

Modeling the political process to forecast the outcomes of hypothetical AI governance proposals

Abstract

International cooperation on AI governance faces a fundamental trust problem: countries like the US and China struggle to assess whether proposed agreements would actually be implemented by their counterpart’s domestic political systems. This uncertainty undermines the credibility of commitments and hinders the establishment of safety regulations needed to prevent catastrophic AI risks. We address this challenge by developing a system that predicts whether AI safety legislation would gain support within a country’s government, enabling both domestic policymakers and international partners to evaluate the political feasibility of proposed regulations. Our approach uses large language models to generate interpretable yes/no questions about legislative bills, then learns legislator-specific perspective representations that capture individual voting patterns on AI policy. We collect and analyze voting records from the 118th and 119th U.S. Congresses (2024-2025), identifying 146 AI safety-related bills. Our model significantly outperforms baseline approaches in forecasting senatorial votes on AI legislation. Additionally, we develop a suite of hypothetical AI governance policies ranging from strict to permissive, using our model to identify political feasibility thresholds—the boundaries between policies likely to pass versus fail. This work provides a concrete tool for improving transparency and trust in international AI governance negotiations.

1 Introduction

One key challenge in AI safety right now is enabling cooperation on governance internationally, especially between the US and China. Each nation is incentivized to continue AI development, because it leads to economic advantage; yet, moving too fast too quickly could lead to loss of control and catastrophe. At some point, the wise path would be to establish common rules and agreement between the countries.

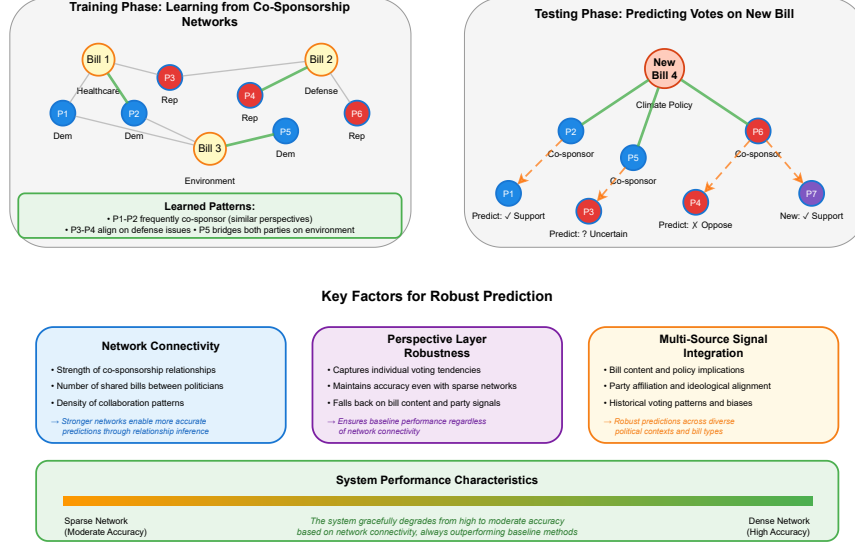


Figure 1: Co-sponsorship network architecture for robust vote prediction. During training (left), the model learns collaboration patterns between politicians across different bills, with green edges indicating strong co-sponsorship relationships. During testing (right), the system predicts votes on new legislation by inferring from known co-sponsors’ network connections (orange dashed arrows). The bottom panel illustrates three key robustness factors: network connectivity strength, perspective layer adaptability, and multi-source signal integration, which together enable graceful degradation from high to moderate accuracy based on network density.

However, there is no common basis of trust in the area of AI. Each country would see that might be in their advantage to push the rules and try to continue to gain an intelligence edge, if doing so would not be discovered. Hence, judging the strength or truth of the other party’s commitments is essential to a lasting and successful agreement. There will also be groups within each country trying to identify the appropriate amount of AI regulation that are likely to succeed in their own domestic environment.

We aim to begin addressing this problem by giving lawmakers in both the US and China, plus any interested third parties, the tools necessary to judge whether proposed AI safety legislation would be supported by a country’s internal system of government. In particular, due to the transparency of the US senatorial system, we begin by modeling each US senator’s approach to

AI legislation. We want to predict possible future responses to new AI safety legislation.

Our contributions are threefold:

- We propose a novel method that learns AI bill representations and legislator-specific voting patterns.
- We demonstrate that our approach significantly outperforms baseline methods on real voting records from the 118th and 119th U.S. Congresses (2024-2025), achieving state-of-the-art prediction accuracy on AI-related legislation.
- We create a set of hypothetical graduated policies that start very strong and grow weaker and weaker, to identify the point at which we predict that members of Congress would approve them. This enables better understanding of the political feasibility of future AI governance proposals.

2 Related Work

AI governance and international cooperation. The challenge of international cooperation on AI development has been extensively studied. Dafoe (2018) outlines key governance challenges, while Armstrong et al. (2016) and Cave and ÓhÉigeartaigh (2018) analyze the dynamics of AI races between nations. Sastry et al. (2024) examines compute governance as a lever for AI regulation, and Gupta et al. (2023) proposes institutional frameworks for international AI governance. Schuett (2023) discusses defensive strategies against AI risks. Our work contributes to this literature by providing tools to assess the domestic political feasibility of proposed international agreements, addressing the trust deficit that undermines cooperation.

Legislative vote prediction. Early work on vote prediction focused on ideal point estimation models such as NOMINATE (Poole and Rosenthal, 1985), which represent legislators and bills in a low-dimensional ideological space. More recent approaches apply machine learning methods including support vector machines (Gerrish and Blei, 2011), random forests (Kornilova and Eidelman, 2018), and neural networks (Eidelman and Kornilova, 2018) to predict votes from bill text and legislator features. Hofmann et al. (2022) apply pre-trained transformers to this task. However, these methods typically focus on general legislative prediction rather than the specific context of AI safety policy, and they do not address the international governance trust problem that motivates our work.

LLMs for social and political modeling. Recent work has explored using large language models to simulate human behavior and decision-making. Argyle et al. (2023) demonstrate that LLMs can simulate human survey responses across demographic groups. Horton (2023) explores LLMs as simulated economic agents, while Fu et al. (2023) and Akata et al. (2023) investigate LLM behavior in negotiation and game-theoretic settings. Spirling (2012) applies text analysis to historical political agreements, and Schrodtt (2014) discusses methodological challenges in quantitative political analysis. Our work extends this line of research by fine-tuning LLMs on legislative sponsorship patterns to predict support for future AI safety policies, with the explicit goal of enabling international cooperation.

3 Method

Our aim is to model the people involved in the political process, to try to estimate how they would respond to potential new AI safety bills in the future. While we want to model the US, China, and EU systems, we restrict our attention to US politicians in this work (see Section 5), specifically US senators, who are empowered to introduce new legislation.

Problem Formulation Let $\mathcal{D} = \{(\mathbf{b}_n, \mathbf{v}_n)\}_{n=1}^N$ denote a dataset of N legislative bills, where $\mathbf{b}_n$ represents the text and metadata of bill n , and $\mathbf{v}_n = \{v_n^l\}_{l=1}^L$ contains the votes of L legislators, with $v_n^l \in \{\text{Oppose}, \text{Support}\}$ representing the vote from legislator l on bill n . Not all legislators vote on all bills, so we introduce a participation indicator $\delta_n^l \in \{0, 1\}$ indicating whether legislator l cast a vote on bill n .

Our goal is to learn a predictive model that, given a bill $\mathbf{b}$ and legislator identity l , can predict the legislator’s vote: $f(\mathbf{b}, l) \rightarrow \{\text{Oppose}, \text{Support}\}$. The key challenge is that legislators have systematically different voting patterns due to party affiliation, ideology, and constituency pressures. Rather than explicitly engineering legislator-specific parameters, we leverage the capacity of large language models to implicitly learn and represent these individual perspectives through fine-tuning on historical voting records.

Our hypothesis is that when provided with sufficient contextual information about both bills and legislators, a fine-tuned LLM can:

1. Understand the political implications of legislative text
2. Learn distinct voting patterns for different legislators
3. Generalize these patterns to predict votes on new legislation

Our architecture consists of three main components: (1) a module that generates targeted yes/no questions about bills from the original bills, (2) a prediction head that aggregates question responses into vote predictions.

Question representation. For each bill n with text $\mathbf{b}_n$, we use a module to create K targeted yes/no questions that capture key political dimensions of the legislation: $\mathbf{q}_n = \{q_n^1, q_n^2, \dots, q_n^K\} = F(\mathbf{b}_n; \mathcal{P})$ where $\mathcal{P}$ is a designed module that extracts sponsors and co-sponsors and their political views corresponding to a specific bill. Each question q_n^k is designed to elicit clear yes/no responses that reflect political positions. We obtain embeddings for each generated question using the same LLM’s representation layer: $\mathbf{e}_n^k = \text{LLM}_{\text{embed}}(q_n^k)$. These question embeddings $\mathbf{e}_n^k \in \mathbb{R}^d$ capture the semantic content of each political dimension.

Political Stance Projection The question embeddings are aggregated and projected through a learned transformation to produce a distribution over K latent political stances. This projection learns interpretable political dimensions that align with the question structure.

Training We train end-to-end using cross-entropy loss over observed votes, where only participating legislators contribute to the loss. During backpropagation, gradients from multiple legislators naturally aggregate at the political stance level, enabling the model to learn representations that account for diverse perspectives while maintaining interpretability.

4 Experiments

4.1 Datasets

We have two main datasets, described below.

AI Safety US Congress We collect a dataset of all bills proposed in the senate in 2024-2025. We then collect all bills related to AI safety. We perform the following:

1. First, we create the question, “Who is in this bill?” followed by a list of names of the sponsor and any co-sponsors for the bill. This dataset is used to fine-tune our model.

2. Second, for each senator mentioned in the bill, we ask whether they would support the bill, constructing a yes/no question pair for each senator. These questions are used to evaluate the success of our method.

We used the GovInfo API gov (2025) to scrape all US senate bills from 2024 and 2025, to the date of our last data collection on Oct 7th, 2025. We performed a search with the `/published` endpoint to retrieve 20,468 bills, then fetched the `packageLink`, `billStatusLink`, and `txtLink`. Our dataset contains a `BILLS` directory for each bill, with `json` files for basic info, all available details, list of members (sponsors) for the bill, and also the full bill text. Our dataset and collection scripts will be made available upon camera-ready.

We combined all this data into a single `json` file, while also recording the `BILLS` identifier used by GovInfo as an ID. We then filtered this dataset in several ways:

1. First, we removed duplicate bills that were presented in different sessions before being passed. We deduplicated based on bill title, and kept only the latest (by timestamp) version of the bill. This resulted in 17,582 bills.
2. Next, we searched for all bills containing the phrase “artificial intelligence”. (Note that we did not account for a possible line break between the words, but this is unlikely to matter since any bill discussing AI mentions it many times.) This gave us 427 AI-related bills.
3. Next, we searched for all bills with the word “risks”. This yielded 1,195 bills.
4. Finally, we generated the intersection to get all AI-related bills that say “risks”, which are likely to be the more safety-oriented bills. This yielded 146 bills.

Gradated New Policies In order to test different potential new AI safety bills, we generated a gradated set of potential policies. We started from a very strict policy that would likely go a long way towards ensuring that superintelligence would not be created, but which is likely far too strict to pass in the senate. Then, we used ChatGPT 5 to generate progressively weaker versions of the bill. The aim was to eventually reach a point where senators (as modelled by our system) would feel comfortable approving the bill.

Table 1: Vote prediction accuracy on US Congress and gradated policies test sets. Gradated policies dataset shows what percentage of senator-bill votes are predicted to be “Yes” across all gradations. Best results in bold.

Method	AI Safety US Congress $\uparrow$	Gradated New Policies $\uparrow$				
		I	II	III	IV	V
GPT4-oss-20B	0.00%	–	–	–	–	–
Ours	62.58%	39.0%	36.0%	32.0%	40.0%	33.0%

We define five levels of gradated policies, from level I (weakest) to level V (strongest). We vary only the names of the bills with some fixed boilerplate content.

4.2 Settings

Compute resources needed:

- One A100 80GB GPU
- One hour of training time
- 30 minutes of test time (with significant portions of bill text)

Because many bills consist of more than 1MB of text, we had to truncate bill text dramatically to fit into the 80GB memory of one GPU. (At one point, our model tried to allocate 144GB of memory at once for a particularly large input.) Fortunately however, bill titles are fairly descriptive, and information about sponsors and co-sponsors was supplied separately via the metadata in our dataset.

4.3 Experimental results

We measured the following results, shown in Table 1. Note: we leverage a Llama 3 judge model to classify our results, to handle cases where the main model outputs a longer answer than just “Yes”/“No”.

We see that the baseline GPT-OSS does not have any information on these congresspeople. The knowledge cutoff for this model is June 2024, so the majority of our bills dataset could not have been included—in fact, even many of the senators would not even have been included in its training data. On the other hand, our fine-tuned model is able to achieve a score of 62.58%

on these questions, which are of the form, “Would senator X support a bill Y with content Z?” (for truncated content Z).

The graduated policies dataset is evaluated by testing each of our policies on the five senators who are most involved in any AI bill. For this dataset, we do not have GPT-OSS results (will include in camera-ready), but expect it to be near zero because of the prior results for this model. Our model shows a reasonable rate of 39% on the weakest gradation (I), and actually a slightly higher rate of 40% on the second-strongest. The toughest policy was still predicted to be successful with at least some senators. Example bill names under each policy are shown in Table 2.

Gradated policy level IV is an anomaly for having such a high percentage of approvals. This may be because many of the generated bills are about financing: freezing capital, insurance, finance registry, etc. It seems that our model believes that these senators are likely to agree to financial control bills.

5 Future Work

In future work, we will aim to model more people with influence over the political process. In the US, this includes people in the administration and tech leaders at AI-related companies. We will also certainly expand our analyses to cover China and the EU as well as the US. Some of these analyses may be ready for a camera-ready version of this paper.

The overall goal is to explore multi-governmental agreements, and to try to model the impact of the personalities involved (and any other relevant factors). We will keep an eye out for opportunities to partner with other research groups to expand the types of analyses we can perform.

6 Conclusion

We introduced a system for predicting legislative support for AI safety policies, addressing a critical challenge in international AI governance: the lack of trust regarding whether proposed agreements would be implemented by domestic political systems. Our approach uses large language models to learn legislator-specific voting patterns on AI-related legislation, enabling prediction of how US senators would respond to new AI safety bills. We collected and analyzed Congress bills from 2024-2025, built models of each senator, and evaluated the system with a graduated set of hypothetical future AI governance policies. We hope our work will provide a concrete approach for improving transparency in international AI governance negotiations.

Table 2: Example titles in the graduated safety policy categories from I (weakest) to V (strongest).

I:	Whistleblower Protection for AI Disclosure Act Financial Disclosure for AI Investments Act AI Model Disclosure Act of 2025
II:	Compute Facility Certification Act Independent Red-Team Evaluation Act Frontier Training License Act
III:	Secure Compute Enclave Standard Act Compute Leakage Penalty Act GPU Serial Registry Act
IV:	Insurance Requirement for Frontier Runs Act AI Finance Registry Act Capital Freeze Authority Act
V:	Frontier AI Prohibition and Capital Control Act Compute Shutdown Emergency Act Frontier AI Moratorium Act

Ethics Statement

Our models could be used to manipulate rather than understand the political process, or could be used against the interests of the originating government. However, we utilize public sources only and there are already many ways to interfere with politics (deepfakes and lobbying, to name a few). Hence, we hope our research will primarily serve its intended purpose as an assistive tool for those concerned with AI safety.

Reproducibility

Code, data, and trained models will be released upon acceptance. All datasets are from public sources. Specifically, we will release:

- **Dataset:** 358MB of combined senate bill data formatted in json
- **Code:** Data collection pipelines, training scripts, evaluation code
- **Benchmarks:** The subsets of data that we used for test and train

Available at: [URL to be added]

References

2025. Govinfo api. <https://api.govinfo.gov/docs/>.

Elif Akata, Lion Schulz, Julian Coda-Forno, and 1 others. 2023. Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, and 1 others. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Stuart Armstrong, Nick Bostrom, and Carl Shulman. 2016. Racing to the precipice: a model of artificial intelligence development. *AI & Society*, 31(2):201–206.

Stephen Cave and Seán S ÓhÉigeartaigh. 2018. An ai race for strategic advantage: Rhetoric and risks. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 36–40.

Allan Dafoe. 2018. Ai governance: A research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford*.

Vladimir Eidelman and Anastassia Kornilova. 2018. Deep learning for legislative roll call prediction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1349.

Yao Fu, Hao Peng, Tushar Khot, and 1 others. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Sean M Gerrish and David M Blei. 2011. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 489–496.

Akash Gupta, Andrew J Lohn, Noam Kolt, and 1 others. 2023. International institutions for advanced ai. *arXiv preprint arXiv:2307.04699*.

- Valentin Hofmann, Janet B Pierrehumbert, and Hinrich Schütze. 2022. Fine-tuning pre-trained transformers for predicting legislative roll-call votes. *arXiv preprint arXiv:2205.08791*.
- John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? *National Bureau of Economic Research Working Paper*.
- Anastassia Kornilova and Vladimir Eidelman. 2018. Party matters: Enhancing legislative embeddings with author attributes for vote prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 510–515.
- Keith T Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 29(2):357–384.
- Girish Sastry and 1 others. 2024. Computing power and the governance of artificial intelligence. *arXiv preprint arXiv:2402.08797*.
- Philip A Schrodtt. 2014. Seven deadly sins of contemporary quantitative political analysis. *Journal of Peace Research*, 51(2):287–300.
- Jonas Schuett. 2023. Three lines of defense against risks from ai. *AI & Society*, pages 1–12.
- Arthur Spirling. 2012. Us treaty making with american indians: Institutional change and relative power, 1784–1911. *American Journal of Political Science*, 56(1):84–97.