effective at narrowing out bills that don't mention the words at hand, they often return numerous bills that mention the keyword but hold little relevance to the user's topic of interest. For instance a user might keyword search for "solar," seeking legislation about investing in solar for public buildings, and may instead receive dozens of bills that reference other things like zoning regulations or earmarks for community organizations that happen to have "solar" in their name. To separate the wheat from the chaff, the user would need to spend substantial time manually investigating each bill.

Question: Can the LLM narrow the results of a broad keyword search to just the most relevant bills on a given topic?

Conclusion: Yes, especially for simply-described policies.

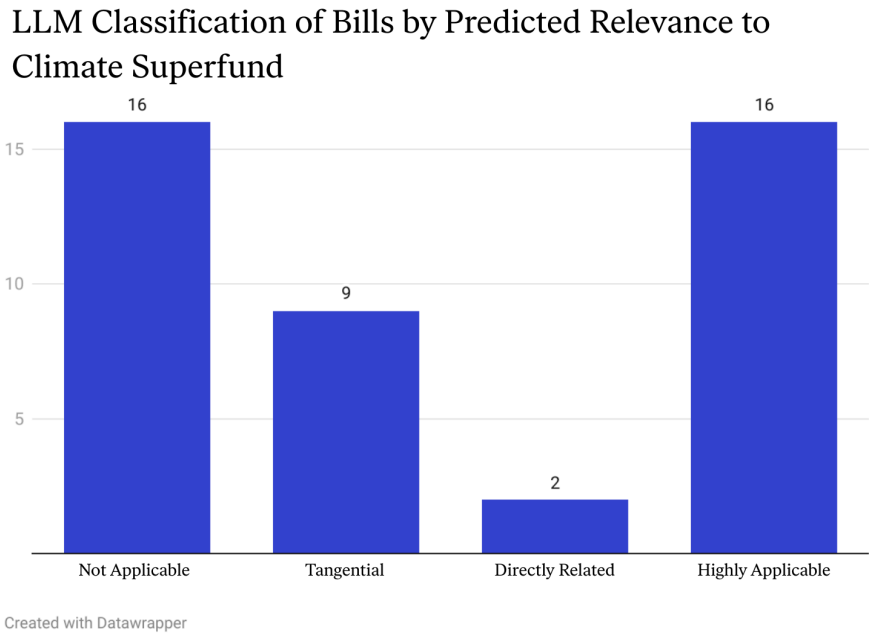
Method:

- Built a sample list using a BillTrack50 SQL query consisting of bills containing the words "climate" or "superfund".

- Wrote a query that clearly and simply defined the policy area that we were looking to identify:
    - We asked the LLM to evaluate the relevance of each bill to "the creation or operation of a polluter-pays fund for climate mitigation."
    - We additionally asked it to score bills from "1-Not Applicable" to "4-Highly Applicable."
- Compared the query results to NCEL's bill list of 23 Climate Superfund bills this session to confirm the accuracy of the LLM's relevance scores.

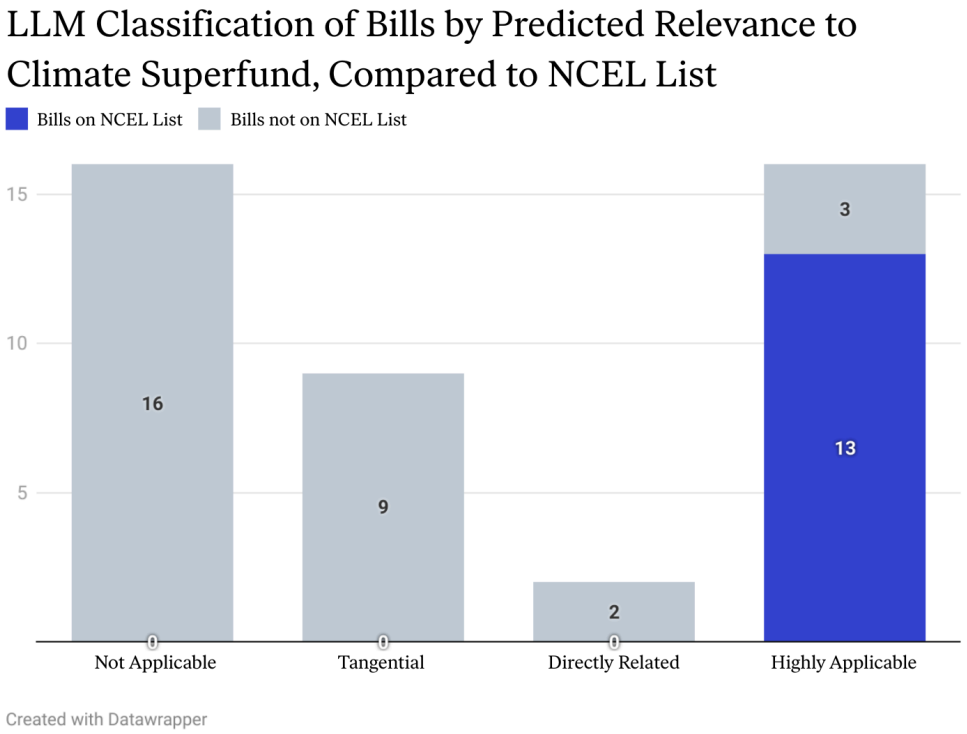
Results:
The LLM ranked 43 bills from "1-Not Applicable" to "4-Highly Applicable."

Overall, 37 % were flagged as "4-Highly Applicable," and 37 % were flagged as ""1-Not Applicable," as seen in the chart below:



Next, we compared the LLM's relevance scores to NCEL's human-developed list of top priority climate superfund bills. If the LLM was accurately evaluating the relevancy of the bills, we would expect to see it classify every single one of NCEL's expert-identified bills as "Highly Applicable" to the issue.

This is exactly what happened. We found that for **all 13 bills** which NCEL had labeled a top priority, the LLM also identified them as "Highly Applicable" – a very promising result.



Of the 13 bills that appeared on both the NCEL bill list and our original list of bills, the LLM correctly identified all 13 of them as "Highly Applicable" and did not miss any "Highly Applicable" bills within the set given to the LLM. It additionally categorized 3 more bills as "Highly Applicable" that NCEL had not.

For every bill that a trusted policy expert had identified as relevant to the creation of a climate superfund, the LLM also identified that bill as relevant.

The model identified 3 bills that were not flagged by NCEL's list that were potentially relevant to Climate Superfund policy (see appendix).

Takeaways:

- When bills were appropriately pre-filtered and the task was tightly scoped, the LLM could reliably classify them with high precision.
- A clearly defined policy objective (i.e., does this bill establish or maintain a polluter-funded climate fund?) simplified interpretation and improved results.
- The LLM correctly identified all partner-priority bills in the input set and contributed 3 new plausible candidates.
- High alignment suggests strong potential for scaling this approach to other narrowly defined climate policy categories.
- The BillTrack50 climate superfund SQL query missed more than half of NCEL's list. This indicates that LLM accuracy is bottlenecked by upstream filtering, indicating we either need to grow our keyword list or run these queries over a more general list of bills.

Test 2: Identify Climate Relevance

The second test assessed a LLM's ability to identify relevance to climate. We gave a LLM 1,756 active bills in the Arizona State Legislature, and asked the LLM to rate them on a scale from 1 (Not Relevant) to 4 (Highly Relevant). Then, we compared these results to a policy expert's assessment.

Question: Can an LLM accurately identify high-relevance bills from a large list of active legislation?

Conclusion: The LLM identified many high-relevance bills, but a few critical bills slipped through the cracks. Overall, scoring was not consistent.

Method:

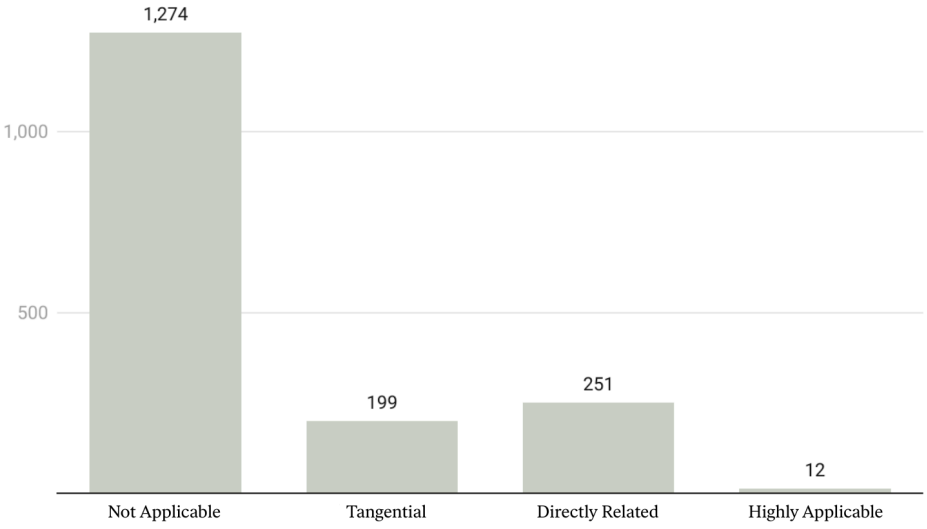
- Generated a list of 1,756 bills that were active in the Arizona state legislature.
- Wrote two LLM queries:
    - The first asked the LLM to identify if the bills in the list of 1,756 were "relevan[t] to state-level climate policy", given a bill summary provided by BillTrack50.
    - The second query asked the LLM to give each bill a "relevance score" ranging from "1-Not Applicable" to "4-Highly Applicable."
- Compared the Climate-Relevant query results to feedback from a policy expert.

Results:

The LLM evaluated that of the 1,756 bills it analyzed, 1,274 of the bills were "Not Applicable" to climate and less than 1% of bills were "Highly Applicable" to climate.

The below chart illustrates how the the LLM categorized the 1,756 bills active in Arizona:

LLM Classification of 1,756 Active Arizona Bills by Predicted Climate Relevance



The LLM classified the vast majority of bills as "Not Applicable" to climate with only 12 classified as "Highly Applicable."

The LLM was accurate for 100% of the "highly applicable" bills. However, it put several climate-critical bills in the "Not Applicable" bucket.

Highly applicable bills ranged from energy to agriculture. From proposed wind farm regulations (HB2223) to a proposal prohibiting government funds from supporting advocacy to reduce meat consumption (SB1195), the LLM picked 12 bills in the "Highly Applicable" category that were definitively related to climate policy. Overall, topics like meat consumption and wind energy were easy for an LLM to identify as climate relevant.

The LLM missed climate-related bills that were oriented more towards government process. For example, the LLM did not identify a bill that proposed restrictive oversight on state agency rulemaking (HB2632).

Takeaways:

- With query refinement, model improvements, or few-shot inclusion of example bills that may seem to not be climate-related, we could likely improve this metric.
- The model would benefit from **stronger priors** — e.g., including examples of known priority bills in few-shot prompt formats, or tuning against our partner and internal lists.
- Ranking thresholds may be salvageable if used for triage (e.g., focusing manual review on Level 3–4 first) but not as standalone filters.

Test 3: Recommend a "support" or "oppose" stance.

Question: Can a LLM accurately recommend a support or oppose stance for climate legislation?

Conclusion: We saw 68.8% Accuracy, with high false confidence.

Method:

- Wrote a query that asked the LLM to score a list of 61 climate bills created by policy experts. The LLM was asked to score these bills in 2 ways: by *category* and by *intensity*.
    - Category scores were: `"support"`, `"oppose"`, `"neutral"`, or `"inconclusive"`
    - Intensity scores indicated how strongly the LLM rated the category: from 0 (neutral/inconclusive) to 3 (strongly support/oppose).
        - For example, if a bill appears to be clearly and extremely anti-climate, the LLM was directed to score it as a "Oppose; 3." If a bill seems related to a climate initiative, but not a large improvement, it was directed to score it as "Support; 1."
- Compared the LLM generated scores to those manually provided by policy experts. The policy experts classified 20 out of 61 bills as 'support', and 41 out of 61 bills as 'oppose'.

Results:
The policy experts and the LLM agreed 68.8% of the time. Notably, the model **incorrectly labeled 15 "oppose" bills as "support"**, often with high confidence. This reflects a pattern of **over-optimism.** The LLM was too generous in interpreting the true intent of the bills, particularly when bill language was vague, technical, or indirectly framed. For example, all of the LLM-ranked 'neutral' bills were ranked by policy experts as 'Oppose'.

	LLM Predicted Oppose	**LLM Predicted Neutral**	**LLM Predicted Support**
Actual Oppose	22	4	15
Actual Neutral	0	0	0
Actual Support	0	0	20

More granular scoring (e.g., predicting levels of support/oppose from -3 to +3) did not improve clarity. In fact, when the model was wrong, it was often **confidently wrong.** For 11 of the bills that policy experts deemed a level 1 'Oppose', the LLM predicted the bills to be a level 2 'Support' (See appendix for full breakdown of the granular scores). This suggests a lack of internal calibration, as the model was too optimistic about the climate impacts of legislation.

Takeaways:

- While the LLM *never misclassified a bill in support of climate action as oppositional*, it **frequently misclassified oppositional bills as supportive** — especially when opposition language was subtle or technical.
- **False positives for support** are concentrated among bills that used misleading or generic climate language.
    - For example, the model struggled to identify problematic utility and corporation commission bills, among others.
- This confirms that **topic-based summaries lack sufficient signal** for opposition identification. Without added sponsor or coalition context, the LLM cannot reliably infer legislative intent that runs counter to climate priorities.
- While earlier results indicated that LLM's are good at **topic assignment,** they struggle with assigning **value** to a bill. The nuance of judging how a policy will harm or help a jurisdiction left the LLM with an overly rosy understanding of bills being proposed.

Results and Next Steps

Future Research Opportunities for AI and Climate Policy: Based on this research, we identified **5 further research opportunities** to explore how these tools can be used for supercharging climate policy.

First, we can improve results with better prompts.

- Problem: These tests found that LLMs require significantly more prompting to identify climate-relevant legislation.
- Potential Next Steps:
    - Work with policy experts to build an AI-enabled keyword annotation system for all legislation that hits state legislatures.
        - Use specific, well-defined LLM prompts to investigate bills. This yields an opportunity for more accurate and detailed bill keyword assignment without having to rely on SQL queries, which do not reliably surface all related bills.
    - Utilize Retrieval-Augmented Generation to improve our ability to associate and identify similar bills across states.
        - For example, similarity search of bill summaries can create additional associations between bills across states, beyond simple keyword associations.

Second, we can improve results by increasing LLM model specificity.

- Problem: LLMs don't always define "climate" or "climate policy" accurately in the abstract.
- Potential Next Steps:
    - Create a domain-specific LLM to improve model accuracy and competency in our specific areas of interest. A domain-specific LLM is an AI model that has been

trained or fine-tuned to perform better on tasks within a narrow subject area—such as law, medicine, climate policy, or clean energy.
    - Generalist LLMs, such as general ChatGPT chats or Claude, draw from the entire internet – so they aren't as specifically tailored towards climate policy goals.
    - "Trained" models can be used to make results faster, less expensive, and more accurate because they only draw from pre-defined sources.
    - A first step could use "semi-trained" approaches, like context-specific chats on ChatGPT.
    - After that, exploring self-hosted models could yield more accurate results.

Third, we can add more "context" data around policies to test if this improves output.

- Explore additional opportunities to leverage AI to understand state legislation such as by inputting bill sponsorship and legislator voting data.

Fourth, we should explore new applications for LLMs and AI. There are other creative uses of LLMs that we are actively exploring after this initial successful project, including:

- Using video transcripts of town halls to pull out key constituent concerns at a large scale – an initial trial run of this use-case yielded promising results.
- Identifying text similarity between bills across states, and expanding that search to other text, including proposed legislation and oppositional language.
- Parsing through state budgets and other omnibus bills to trace climate action at an even more granular level.
- Tracing iterations of legislation, alongside sponsorships and voting records, to track how bills become law.

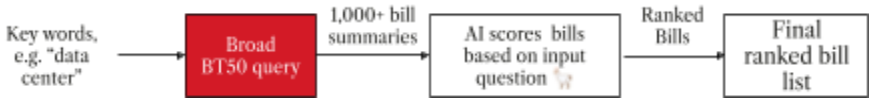
Lastly, LLMs are changing rapidly.

- Problem: Further research and iterative testing is required to understand new LLM capabilities as they come to market.
- Potential Next Steps:
    - Test different AI models, including vector similarity search.
    - Test different prompt approaches and layering.
    - Re-run each test iteratively as AI evolves – models are getting better every day.

Now that we have seen promising results from our early tests in phase one, Climate Cabinet Education will advance this project further. Given the unstable trajectory of the federal government, it has never been more important to supercharge state and local policies.

Appendix

Bill Pipeline Design



Queries

Topic	BT50 keywords	AI prompt
Climate superfund (Topic 1)	Climate superfund	You are analyzing legislation to determine its relevance to the creation or maintenance of a climate superfund — i.e., a fund that polluters must pay into that is used to fund climate mitigation and adaptation projects. Please read the following bill text and respond with a JSON object containing only the following fields: 1. applicable: true or false — Is the bill related in any meaningful way to a climate superfund? 2. level_of_applicability: 1 to 4 — Rate the degree of relevance to a climate superfund (1 = not at all, 2 = tangential, 3 = directly applicable, 4 = highly focused on this issue). 3. explanation: A brief rationale (1-2 sentences) explaining why you chose that applicability and score.
Arizona 2025 Bills (Topic 2, 3)	climate, transportation, energy, utility, justice, budget, approval, land, construction	You are analyzing legislation to determine its relevance to state-level climate policy, especially policies that are important for energy resources, building efficient transportation systems, insurance as it relates to climate risk, utilities, water resources, land use, and limiting greenhouse gas emissions in any way. Please read the following bill text and respond with a JSON object containing only the following fields: 1. applicable: true or false — Is the bill related in any meaningful way to critical climate policy issues? 2. level_of_applicability: 1 to 4 — Rate the degree of relevance to critical climate policy issues (1 = not at all, 2 = tangential, 3 = directly applicable, 4 = highly focused on this issue). 3. key_issues: A list of climate-relevant keywords that apply to this bill.
Climate Priority Bills (Topic 3)	Climate Policy expert list of important, active climate-relevant bills.	You are analyzing legislation to determine whether a pro-climate organization should support or oppose this bill. Please read the following bill summary and provide a JSON object with the following fields. Do not include any other text in your response. Note that we have limited resources, so rank support or oppose carefully and consider the scope and effects of each bill. Keep an eye out for tricky language that may be used to mislead the reader - we care about preventing climate change and ensuring a just transition to a sustainable future. We prioritize long-term solutions that will have the most impact on reducing emissions and mitigating climate change. 1. support_or_oppose: "support", "oppose", "neutral", "inconclusive"

		2. level_of_support_or_oppose: 0, 1, 2, or 3 3. explanation_of_support_or_oppose: A brief explanation of why the pro-climate organization should support or oppose the bill

TEST 1:

Bills identified as 'Climate Superfund' that were not identified by NCEL's partner list:

- Companion bills HB1088/SB882 in MD:
  - "Coal Transportation Fee and Fossil Fuel Mitigation Fund (Coal Dust Cleanup and Asthma Remediation Act)".
- NY A01474
  - 'Relates to the climate change adaptation cost recovery program and requirements for climate change adaptive infrastructure projects'

TEST 2:

Full Query Overview

LLM Ranking	Description	Count
Level 1	Not relevant	1292
Level 2	Tangentially relevant	201
Level 3	Directly relevant	251
Level 4	Highly focused and relevant	12
Total		1756

LLM Level of Bill Relevancy (Not relevant to Highly relevant) Results

AI score of bill relevancy to climate	Expert-ranked high-priority bills	Bills not on expert list	Total bills at relevancy level	% of bills on expert list
Not Applicable	2	1273	1274	0.16%
Tangential	2	197	199	1.01%
Directly Related	7	243	251	2.88%
Highly Applicable	2	10	12	16.67%

TEST 3: LLM vs. Policy Expert Bill Scoring

LLM Prediction vs. Actual Bill Support (+3) / Oppose (-3)								
	LLM prediction of whether a pro-climate organization would support (+3) or oppose (-3) a bill							
Policy experts determination of whether a pro-climate organization would support (+3) or oppose (-3) a bill		**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
	3	0	0	0	0	0	0	0
	2	0	0	0	0	0	4	5
	1	0	0	0	0	0	3	6
	0	0	0	0	0	0	1	2
	-1	3	12	0	4	0	**11**	1
	-2	4	3	0	0	0	1	0
	-3	0	0	0	0	0	1	0

Note the 11 bills predicted as a '2 - Support' that a policy expert ranked as a '-1 - Oppose'.