Democratizing Government Discourse: Building an AI-Powered Robust Summarization and Bias Detection System for Congressional Hearings

Karthik Sreedhar, Anusha Sonthalia, Diego Rivas Lazala

Abstract

Congressional hearings are a foundational mechanism for public policy development, yet their dense, technical language and volume render them inaccessible to most citizens. This limits public understanding of key issues such as internet regulation, privacy, and cybersecurity—areas with significant civic impact. In this project, we introduce a web-based system designed to democratize access to congressional discourse by leveraging state-of-the-art natural language processing (NLP) techniques. Our tool supports three modes of input—URLs, PDFs, and raw text—and enables users to generate structured, policy-relevant summaries or identify linguistic bias in hearing transcripts.

The summarization module uses few-shot prompting with GPT-4 to produce concise, six-part summaries that include the hearing title, committee and date information, goal, key points, outcome, and a policy connection section. Benchmarking using ROUGE scores showed that GPT-based summarization outperforms extractive and traditional abstractive models in readability and domain specificity. In parallel, the bias detection module applies prompt-engineered LLMs to classify statements into categories such as ideological, factual, group, or adversarial bias. The system incorporates chain-of-thought prompting and user feedback collection to support transparent and participatory evaluation.

Together, these modules demonstrate how large language models can enhance civic accessibility and institutional transparency. While our current implementation focuses on internet policy, the architecture is broadly extensible. Future work will involve user studies to evaluate effectiveness and expansion into higher-level bias analysis. This system serves as a step towards more interpretable, accountable, and user-centered tools for democratic engagement.

1 Introduction

Congressional hearings are one of the most important public mechanisms through which lawmakers deliberate, propose, and refine legislative action. They represent the formal interface between policy experts, elected representatives, and the public record. Despite their centrality to democratic governance, the content of these hearings remains largely inaccessible to the general public. Official transcripts are long, complex, and written in legalistic or technical language that is difficult to interpret without substantial context or domain knowledge. For hearings related to internet and network policy—topics that directly affect issues like broadband access, online privacy, net neutrality, and cybersecurity—this information gap poses a significant barrier to civic participation. The inability of citizens to easily understand these discussions limits their ability to advocate, vote, or organize in response to new legislation. As such, there is a pressing need for tools that can make congressional discourse more interpretable and actionable.

Recent advances in natural language processing (NLP), particularly the emergence of large language models (LLMs), provide a promising path toward solving this challenge. LLMs like GPT-3 and GPT-4 have demonstrated the ability to perform a variety of high-level language tasks—ranging from summarization to question answering—with little to no fine-tuning when properly prompted (Goyal et al., 2023; Basyal and Sanghvi, 2023). These models can condense complex language into more accessible summaries and identify linguistic features such as sentiment or framing. Building on these capabilities, we designed an interactive, web-based system that allows users to upload or paste transcripts of congressional hearings and receive structured, policy-relevant summaries. The system supports three types of input—direct URLs (such as from congress.gov), PDF uploads, and raw text—and routes the content through a backend GPT-based summarization module.

This summarization module uses a few-shot prompting strategy to guide GPT-4 toward producing summaries in a consistent and readable six-part format. Each summary includes the hearing title, committee and date information, the stated legislative goal, a list of key points raised, the hearing outcome, and a concluding section that connects the discussion to broader internet or network policy. These components were designed to strike a balance between clarity and completeness, offering both narrative and analytical insight into the hearing's content. The choice of GPT-4 was informed by preliminary benchmarking using ROUGE scores (Lin, 2004), which demonstrated that the model outperformed traditional extractive and supervised abstractive approaches like BART and PEGASUS on both accuracy and readability (Lewis et al., 2020; Zhang et al., 2020). The few-shot approach further ensured output stability and structural uniformity, even across hearings with vastly different policy themes.

In parallel, the system includes a bias detection module that identifies potential forms of bias in the hearing transcript. Bias detection plays a crucial role in preserving the integrity of democratic processes and public discourse. In an era where information flows rapidly through various channels, the presence of bias—whether intentional or inadvertent—can significantly influence public perception and policy decisions. Unchecked biases in media and political communication can erode trust in institutions and deepen societal divides.

Current approaches to bias detection encompass a range of methodologies, from manual content analysis to advanced computational techniques. Traditional methods involve human reviewers assessing content for bias, which, while thorough, are time-consuming and may lack scalability (Spinde et al., 2021). To address these limitations, researchers have developed automated systems leveraging natural language processing and machine learning algorithms to identify and classify bias in textual data (Spinde et al., 2021; Rodrigo-Ginés et al., 2024; Hamborg et al., 2019).

Although most automated bias detection efforts focus on news media, applying these techniques to congressional hearings introduces new complexities. Legislative transcripts are structured differently than journalistic texts and are driven more by rhetorical debate than narrative reporting. This shift in discourse style calls for tailored detection strategies capable of parsing political framing, partisan signaling, and institutional language.

This module draws on a mix of prompt templates and rule-based patterns derived from multidisciplinary research. It is designed to classify statements within categories of bias and specific sub-types including racial bias, misinformation, or party bias. This functionality allows users to critically assess how language may influence the framing of an issue, offering a deeper interpretive layer beyond simple content summary.

By combining these two modules, our system offers a proof-of-concept for how state-of-the-art NLP can be used to increase transparency and interpretability in the democratic process. The web interface is simple and intuitive, but it is backed by a powerful architecture that leverages cutting-edge language models in a structured, policy-aware framework. While the current implementation focuses specifically on internet and network policy, the methodology is broadly extensible to other issue areas and transcript corpora. We envision this project as a small but important step toward building tools that help close the gap between the complexity of government and the public's ability to understand and act on it. In the future, we aim to put our system in front of real users

2 Related Work

2.1 Textual Summarization Approaches

Text summarization is a fundamental natural language processing (NLP) task aimed at generating concise and informative summaries of larger texts. Broadly, summarization approaches are classified into two categories: extractive and abstractive methods. Extractive summarization selects and reorders subsets of sentences from the original document, whereas abstractive summarization generates new sentences that paraphrase the source text using learned language representations.

2.1.1 Extractive Approaches

Extractive summarization techniques identify and select salient sentences directly from the input document. Traditional statistical methods include TF-IDF-based sentence scoring and graph-based algorithms like LexRank and TextRank, which model sentence similarity as edges in a graph and rank importance via eigenvector centrality (Erkan and

Radev, 2004). More recent approaches incorporate pretrained language models such as BERT, which capture semantic similarity and contextual relevance to improve extraction quality (Liu and Lapata, 2019). These methods often perform well on factual precision and can retain grammatical correctness, since the extracted sentences are unmodified. However, they may struggle with coherence, flow, and abstraction, and often produce redundant or overly long summaries.

2.1.2 Abstractive Approaches

Abstractive summarization methods aim to produce novel textual output that paraphrases and compresses the source content. Early approaches relied on encoder-decoder architectures trained on large corpora of document-summary pairs (See et al., 2017). More recently, transformer-based models like BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020) have achieved great performance through pretraining on large-scale datasets using denoising or gap-sentence objectives. In parallel, large language models (LLMs) such as GPT-3 and GPT-4 have demonstrated strong zero-shot and few-shot summarization capabilities, often outperforming extractive and even fine-tuned abstractive models across several domains (Goyal et al., 2023; Zhang et al., 2023; Basyal and Sanghvi, 2023). These models can be prompted with contextually rich instructions and examples, enabling fine control over summary structure and length. However, they may occasionally hallucinate facts or deviate from the original intent, making robust evaluation and verification essential.

2.1.3 Evaluating Summarization Approaches

Evaluation of summarization quality can be done using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric (Lin, 2004), a suite of measures that compare automatically generated summaries against human-written reference summaries. ROUGE-N scores evaluate the overlap of n-grams, with ROUGE-1 and ROUGE-2 commonly assessing unigrams and bigrams, respectively. ROUGE-L, another variant, measures the longest common subsequence between generated and reference texts. Higher ROUGE scores generally indicate greater fidelity to human-written content, though they may not fully capture readability, coherence, or factuality.

2.2 Bias Detection Approaches

Bias detection has long played a crucial role in journalism and democratic discourse, originally relying on manual analyses grounded in cognitive science, linguistics, and political theory. With the advent of artificial intelligence and machine learning, the field has evolved to include automated systems capable of processing large volumes of text and identifying subtle patterns of ideological, linguistic, or structural bias. In particular, large language models (LLMs) have expanded our ability to analyze human discourse with greater contextual awareness and semantic depth.

Recent systematic reviews (Spinde et al., 2023; Rodrigo-Ginés et al., 2024; Hamborg et al., 2019) have surveyed approaches to automated media bias detection and emphasized the utility of interdisciplinary methods. These studies highlight a trend toward increasingly sophisticated models, particularly those based on transformer architectures, and emphasize the need for high-quality annotated datasets to ensure reliable detection. They also underscore the challenge of modeling bias across differing political and cultural contexts and the importance of integrating insights from political science and media theory.

Media bias detection methodologies typically commence by establishing a clear definition of bias and a corresponding categorization framework. Subsequently, they employ specific techniques to identify and analyze bias within media content. The following sections delineate these methodologies, providing insights into their approaches and applications.

2.2.1 Manual Approaches

Manual approaches to media bias rely on human judgment, often involving trained analysts or structured rating frameworks. For example, AllSides uses a combination of expert editorial review, blind bias surveys, and third-party data to rate political leanings of media outlets (all). While these methods are transparent and context-aware, they are resource-intensive and may suffer from reader bias, limiting their scalability and reliability.

2.2.2 Aggregator Approaches

Aggregator-based platforms compile bias ratings from multiple sources to produce composite media bias scores. Ground News, for instance, incorporates data from AllSides, Ad Fontes Media, and Media Bias Fact Check to classify bias across outlets and articles (gro). Aggregator models enhance robustness by integrating perspectives, but depend on the methodologies and update frequencies of contributing organizations.

2.2.3 Traditional ML Approaches

Early machine learning methods for bias detection framed the task as a supervised classification problem. Algorithms such as support vector machines, logistic regression, and Naive Bayes were trained on labeled datasets, using handcrafted features like n-gram frequency, sentiment scores, and part-of-speech patterns (Hamborg et al., 2019; Rodrigo-Ginés et al., 2024). These approaches offered interpretability and decent performance on structured inputs, but struggled with contextual nuance and were highly dependent on feature engineering and training data quality. For something as nuanced and difficult to define as bias, feature engineering for these models was a very hard-to-solve problem.

2.2.4 Transformer-Based Approaches

Modern transformer-based models have significantly advanced the field of bias detection by capturing long-range dependencies and contextual relationships in text. Models like BERT and RoBERTa, when fine-tuned on political or ideological classification tasks, have shown strong results in detecting subtle rhetorical cues (Spinde et al., 2023). Multi-task learning and domain-specific pretraining on media bias datasets have further improved accuracy. However, these models are resource-intensive and require large annotated corpora, which at the moment is only available for bias detection the media space. As previously mentioned, for congressional hearings the problem is slightly different.

2.2.5 Prompt Engineering Approaches

Prompt-based methods leverage LLMs like GPT-3 and GPT-4 by crafting tailored prompts that guide the model to identify and explain bias. These approaches are flexible and do not require extensive retraining, making them suitable for quickly adapting to new domains or document types. For example, the University of Pennsylvania's Media Bias Detector employs prompt engineering combined with human-in-the-loop feedback to evaluate ideological framing in media texts (upe). While promising, prompt-based systems require careful calibration and risk inheriting or amplifying biases from the base LLM itself.

3 System Design

Our initial system design included a full transcription pipeline that would automatically convert video or audio recordings of congressional hearings into text for downstream summarization and analysis. However, during early development, we discovered that high-quality, human-edited transcripts of nearly all recent U.S. congressional hearings are publicly available through the Library of Congress at congress.gov. These transcripts include speaker labels, timestamps, and are formatted consistently, making them highly suitable for automated text processing. As a result, we shifted our focus from building a transcription model to developing tools that analyze and summarize these transcripts using natural language processing techniques.

3.1 Summarization Module

To generate policy-relevant summaries of congressional hearing transcripts, we developed a summarization module based on large language model (LLM) prompting. During experimentation, we compared the outputs of multiple summarization strategies—both extractive and abstractive—using a Jupyter notebook-based evaluation pipeline. We used the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric (Lin, 2004) to assess n-gram overlap between model outputs and reference summaries, with ROUGE-1, ROUGE-2, and ROUGE-L scores serving as metrics.

Initial tests with extractive approaches and supervised abstractive models (e.g., BART, PEGASUS) yielded summaries that were often disjointed or insufficiently tailored to legislative content. By contrast, OpenAI's GPT models—specifically GPT-4—performed best under a few-shot prompting paradigm. This approach involved crafting custom prompts with in-context examples to steer the model toward consistently formatted, comprehensive summaries. We observed that GPT-based summarization produced higher ROUGE scores and clearer, more domain-appropriate results compared to baseline methods, aligning with recent comparative findings in the literature (Basyal and Sanghvi, 2023; Goyal et al., 2023).

The final design of the summarization module integrates a backend GPT interface, which accepts either raw text, a transcript URL, or a PDF file. The user-facing interface allows for transcript submission, after which the backend invokes the LLM to generate a structured summary. Each summary adheres to a standardized format consisting of the following components:

- **Title**: The official title of the hearing.
- **Committee, Date, and Meeting Info**: Identifies the congressional committee, the date, and whether the hearing was held in person, virtually, or in a hybrid format.
- **Hearing Goal**: A short explanation of the main purpose or focus of the hearing.
- **Key Points**: A bulleted list of 2–4 major takeaways, including proposed bills, themes raised by witnesses, or debated issues.
- **Hearing Outcome**: Summarizes the result of the hearing—e.g., consensus reached, actions proposed, or legislation advanced.
- **Connection to Internet Policy**: A closing section explaining how the hearing relates to broader internet and network policy, such as regulation, privacy, digital equity, or cybersecurity - relating our project to this course.

This structure was designed to balance clarity, comprehensiveness, and policy relevance. By standardizing summary output, the module helps users—especially non-technical stakeholders—quickly understand the legislative implications of complex governmental discourse. In addition, the consistency of the format supports future integration with bias detection and topic classification modules in the broader system.

3.2 Bias Detection Module

3.2.1 Framework Definition

In adapting bias detection methodologies discussed to the context of congressional hearings, we recognized the need to tailor existing frameworks to address the unique characteristics of legislative discourse. Unlike traditional media, congressional hearings involve structured dialogues among policymakers, necessitating a specialized approach to identify and categorize bias effectively.

Our bias detection framework is informed by comprehensive analyses of media bias detection methodologies (Spinde et al., 2023; Rodrigo-Ginés et al., 2024; Hamborg et al., 2019). Drawing from these studies, we identified four primary categories of bias pertinent to congressional hearings:

- **Group Bias:** Bias based on demographic characteristics. Types in this category include racial, gender, class, etc.
- **Ideological Bias:** Bias stemming from alignment to belief systems. Types in this category include political, religious, etc.
- **Factual Bias:** Bias created by misrepresenting information. Types in this category include omission, misinformation, opinions-as-fact, etc.
- **Adversarial Bias:** Bias from confrontational tactics in debate. Types in this category include loaded questions, character attacks, etc.

This categorization facilitates a nuanced analysis of bias within the unique setting of congressional hearings, where the interplay of political ideologies and rhetorical strategies is prominent.

3.2.2 Detection Method

Given the scarcity of annotated datasets specific to congressional hearings, we adopted a prompt engineering approach utilizing Large Language Models. This strategy aligns with methodologies employed in projects like the University of Pennsylvania's Media Bias Detector (upe), which harnesses LLMs for bias analysis.

Our approach involved crafting prompts that guide the LLM to identify and classify instances of bias within hearing transcripts. We experimented with varying levels of specificity in these prompts:

- **Minimal Guidance:** Initial prompts provided broad definitions of bias categories, allowing the LLM to utilize its latent knowledge.
- **Simple Definitions:** Incorporating the detailed descriptions of categories as defined above aiming to give structure to the LLM's responses.
- **Enhanced Definitions:** Detailed descriptions of categories as defined above and examples for each bias type aiming to improve consistency in the LLM's responses.

This iterative process revealed a trade-off: while detailed prompts enhanced consistency, they occasionally constrained the LLM's ability to detect nuanced or context-specific biases. Conversely, broader prompts allowed for greater flexibility but resulted in less structured outputs.

Finally, to enhance the transparency and reliability of the LLM's assessments, we integrated chain-of-thought prompting. This technique encourages the model to articulate its reasoning process, providing justifications for its bias classifications. Such an approach not only aids in understanding the model's decision-making but also aligns with best practices in prompt engineering.

3.2.3 System Scope

In developing our bias detection system for congressional hearings, we have deliberately scoped the analysis to the statement level. This decision allows us to build the foundations of a system that could be scaled to provide biases in larger scopes.

In traditional media bias detection, analyses often span multiple levels:

- **Statement Level:** Assessing individual sentences or claims for
- **Article Level:** Evaluating the overall bias in a news article
- **Author Level:** Analyzing an author's body of work to identify consistent bias patterns
- **Media Outlet Level:** Determining the overarching bias tendencies of a media organization based on aggregated content and content placement

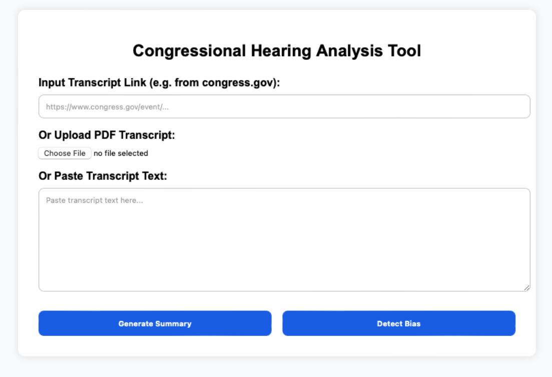
Similarly, in the context of political discourse, bias can be examined at various levels:

- **Statement Level:** Focusing on individual phrases or questions within a single speech or debate.
- **Speech Level:** Analyzing entire speeches to discern overarching themes.
- **Hearing Level:** Assessing the overall bias present in an entire hearing session, including biases for speaker time
- **Speaker Level:** Assessing a particular individual's discourse over multiple hearings for recurring patterns or trends
- **Party Level:** Evaluating collective discourse of a political party to identify systemic bias

Our initial focus on the statement level allows for precise identification of bias instances, creating the building blocks for a larger system. While our current solution is tailored for statement-level bias detection, the architecture is designed with scalability in mind. Upon successful validation of our results and accumulation of sufficient data, the system can be extended to analyze bias at higher levels. This modular approach ensures that our system remains adaptable, allowing for progressive enhancement as more data becomes available and as the models mature.

4 System Walkthrough

To enable user-friendly access to our analysis tools, we built a web-based interface that allows users to input congressional hearing transcripts in one of three formats: a direct URL (typically from congress.gov), a PDF file, or raw pasted text. As shown in Figure 1, the user can then choose between two backend analysis modules: the summarization module or the bias detection module. This flexible input system ensures compatibility with real-world data formats used by government repositories, while the modular backend architecture supports extensibility and future development.



Congressional Hearing Analysis Tool

Input Transcript Link (e.g. from congress.gov):

https://www.congress.gov/event/...

Or Upload PDF Transcript:

Choose File no file selected

Or Paste Transcript Text:

Paste transcript text here...

Figure 1: Main landing page of the Congressional Hearing Analysis Tool. Users may input hearing transcripts as a link, PDF, or raw text, then select summarization or bias detection.

Our system's code is publicly available at https://github.com/karthiksreedhar/step-final. A demonstration video showcasing the system's functionality and full usage is included in the repository's README for reference.

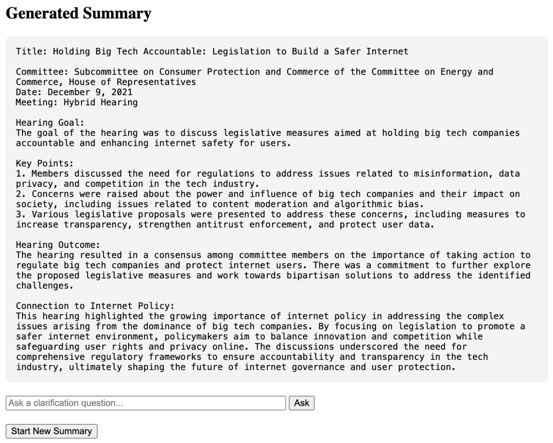
4.1 Summarization Module

Once a user reaches the main interface of the system (Figure 1), they are presented with three ways to input a congressional hearing transcript: by pasting a direct URL (such as from `congress.gov`), uploading a local PDF file, or entering raw text into a freeform input field. This allows flexibility in how users interact with the tool depending on the format of hearing data they have available.

After submitting a transcript and selecting the "Generate Summary" button, the user is redirected to a new page where the system outputs a structured summary of the hearing. This summary is organized into six clearly defined sections: **Title**, **Committee, Date, and Meeting Info**, **Hearing Goal**, **Key Points**, **Hearing Outcome**, and **Connection to Internet Policy**. The format ensures readability and consistency across different hearings, even when topics and content vary widely.

Figure 2 shows a representative summary generated by the system for the hearing titled *"Holding Big Tech Accountable: Legislation to Build a Safer Internet"*, conducted by the House Subcommittee on Consumer Protection and Commerce on December 9, 2021. The hearing transcript is available at congress.gov. This output demonstrates the system's ability to distill complex legislative discourse into digestible components.

Beneath the summary, users are also provided with optional tools to further engage with the content. They can ask clarification questions in natural language (e.g., "What were the concerns about data privacy?"), or return to the main page to analyze a different hearing. This interactive layer enhances user engagement while keeping the system lightweight and focused.



Generated Summary

```
Title: Holding Big Tech Accountable: Legislation to Build a Safer Internet

Committee: Subcommittee on Consumer Protection and Commerce of the Committee on Energy and
Commerce, House of Representatives
Date: December 9, 2021
Meeting: Hybrid Hearing

Hearing Goal:
The goal of the hearing was to discuss legislative measures aimed at holding big tech companies
accountable and enhancing internet safety for users.

Key Points:
1. Members discussed the need for regulations to address issues related to misinformation, data
privacy, and competition in the tech industry.
2. Concerns were raised about the power and influence of big tech companies and their impact on
society, including issues related to content moderation and algorithmic bias.
3. Various legislative proposals were presented to address these concerns, including measures to
increase transparency, strengthen antitrust enforcement, and protect user data.

Hearing Outcome:
The hearing resulted in a consensus among committee members on the importance of taking action to
regulate big tech companies and protect internet users. There was a commitment to further explore
the proposed legislative measures and work towards bipartisan solutions to address the identified
challenges.

Connection to Internet Policy:
This hearing highlighted the growing importance of internet policy in addressing the complex
issues arising from the dominance of big tech companies. By focusing on legislation to promote a
safer internet environment, policymakers aim to balance innovation and competition while
safeguarding user rights and privacy online. The discussions underscored the need for
comprehensive regulatory frameworks to ensure accountability and transparency in the tech
industry, ultimately shaping the future of internet governance and user protection.
```

Figure 2: Example output of the Summarization Module for the hearing "Holding Big Tech Accountable: Legislation to Build a Safer Internet" (Dec 9, 2021).

4.2 Bias Detection Module

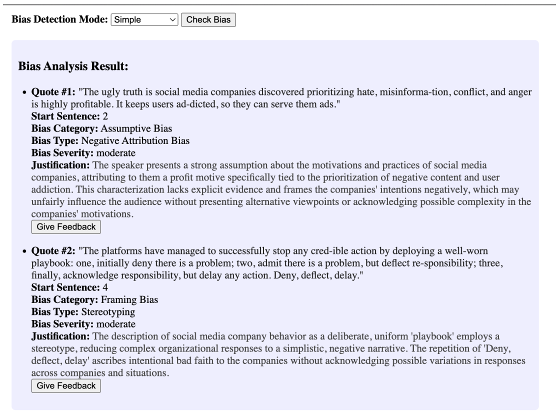
Once a user navigates to the main interface of the system (Figure 1), they are presented with flexible options for transcript input: users may upload a PDF of a congressional hearing, or directly copy and paste transcript text into the provided interface. This design accommodates the varied formats in which hearing data may be accessed. After submitting a transcript and selecting the "Detect Bias" button, the user is redirected to a new page where the system outputs the uploaded transcript for selection.

To initiate bias analysis, the user selects a specific portion of the transcript they wish to examine. They are then prompted to choose a bias detection mode—*Simple (minimal guidance)*, *More Context (simple definitions)*, or *Most Context (enhanced definitions)*—as described in Section 3.2.2. These modes correspond to different levels of prompt specificity provided to the underlying LLM, ranging from minimal guidance to detailed, framework-driven definitions. This approach reflects our iterative methodology, balancing the flexibility of broad prompts with the consistency afforded by enhanced definitions.

Upon submission, the system processes the selected text and, within a few seconds, returns a structured analysis. For each identified biased quote, the output includes: the quote itself, its starting sentence, the bias category (as defined in section 3.2.1: Group, Ideological, Factual, or Adversarial), the specific bias type, a severity rating (mild, moderate, or severe), and a justification articulating the model's reasoning. This chain-of-thought prompting not only increases transparency but also aligns with best practices in LLM-based bias detection.

Figure 3 illustrates a typical output, where the system highlights biased statements within the selected transcript segment and provides detailed justifications for each classification. This granular feedback enables users to understand both the presence and nature of bias within legislative discourse.

To help refine the system and collect data for further development, users are invited to provide feedback on each analyzed quote as pictured in Figure 4. The feedback interface allows users to rate the quality of the bias analysis (on a scale from "Excellent" to "Very Poor"), specify their own assessment of bias severity and type, and optionally comment on their rating. This participatory layer is critical for iterative improvement: user feedback is systematically collected and analyzed to identify edge cases and inform future enhancements to the



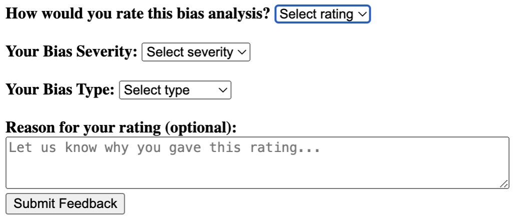
Bias Detection Mode: Simple

Bias Analysis Result:

- **Quote #1:** "The ugly truth is social media companies discovered prioritizing hate, misinforma-tion, conflict, and anger is highly profitable. It keeps users ad-dicted, so they can serve them ads."
  Start Sentence: 2
  Bias Category: Assumptive Bias
  Bias Type: Negative Attribution Bias
  Bias Severity: moderate
  Justification: The speaker presents a strong assumption about the motivations and practices of social media companies, attributing to them a profit motive specifically tied to the prioritization of negative content and user addiction. This characterization lacks explicit evidence and frames the companies' intentions negatively, which may unfairly influence the audience without presenting alternative viewpoints or acknowledging possible complexity in the companies' motivations.
- **Quote #2:** "The platforms have managed to successfully stop any cred-ible action by deploying a well-worn playbook: one, initially deny there is a problem; two, admit there is a problem, but deflect re-sponsibility; three, finally, acknowledge responsibility, but delay any action. Deny, deflect, delay."
  Start Sentence: 4
  Bias Category: Framing Bias
  Bias Type: Stereotyping
  Bias Severity: moderate
  Justification: The description of social media company behavior as a deliberate, uniform 'playbook' employs a stereotype, reducing complex organizational responses to a simplistic, negative narrative. The repetition of 'Deny, deflect, delay' ascribes intentional bad faith to the companies without acknowledging possible variations in responses across companies and situations.

Figure 3: Example output of the Bias Detection Module for the hearing "Holding Big Tech Accountable: Legislation to Build a Safer Internet" (Dec 9, 2021).

bias detection framework.



How would you rate this bias analysis? Select rating

Your Bias Severity: Select severity

Your Bias Type: Select type

Reason for your rating (optional):
Let us know why you gave this rating...

Figure 4: Example of the feedback template after the Bias Detection Module for the hearing "Holding Big Tech Accountable: Legislation to Build a Safer Internet" (Dec 9, 2021).

By grounding our detection methodology in established media bias frameworks (section 3.2.1) and adapting them to the unique context of congressional hearings, our system delivers nuanced, transparent, and user-engaged bias analysis. The integration of prompt engineering strategies (section 3.2.2) ensures that the LLM's assessments are both reliable and adaptable to the complexities of legislative discourse.

5 Discussion

This work represents an important step toward the broader goal of making legislative processes more transparent, interpretable, and accessible to the public. As congressional hearings shape national policy on critical topics such as internet regulation, privacy, cybersecurity, and content moderation, it is vital that both expert and non-expert stakeholders are able to understand and critically engage with these discussions. Our system addresses this need by combining summarization and bias detection into a single platform, demonstrating how large language models and bias-aware NLP techniques can be operationalized to help users navigate complex political discourse (Spinde et al., 2023; Rodrigo-Ginés et al., 2024; Hamborg et al., 2019).

By focusing on internet and network policy, we provide a proof-of-concept for a tool that is both topical and extensible. The summarization module offers a structured and consistent entry point into long and dense transcripts, while the bias detection module provides interpretive insight into the rhetorical framing and ideological undercurrents of the text. Together, these components allow users not just to consume political information, but to question and reflect on how it is communicated.

5.1 Bias Detection as a Participatory Process

The development of our bias detection system for congressional hearings was shaped by two major factors: the unique characteristics of legislative language and the scarcity of annotated datasets for this domain. Unlike journalistic or social media text, congressional transcripts are highly formal, often multi-speaker, and shaped by institutional norms rather than commercial incentives. These traits complicate the application of off-the-shelf bias detection methods (Hamborg et al., 2019).

To address the data scarcity challenge, we adopted a user-centric approach. Our system incorporates mechanisms for users to annotate perceived bias within transcript excerpts. This participatory design serves two purposes. First, it allows us to validate model outputs by comparing them against human judgments. Second, it enables the gradual creation of a domain-specific dataset that can be used to improve bias detection models over time. This design aligns with recent calls for human-in-the-loop frameworks in bias detection (Spinde et al., 2023; Rodrigo-Ginés et al., 2024).

5.2 Toward Deeper Bias Analysis

The architecture of our system is designed for scalability toward higher-level analysis. Currently, we operate at the statement level—highlighting biased sentences or phrases within hearings. As more user-generated annotations are collected, future iterations of the system can expand to include the higher levels of bias including speech, hearing, speaker, and party level bias as discuss in section **??**. This progression will enable more robust insights into the structure and dynamics of political discourse. It also opens the door to quantitative studies on fairness, accountability, and polarization in congressional rhetoric.

With a sufficiently large annotated corpus, we anticipate developing transformer-based models specifically fine-tuned for legislative bias detection. Domain-adapted models have already shown significant gains in specialized fields such as biomedical and legal NLP (Beltagy et al., 2019; Chalkidis et al., 2020), and we expect similar improvements in political discourse analysis.

5.3 Limitations and Future Work

While our system demonstrates strong potential, it has several limitations. First, our evaluation of summary quality and bias detection has been largely internal and qualitative. Future work should include a structured user study to assess how end users perceive the accuracy, clarity, and usefulness of the summarization and bias annotations. Such a study could also help assess the system's effectiveness in informing or altering users' perceptions of legislative content.

Second, the use of large language models raises issues of transparency and fairness. LLMs are known to reflect and sometimes amplify harmful biases found in their training data (Bender et al., 2021; Abid et al., 2021; Dhamala et al., 2021). We use prompt engineering and human review to mitigate this risk - future work could benefit from adversarial testing, detoxification techniques (Welbl et al., 2021), and robust auditing pipelines.

Finally, our current deployment focuses on English-language U.S. federal hearings. Extending the system to state-level governments, multilingual transcripts, or other domains (e.g., healthcare, education) would require retraining or adjustment, but the general architecture remains flexible.

5.4 A Step Toward Civic NLP

In sum, our work contributes to the growing field of civic NLP—tools and methods developed to improve democratic access to language-based political processes. By transforming raw congressional hearing transcripts into structured summaries and bias-aware annotations, we provide users with both digestible content and interpretive depth. As concerns about institutional trust and media polarization persist, tools like ours have the potential to promote transparency, critical engagement, and informed civic participation.

6 Conclusion

In this project, we presented a modular, user-centered system that leverages large language models to enhance public accessibility and interpretability of congressional hearings. By combining a structured summarization module with a bias detection module, the platform enables users to digest complex legislative content and critically evaluate rhetorical framing in political discourse. Our approach prioritizes usability without compromising analytical depth—supporting multiple input formats, delivering consistent outputs, and incorporating participatory feedback mechanisms.

While our current implementation focuses on hearings related to internet and network policy, the architecture is broadly adaptable to other domains. The few-shot prompting design enables scalable summarization, and our bias framework provides a foundation for future analysis at broader levels, including party or speaker-level trends.

Looking ahead, we plan to conduct user studies to evaluate the system's real-world utility, refine bias detection through iterative feedback, and extend our approach to more languages and contexts. Ultimately, we envision this tool as a foundational step in building a broader ecosystem of civic NLP technologies—ones that make government more legible, accountable, and responsive to the people it is supposed to represent.

References

Allsides media bias rating methodology. https://www.allsides.com/media-bias/media-bias-rating-methods. Accessed: 2024-05-16.

Ground news bias rating methodology. https://ground.news/rating-system. Accessed: 2024-05-16.

Media bias detector methodology. https://mediabiasdetector.seas.upenn.edu/methodology/. University of Pennsylvania, Accessed: 2024-05-16.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. *arXiv preprint arXiv:2101.05783*.

Lochan Basyal and Mihir Sanghvi. 2023. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models. *arXiv preprint arXiv:2310.10449*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *EMNLP*, pages 3615–3620.

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

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. *arXiv preprint arXiv:2101.11718*.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Tanya Goyal, Aayush Sharma, Nanyun Peng, and Ani Nenkova. 2023. News summarization and evaluation in the era of gpt-3. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11098–11114.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Timo Spinde, Smi Hinterreiter, Fabian Haak, Terry Ruas, Helge Giese, Norman Meuschke, and Bela Gipp. 2023. The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias. *arXiv preprint arXiv:2312.16148*.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.