

Analysis of State-Level Legislative Process in Enhanced Linguistic and Nationwide Network Contexts

Maryam Davoodi
Purdue University
mdavoodi@purdue.edu

Dan Goldwasser
Purdue University
dgoldwas@purdue.edu

Abstract

State bills have a significant impact on various aspects of society, including health, education, and the economy. Consequently, it is crucial to conduct systematic research on state bills before and after they are enacted to evaluate their benefits and drawbacks, thereby guiding future decision-making. In this work, we developed the first state-level deep learning framework that (1) handles the complex and inconsistent language of policies across US states using generative large language models and (2) decodes legislators' behavior and implications of state policies by establishing a shared nationwide network, enriched with diverse contexts, such as information on interest groups influencing public policy and legislators' courage test results, which reflect their political positions.

1 Introduction

State policies affect our lives in many ways, such as our health, taxes, and rights. State legislatures can protect or limit our freedoms in areas like voting, reproduction, labor, and more. State laws also impact national issues, such as federal courts deferring to states on constitutional matters and federal laws responding to state laws (see Appendix A).

Thus, our goal is to build an NLP/ML framework for understanding state-level legislative process in the US. This can help the public to know how bills affect them, check if legislators vote consistently with their public views, and find out who influence them. While promising NLP/ML studies, e.g., (Feng et al., 2022; Davoodi et al., 2022), predict bill or vote outcomes using bill text and legislator information, they have limitations: most focus on centralized federal-level process, they miss out some state-level contextual information such as influence of local lobbyists, they use low quality and inconsistent state data, they lack support for non-vote prediction tasks, or they often use traditional NLP/ML architectures. Our framework addresses

these limitations and improves the understanding of state legislative process and its implications:

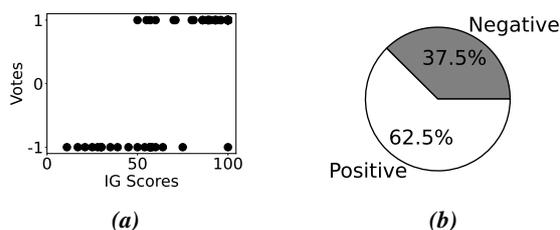


Figure 1: Correlation between (a) Interest group (IG) scores of legislators and their voting on a bill favoring the IG and (b) Legislators' responses to a courage test question and their voting pattern on a related bill.

Normalizing the language of state legislation.

Our framework's first pillar addresses the inconsistency, verbosity, and complexity of state-level legislative data (LegiScan, 2023; NCSL, 2024). State bills vary in language and format for the same topic, making cross-state policy analysis difficult. They are also lengthy and detailed, obscuring the main points and implications of the bills. State bills often require domain knowledge to comprehend, limiting their accessibility for the public. Therefore, we normalize and optimize state legislative data by using Generative LLMs, e.g., OpenAI GPT (OpenAI, 2023). This involves generating consistent summaries of bills, enriching legislator's data, and identifying potential winners/losers of bills. This way, we can better compare similar policies across different states and topics. We evaluate the quality of LLM summaries by comparing them with expert-written ones and explore the feasibility of using open-source LLMs (Appendix C).

Capturing inter- and intra-state contexts. We also aim to understand how state bills relate to other external factors (e.g., cross-state effects). Thus, our framework adheres to two design principles: (1) We study all 50 US states, unlike most recent studies that focus on a few states (Davoodi et al., 2020). (2) We find and include important inter-state and intra-state entities that affect

state policies and were ignored before. Specifically, we include *Interest Groups (IGs)* that lobby legislators for their values and affect the making and passing of bills on issues like health care. IGs assign scores to legislators based on their policy alignment, which can indicate their future voting patterns (Figure 1a). Also, we use *Political Courage Tests (PCTs)* that ask candidates to state their positions on issues related to their session, which can reveal the genuine preferences of legislators who may conceal them for electoral/partisan motives (Figure 1b, Appendix D).

Efficient abstraction and embedding. To operationalize our work, we construct a new text-attributed graph abstraction to represent different entities and relations in state legislative process in a highly normalized linguistic and broadened network contexts (Figure 2). On top of this abstraction, we enable diverse prediction tasks, beyond predicting legislators’ votes, such as predicting legislators’ courage test results and interest group impacts that both have real-world use-cases (Section 6). We argue that such tasks are interdependent and can benefit from joint learning and representation. To efficiently deliver these, we develop a new shared text and graph embedding architecture (Figure 3), designed to handle longer legislative texts and unseen entities more effectively. The first in kind that combines the benefits of an efficient non-generative LLM, OpenAI Ada (OpenAI, 2023), with that of a widely-adopted heterogeneous GNN, RGCN (Schlichtkrull et al., 2018), to jointly generate contextualized text and graph representations for nodes and relations.

Technical contributions: In summary, we present the following technical contributions:

- Modeling state policies in wider intra-/inter-state contexts overlooked by prior studies (Section 3).
- Using LLMs to solve linguistic issues in state legislative data for the first time (Section 4).
- Creating a new dataset and nationwide abstraction of state legislative process and defining two new prediction tasks (Section 5).
- Developing an efficient joint LLM and GNN embedding and prediction architecture, better supporting long texts and unseen nodes (Section 6).
- Showing our architecture’s superior performance over robust text and graph baselines (Section 7).

2 Related Works

We discuss promising studies that have influenced our work (Further discussion in Appendix D).

Bill progression classification. One research direction aims to predict the likelihood of a congress bill passing a certain legislative stage, e.g., clearing a committee, which is a challenge that most bills fail to meet. For instance, Yano et al. 2012 use features related to the bill’s significance and bill’s co-sponsors, to estimate the committee outcome. Eidelman et al. 2018 conduct a similar analysis for state bills and show the benefits of integrating bill text with minimal contextual info. on legislators.

Roll-call (vote) classification. A related research direction is to model how federal legislators vote. Kraft et al. 2016; Patil et al. 2019 capture ideological preferences of US congress members for this task. Karimi et al. 2019; Kornilova et al. 2018 incorporate bills’ text and metadata of the sponsors. Feng et al. 2022; Mou et al. 2021, 2023 build better models for Congress legislators based on their info. on social networks (Wikipedia, Twitter) and expert knowledge websites (Think tanks). While Davoodi et al. 2020, 2022 conduct similar studies for 3 US states but with key limitations (see Section 1).

LLM applications. The evolution of LLMs, like OpenAI’s GPT and Ada, has significantly improved core NLP tasks and enabled new applications. For instance, GPT-3 made accurate data annotation possible despite biases (Ding et al., 2022). GPT-3.5 enhanced this with better summarization of complex user reviews (Bhaskar et al., 2023). GPT-4 takes it further, enabling assessments of second-language proficiency aligned with standards such as the CEFR (Yancey et al., 2023).

3 Broadening Political & Social Contexts

Our goal is to build a model that can better understand legislators’ behavior and the impact of policies within and across states. Thus, we abstract the state legislative process (detailed in Appendix A) using a text-attributed graph and enrich it with diverse political-social contexts, derived from our ablation study (Section 7). In the next section, we apply our generative LLM optimization on the graph to handle the linguistic challenges.

3.1 Nationwide Legislative Graph: Properties

Our proposed legislative graph abstraction, shown in Figure 2, comes with four unique properties: Our model is nationwide, interconnecting all 50 states,

allowing us to compare legislative activities across different states, e.g., abortion regulation (Ballotpedia, 2024). It incorporates multiple types of heterogeneous entities, such as legislators, interest groups, and political courage tests, with varying levels of influence and involvement in the policy process. It contains rich textual information for each node in the graph, such as bills with titles and summaries or interest groups with mission statements. This enables us to extract meaningful insights from the data. Additionally, it captures the relationships among the nodes, such as co-sponsorship, endorsement, or opposition, that reflect the alignment or conflict between them. Next, we describe some key components of our graph.

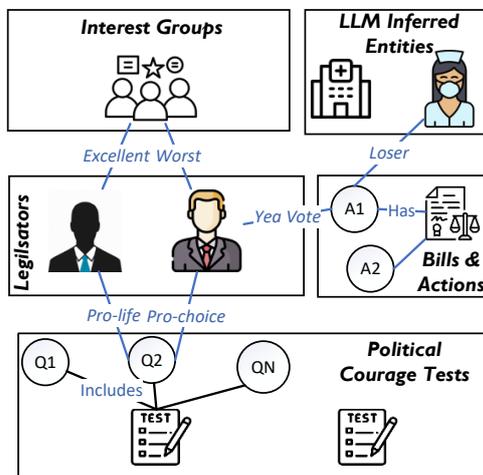


Figure 2: LLM-optimized, contextualized, and nationwide legislative graph for analyzing state-level policies.

3.2 Modeling Legislators & Revisiting Bills

Legislators and bills are the most important components of the legislative process. A state bill has metadata such as title, official synopsis, status, and sponsors, as well as a list of actions that track its progress through different stages (e.g., introduction, committee referral, floor vote). We add both bills and their actions as nodes in the graph and connect them with edges that indicate the temporal order of the actions. We also add edges between legislators and actions to represent their voting behavior and between legislators and bills to model sponsorship.

Adopting bill highlights for normalizing state bills. A major obstacle for legislative data analysis in the US is the inconsistency among different states (LegiScan, 2023; NCSL, 2024). Each state has its own way of writing, naming, summarizing, and categorizing bills, making it hard to

compare and contrast them. Also, we have seen bill titles and synopses can be missing, misleading or vague. Unlike existing works, we propose to use bill highlights or summaries of legislation written by a group of legislative researchers as a better alternative in the graph as well as *ground-truth* to evaluate our LLM generated ones (Section 4). These researchers (VoteSmart, 2023) read the text of the bill and then write consistent summaries that extract the most important information from each bill and present it in a clear way (Examples in Appendix C). Bill highlights are more informative and accurate than bill titles and synopses. Also, they normalize our data across different US states and make it easier to understand the impact and implications of each bill. *Our experiments show bill highlights can boost the performance of two bill classification tasks by up to 3.5 F1 points, compared to the official titles and synopses* (Appendix B.4).

3.3 Incorporating Interest Groups (IGs)

Our legislative graph incorporates the data of interest groups (IGs) that operate at different levels (Figure 2): *within a state, across multiple states, and nationwide*. This allows us to gain insights into state policies, which were overlooked in previous NLP/ML studies at the state level. IGs advocate for causes or groups. They lobby directly or indirectly to sway legislators’ votes. They offer information, expertise, or resources to help legislators. They also influence public opinion through media or petitions to pressure legislators to align with their interests. To capture all these interactions, we represent IGs as nodes in the graph and include a textual description of their activities, drawn from their public mission statement or other sources on the Internet, into these nodes. For example, this is a part of textual info. associated with an IG in Indiana:

Name: "Florida Planned Parenthood PAC", **Description:** The Florida Planned Parenthood PAC is the political arm of Planned Parenthood organizations in Florida and is a separate segregated political action committee, which supports state candidates who support access to safe/legal abortion and family planning health services...

IG relations (scores). We also study how IGs affect legislators through their public ratings of legislators. An IG score is a percentage that interest groups give to legislators based on various factors, e.g., money donations and historical voting alignment (See Appendix B). We collect these scores for all legislators and model them as relations in the graph (Figure 2). IGs publish these scores to influ-

ence legislators’ future actions. We divide scores into four categories: Worst: 0-25 Bad: 26-50%, Good: 51-75%, Excellent: 76-100%; and create corresponding relation types between IGs and legislators in the graph. This is a sample IG score:

"RatingText": "Representative Bob Heaton supported the interests of the Citizens Action Coalition of Indiana 32 percent in 2021.", "rating": "32", "Id": "1062", "timespan": "2021", "ratingName": "Lifetime Positions", "Category": "Agriculture and Food"

3.4 Incorporating Political Courage Tests

To better understand state legislators’ voting behaviour, we need to know their motivations, priorities, and commitments. Political courage tests (PCTs), detailed in Appendix B.3, can reveal these by asking them challenging questions—more effective than their social network information (Appendix D). These questions measure how much legislators respect democracy, human rights, and the rule of law. They also test how they deal with unpopular issues, IGs, and minority rights. Some frequently asked PCT questions are as follows:

- (1) Do you support the prohibition of public funds for organizations that perform abortions?
- (2) Should background checks be required on gun sales between private citizens at gun shows?
- (3) Do you support pro-life or pro-choice legislation?

PCTs do not provide a clear way to compare different legislators or to identify patterns and trends in their behavior. This is where our graph modeling can offer a significant advantage. By representing the PCT and its questions as nodes in a graph, and linking them to the legislators who answer them (Figure 2), we can create a rich and dynamic representation of the political landscape. This allows us to (a) better identify the similarities and differences between legislators based on their answers; (b) easier analyze the relationships and influences among legislators, IGs, and issues; and (c) better predict how legislators might react to new legislation based on their past performance. In this work, our focus is on 2-choice (Yes/No) and 4-choice questions, making up 95% of the tests on average.

4 LLM-Based Legislative Graph Optimization

Before utilizing the legislative graph for training and inference tasks, we implement a data optimization phase. This phase is realized using a generative Large Language Model (LLM) and includes: (a) normalizing the textual attributes of nodes to address linguistic challenges such as their verbosity,

complexity, and inconsistencies across states; (b) enhancing the nodes with additional textual information where necessary; and (c) uncovering and formalizing latent relationships embedded within the nodes’ textual data. These steps are crucial for improving performance, as they enable a more cohesive understanding of policies and legislative actions related to the same topic at both intra- and inter-state levels (Section 7).

4.1 LLM Approach to Bill Summarization

In Section 3, we discussed the benefits of adopting bill highlights written by researchers, which serve to normalize the language of state policies across the US. However, our dataset includes these highlights for only a selection of state bills (see Section 5). To enhance practicality and performance, it is necessary to scale these highlights to a larger number of bills, including new ones—for instance, to potentially retrain our model with an expanded dataset or to process new bills at inference time. For this purpose, we utilize a generative LLM to produce highlights for state bills, a task that would be highly tedious for humans. LLMs assist us by reading the full text of a bill and producing a concise summary that capture its main points, objectives, and impacts. We experimented with different prompting strategies for GPT-3.5-turbo-16k to summarize bills. Initially, we posed separate questions and compiled the responses, but this approach resulted in repetitive content. As a result, we refined our prompt to extract a comprehensive highlight, including the potential winners and losers of a bill, in a single query. The advantages of identifying winners and losers will be detailed shortly. We have assessed the quality and efficacy of these AI-generated highlights in Appendix C.

Prompt: What are the detailed highlights, as well as the potential winners and losers, of House Bill HB 5 titled ‘Prohibits Abortions After 15 Weeks’ in Florida?

Highlights (Full version in Appendix C.3)

1. **Abortion Prohibition:** The bill prohibits abortions after 15 weeks of pregnancy, except in cases where the mother’s life is at risk or if there are lethal fetal anomalies.
2. **Criminal Penalties:** It imposes criminal penalties on healthcare providers who perform abortions after 15 weeks, with potential charges ranging from a third-degree felony to a first-degree felony,...
3. **Informed Consent and Ultrasounds:** The bill requires healthcare providers to perform an ultra-

sound and allow the pregnant person to view it before performing an abortion. ...

4. **Reporting Requirements:** The bill imposes additional reporting requirements on healthcare providers who perform abortions, including reporting specified information to the state's Agency for Health Care Administration.

Potential Winners:

1. **Pro-Life Advocates:** Supporters of the bill argue that it protects the rights and lives of unborn children, and they view any limitation on abortion as a step in the right direction.
2. **Religious and Conservative Groups:** ...

Potential Losers:

1. **Abortion Rights Advocates:** Those who support a woman's right to choose may oppose this bill, as it limits access to abortions beyond the 15-week.
2. **Healthcare Providers:** ...

4.2 LLM Augmentation of Legislative Data

LLMs can help us further enhance the state data quality by filling in the gaps in two scenarios: (1) We apply GPT to enrich or augment the textual features of entities. For example, some actors (e.g., bill sponsors) may have hidden agendas or motivations that are not clearly expressed in the bill or its related documents. LLMs can assist by generating more information, better exposing the background, affiliations, and goals of these actors. (2) State-level data is not flawless, and some entities may lack certain attributes (e.g., the mission statement of IGs). We use GPT to create synthetic information for such entities. To achieve this data augmentation, we have devised prompts, as follows:

Prompt: What are the highlights of the political interests of Legislator Scott Thomas Wilk in CA?

Scott Thomas Wilk is a Republican legislator from CA who represents the 21st Senate District. Some of the highlights of his political interests and priorities include:

1. **Education:** Wilk has been a strong advocate for education and improving educational opportunities for Californians. He has expressed support for increasing funding for K-12 schools, ensuring access to quality higher education, and promoting career and technical education.
2. **Veteran Support:**
3. **Water and Natural Resources:**
4. **Fiscal Responsibility:** Wilk emphasizes the need for fiscal responsibility in the California state government. He supports reducing government waste,

lowering taxes, and promoting policies that encourage economic growth and job creation.

Prompt: What is the mission statement of Florida Planned Parenthood PAC?

The mission statement of Florida Planned Parenthood PAC is to protect, promote, and enhance reproductive health care access and rights for all individuals in Florida. They strive to elect candidates who support comprehensive reproductive health care, including access to safe and legal abortion, affordable birth control, and comprehensive sex education

4.3 Extracting Winner/Loser Relations

The third use-case of LLM in our study is to uncover hidden relations within the legislative process across the US. Specifically, as part of generating bill highlights, we utilize LLM to extract winner and loser entities produced by the bill. These entities represent the groups or individuals that either benefit from or are adversely affected by the bill's proposed policy. We depict them as nodes in the legislative graph, along with a brief summary of their interests (see Figure 2). Additionally, we establish links between these entities and relevant bill actions using winner/loser relations. Our ablation study demonstrates the value of these synthetic relations (Section 7), enabling us to compare and contrast bills across different US states, identifying similarities and differences in their content and impact. Furthermore, our LLM method outperforms expert MTurk workers in the same annotation task (Appendix C.2). Recent research (Davoodi et al., 2022) relies on such human-based winner/loser analysis of state bills, which is not scalable.

5 Data Collection

To realize and optimize the legislative graph, we have built a scraper that can extract various data on state legislation, legislators, interest groups, and courage tests from different sources, particularly, Vote Smart (VoteSmart, 2023). Specifically, we have obtained 7K pieces of legislation covering the period from 2019 to 2022. These bills have undergone 16.9K actions. We analyze the 9K main and second floor actions (see Appendix A.1). VoteSmart does not include every legislation from every state legislature; Rather, their research team tracks topical legislation, as well as, unique and novel legislation (i.e. legislation that addresses things never previously addressed by any state legislature). Moreover, we have obtained different information on 7237 legislators, including their biogra-

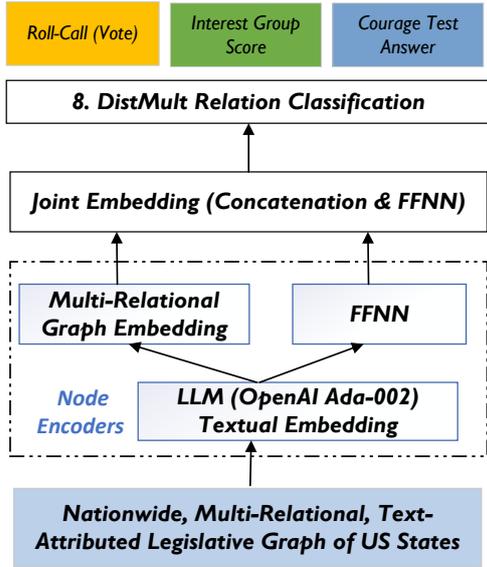


Figure 3: Joint LLM and GNN embedding architecture on nationwide legislative graph (Properties in Sec. 6.2)

phy, political party, religion, and etc. Furthermore, we collected detailed information on 1421 Interest Groups (IGs) including their textual attribute such as mission. We have observed 61K ratings between different pairs of legislators and IGs. Combining the courage test results from all US states, we have observed legislators answering 2216 unique questions. In total, 902 legislators participated in a test. Overall, we have also LLM-extracted 300 entities and 3K winner/loser relations after pruning single-time used entities. As part of our future work, we plan to use LLM to further scale the number of bills in our dataset. For details on data collection and processing, see Appendix B and the Ethics Section.

6 Classification Tasks & Architecture

Next, we explore three important political science applications as relation prediction tasks on the legislative graph (Figure 2). Our optimized legislative graph, utilizing generative LLM techniques, already enhances model performance in these tasks (as detailed in Section 7). To achieve further gains, crucial in our problem domain (see the Ethics section), we develop an efficient joint text and graph (LLM + GNN) embedding architecture.

6.1 Political and Legislative Tasks on Graph

1. Roll-call (RC) Prediction. Our first task involves predicting legislators’ votes on bill actions and subsequently *aggregating* them to determine the fate of each action. This is key to understand

the dynamics of state politics. It reveals what affects legislative behavior, such as party, ideology, voters, interest groups, etc. It also can inform the public on the outcome and implications of state legislation prior to voting.

2. Political Courage Test (PCT) Prediction.

It is crucial to know legislators’ answer to PCT questions, which measure their risk-taking on controversial issues. However, many legislators avoid these tests (Appendix B.3). Thus, predicting their answers can help to: (a) Know their true values, goals, and actions. (b) Compare different legislators (e.g., in parties) and spot any agenda inconsistencies. (c) Motivate legislators to take the test and share their views, thereby gaining trust with voters.

3. Interest Group (IG) Impact Prediction.

Our final task is to predict the IG scores of legislators, which is useful in at least two scenarios: (a) Interest groups (IGs) can use these scores to identify allies or opponents among newly elected legislators by leveraging the broader context in the legislative graph (e.g., winners/losers, courage tests). (b) The public can estimate how IGs interact with their legislators, especially when one or both parties are new and the score has not yet been established.

Dependency. These three relation type prediction tasks are interdependent and leverage the same shared textual and graph contexts within the legislative graph. For instance, the outcomes of the PCT and IG scores can hint at the voting behaviors of legislators (see Figure 1). Thus, we employ a single architecture to address all these tasks.

6.2 Efficient Joint LLM and GNN Embedding

We predict these relation types with an efficient joint text and graph architecture (Figure 3) that has three novel aspects: **(1)** It handles longer legislative texts by using the latest non-generative LLM, while exploiting our bill text highlighting system (Section 7). **(2)** It better deals with unseen nodes in the graph, by utilizing the LLM representation of nodes (Ethics Section). **(3)** It jointly generates and optimizes text and graph embedding, without losing the textual context during in graph embedding like previous methods (Feng et al., 2022; Mou et al., 2021). Next, we explain our architecture from bottom to up.

6.2.1 LLM-Based Text Embedding

Our architecture (Figure 3) leverages LLMs to produce textual representations of the nodes and relations in the legislative graph. Each node has some

text associated with it, e.g., GPT bill highlight, legislator’s political interest, and etc. We prepare and feed this text to LLM to make and freeze a textual embedding for each node. While we support different LLM providers (the Ethics Section), we chose to use `text-embedding-ada-002` from OpenAI. Ada can encode up to 8K tokens into a 1538-dimensional vector and has a longer max sequence length than models such as RoBERTa (Liu et al., 2019), thus enabling our architecture to fully exploit GPT bill highlights.

6.2.2 LLM-Attributed Graph Embedding

We leverage the nationwide context/structure of our graph by passing it through a popular heterogeneous GNN, RGCN (Schlichtkrull et al., 2018). Our goal is to obtain effective graph embeddings for nodes and relations, thereby enriching RGCN training with LLM (OpenAI Ada in Figure 3). In this approach, we extract the text embedding of each node from OpenAI Ada and use it as the initial graph embedding in RGCN for boosting performance. The RGCN model performs graph convolutions, allowing it to learn representations of nodes within our legislative graph, which contains multiple relations. Specifically, RGCN aggregates and weighs input features from nodes and their neighbors based on relation types. It then applies a convolutional layer to refine these features, creating new representations. By repeating these steps across layers, we achieve deeper and more expressive node embeddings. RGCN effectively captures the rich and complex information present in our the legislative graph data using this formula:

$$h_i^{(l+1)} = \text{ReLU} \left(W_0^{(l)} h_i^{(l)} + \sum_{r \in R} \sum_{j \in N_i^r} \frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)} \right)$$

Here, $h_i^{(l)}$ is the hidden representation of node i at layer l , $W_0^{(l)}$ and $W_r^{(l)}$ are learnable weight matrices for layer l , R is the set of relation types, N_i^r is the set of neighbors of node i under relation type r , and $c_{i,r}$ is a normalization constant.

6.2.3 Joint Text & Graph Optimization

We aim to *retain* the important linguistic and textual contexts of nodes, which can be lost through RGCN layers. So we feed nodes’ LLM-based (OpenAI Ada) text embedding through FFNN to preserve it. Then, we join and optimize the text and graph embedding of the node from RGCN and FFNN. Finally, we predict the relation types,

e.g., (legislator, bill action), (legislator, PCT question), and (IG, legislator). We use DistMult (Yang et al., 2014), a factorization model that measures the plausibility of a triple (entity1, relation, entity2) by computing the dot product of their embeddings with relation-specific weights. Here, our loss function consists of (a) cross-entropy loss of the relation classification; (b) L2 regularization of the RGCN weights; and (c) L2 regularization of the FFNN.

7 Experiments

This section assesses our framework’s core aspects: the joint LLM and GNN embedding architecture’s performance gains against strong baselines; the benefits of normalizing and optimizing state-level legislative data with a generative LLM; a performance comparison over three legislative tasks; an ablation study on integrating wider social and political context into the graph; and a performance analysis using data splits that mimic real-world legislative conditions.

7.1 Experimental Setup

Implementation. We used the Deep Graph Library (DGL, 2023) to build both our model and baselines. OpenAI served as our LLM provider, chosen for its efficiency (refer to our feasibility study on open-source providers in Appendix C.4). We used `text-embedding-ada-002` for generating embeddings and `GPT-3.5-turbo-16k` models¹ for our legislative graph optimization. The embedding dimension of Ada is 1536. Both the FFNN and the RGCN (4-layer) models convert the Ada embeddings into a 256-dimensional space. We use Adam optimizer with a learning rate of 0.01.

Data Splits. To evaluate our model, we divide the graph (formed over 50 US states) into train/eval and test sets based on the bill nodes, using 80% of them for training and validation. We select the 20% of the test nodes in three ways: (1) Randomly; (2) By date, where we use bills after a certain date for testing; and (3) By party, where test bills come from either Blue or Red states (Appendix B). For these splits, for each observed (positive) relation, we generate a corresponding negative sample.

Metric. We choose Macro F1 over accuracy due to imbalanced data in our prediction tasks.

7.2 Details of Baseline Models

We build and study three types of robust baselines:

¹We wrote this paper before GPT 4 was available.

Type	Embedding	Raw Graph			LLM-Opt Graph		
		RC	IG	PCT	RC	IG	PCT
Text	BoW	59.8	59.3	58.4	62.8	64.2	61.9
	RoBERTa	67.3	65.2	63.8	69.1	68.6	66.3
	Ada	72.1	71.3	70.8	75.1	74.3	72.5
Graph	DeepWalk	63.3	61.8	61.3	65.7	63.6	62.2
	GCN	65.2	63.7	62.4	66.5	64.2	63.8
	RGCN	70.2	69	68.1	72.2	71.3	69.4
Joint	Ada+ Txt.	74.4	75.3	72.8	77.8	77.1	74.9
	RGCN						

Table 1: Random split. Performance (Avg. Macro F1 \times 100 in 10 runs) of models with and without our LLM-based legislative graph opt. in Roll-call (RC), Interest Group (IG), and Political Courage Test (PCT) tasks.

- Text models:** We create text classifiers for each task using a Random Forest on embeddings from three models: (a) **BoW** utilizes the top 5K uni/bi-grams. (b) **RoBERTa-Large**, a small language model (SLM). (c) **Ada**, the latest OpenAI non-generative LLM. For the roll-call (RC) task, we directly determine bill actions’ fate by computing the embedding of (bill highlight, sponsors), without vote aggregation. When predicting interest group (IG) scores, we calculate the embedding of (legislator, IG). For predicting courage test (PCT) responses, we generate embedding of (legislator, PCT question).
- Graph models:** We train relation classifiers with edge embeddings from: (a) **DeepWalk** (Perozzi et al., 2014), a model that ignores relations and forms SkipGram on random walk node sequences. (b) **GCN**, a standard 4-layer GCN model with random node features. (c) **RGCN**, the relational version of the GCN.
- Joint models:** In our ablation study (Section 7.6), we examine weaker versions of our joint text and graph model, limited to the roll-call prediction task. Specifically: (a) **EX-IG** excludes interest group (IG) relations and nodes. (b) **EX-PCT** omits the courage test contexts. (c) **EX-WL** removes the winner/loser (WL) relations and nodes.

7.3 Benefit of Joint Text and Graph Model

To understand the benefit of our joint LLM and GNN model in isolation, we first *disable* our generative LLM-based optimization on the legislative graph for all the models. We look at the roll-call task results in Table 1 for the random split on the raw graph and notice the following: (1) Comparing the text-based models, Ada (LLM) achieves

4.8 points higher F1 than RoBERTa (SLM). This is because Ada is trained on a larger text corpus and is speculated to have $7.63 \times$ more parameters. More importantly, Ada handles longer legislative texts (i.e., researcher-written bill highlight) by having $16 \times$ larger max sequence length (both having similar embedding size, 1024 for Roberta and 1536 for Ada). (2) Among the graph models, RGCN better captures different relation types in the graph, thus delivering higher performance compared to GCN/DeepWalk that are relation agnostic models. (3) Our joint model outperforms the leading individual text or graph models by 2.3 and 4.2 points, respectively. It effectively maintains the textual and relational contexts of policy-making in the legislative process, thereby merging the strengths of the individual models.

7.4 Effect of LLM-Based Graph Optimization

Next, we *activate* our generative LLM-based optimization on the legislative graph and revisit the roll-call task as shown in Table 1. We observe that: (1) Our optimization improves all baseline models and our own, owing to the normalization of legislative language across US states. (2) The optimization has a reduced impact on RoBERTa compared to Ada. RoBERTa is further hindered by its shorter max sequence length, which limits its ability to utilize the GPT-generated highlights that are, on average, longer than those authored by the researchers in the previous experiment. Conversely, the Ada model sees a 3-point increase in F1 score, due to its advanced features, including a larger number of parameters, enhanced training data, etc. (3) Baseline graph models fail to consider the textual attributes of nodes improved by our optimization, leading to only a modest enhancement derived from the winner/loser relationships identified by GPT. (4) Our joint text and graph model continues to surpass other models in this task, achieving an F1 score of 77.8.

7.5 Analysis of PCT and IG Score Prediction

Next, we focus on the performance of the other two tasks presented in Table 1. For all models, regardless of the inclusion of our LLM-based graph optimization, we observe that predicting legislators’ responses to a political courage test (PCT) is more challenging than predicting their scores for a certain IG. These PCT questions require legislators to declare their positions on divisive issues during state elections. At times, they address very

specific issues, for which the legislative graph may not provide adequate context (e.g., related bills and votes). This lack of context poses a challenge to the models’ learning capabilities and their ability to make accurate predictions. However, IGs have broader agendas (e.g., see the IG mission statement in Section 3). As a result, all models show improved performance on this task, benefiting from a more comprehensive contextual information.

	Joint Model Variations			
	BEST	EX-IG	EX-PCT	EX-WL
Avg. Macro F1	77.8	75.7	76.2	76.3

Table 2: Weaker versions of the joint model (random split). Ablation study on the impact of different contextual information in the LLM optimized legislative graph.

7.6 Ablation Study on Legislative Graph

Our legislative graph abstraction, which incorporates various social and political contexts, enhances the understanding of the state-level legislative process. These contexts improve both baseline and joint models across the three prediction tasks. Here, we detail how each component affects the F1 score in the roll-call task (Table 2). In our joint model with the LLM-optimized graph (F1 of 77.8, Table 1), removing IG relations (**EX-IG**) decreases the F1 score by 2.1 points, while excluding courage test relations (**EX-PCT**) and winner/loser relations (**EX-WL**) reduces it by 1.6 and 1.5 points, respectively. Removing multiple contexts simultaneously can diminish the advantages of our joint model over the best text model (Ada). Overall, the text models are unaffected by the exclusion of these contexts, except when utilizing the textual winner/loser information in the GPT bill highlights. To preserve the joint model’s superiority, adjusting our loss function has proven effective. We can apply weights to the graph and text components’ losses in our model to rely more on the textual embeddings. Note that the baseline graph model (RGCN) experiences a comparable performance drop when different contextual information is excluded.

7.7 Impact of Party- and Time-based Splits

Finally, we compare different models with the time and party split over the LLM-optimized graph. Here we present the results for the best model in each category (shown in Table 3) and make some observations: (1) All models show a lower F1 score when they face a more realistic situation that may include significantly higher unseen nodes—not observed in training data, e.g., newly elected legisla-

Type	Embedding	Time	Party (Blue)	Party (Red)
Text	Ada	72.1	73.7	74.2
Graph	RGCN	68.5	69.2	70
Joint	Ada+ Txt. RGCN	74.8	75.8	76.5

Table 3: Time and party-based graph splits. Performance (Avg. Macro F1 \times 100 in 10 runs) of best models in each category for the Roll-call task when LLM-based legislative graph optimization is enabled.

tors. (2) The time-based split gave us more unseen nodes, which leads to more performance loss. (3) In the party-based splits, we had more Red states than Blue states in our training data (in both settings), so we observed a slightly higher F1 score when the test bills are from Red states. (4) Generally, the baseline graph model (RGCN) performs significantly worse than the text models. This is attributed to its default method of managing and approximating the embeddings of unseen nodes, which is suboptimal and involves combining the embeddings of potential neighboring nodes based on relational weights. Our joint model handles unseen nodes better thanks to our LLM-based entity alignment (Ethics Section), thereby outperforming both the graph and text baselines.

8 Summary

In this paper, we introduce an efficient framework for analyzing the state-level legislative process. This framework interconnects legislative activities across U.S. states by creating a national network that includes often-overlooked elements such as interest groups and political stances. Moreover, it is the first to tackle the linguistic complexities of state legislative data by utilizing generative LLMs to normalize, summarize, and enhance this information for improved policy analysis. In addition, we develop a legislative graph abstraction and define several key prediction tasks concerning the interactions between legislators, bills, and interest groups. We design an effective joint LLM and GNN architecture to handle these tasks, which is built to accommodate unseen nodes and support longer texts, thus further leveraging the enhanced linguistic and network contexts. The next section discusses framework limitations and ethical aspects.

Acknowledgement

We would like to thank the reviewers for their constructive feedback. This work was partially supported by an NSF CAREER award IIS-2048001.

Limitations and Ethical Considerations

One of the main challenges of our work is to analyze state legislation in a responsible and ethical way. We are aware that the outcomes of our research and the applications of our models can have significant impacts on how people perceive and evaluate government policies on crucial issues such as health and education. Thus, we want to address limitations and ethical aspects of our work, both in terms of the data we use and the model we develop:

A. Model-Specific Challenges & Considerations

Scaling training and inference. We aim to improve the inference speed of our joint LLM and GNN model (Yang et al., 2020), which is heavier than a pure LLM or GNN approach. To achieve this, we will explore different techniques, such as graph knowledge distillation (KD), to transfer knowledge from a large and complex teacher model to a smaller and simpler student model. Moreover, we will use more efficient distributed training platforms, such as DeepSpeed (Microsoft, 2023a), offering lower memory and bandwidth requirements than DGL, while maintaining the accuracy and convergence of the original models.

Choice of specific LLM and GNN: Our joint text and graph architecture supports various LLM and GNN providers. We chose OpenAI GPT/Ada and RGCN as our LLM and GNN models, respectively, due to their popularity and performance. This allowed us to stay focused on our contributions, e.g., the benefits of optimizing legislative graph data using generative LLMs; and handling key political science tasks using our shared LLM and GNN architecture. For the latter, we compared our approach with the solo LLM models, which are unable to prioritize key contexts and players in the legislative process, and the solo GNN models, which miss out on crucial linguistic dependencies and information. In Appendix C.4, we explore the possibility of using open-source models, e.g., Phi-2 and Llama-2, in our framework.

Improving handling of unseen entities. RGCN uses neighbor estimation and inductive inference to handle unseen nodes. To realize that in our framework, initially, we established neighborhood relations for unseen nodes by performing heuristic exact matches on their textual attributes (e.g., party of legislators). Later, we improved this by devising an LLM-based entity alignment algorithm. To determine neighbors of an unseen node, we sample a

finite set of nodes of the same type in the same US state. Then, we compute cosine similarity over the node’s LLM (Ada) text embedding and that of other nodes in the graph. This gives our text-attributed RGCN a very effective set of neighbors to generate the graph embedding of each unseen node based on that of its neighbors. Injecting the unseen nodes in the train graph and further fine-tuning the model could be a future improvement.

B. Data-Specific Limitations & Considerations

Select bill actions. In this study, we have focused on “floor vote” or “roll call vote” that is a vote taken by the full membership of either chamber of the state legislature on a bill that has been debated on the floor. Our framework supports other bill actions, detailed in Appendix A.1, e.g., Introduction, Committee Review, Governor’s Action.

Handling abstain votes. We did not consider abstain votes (marked as absent or N/A) in our study. While our proposed model can handle them, we did not study them as they are very rare. We think predicting abstain votes is an interesting problem that deserves further exploration in future work.

Data source licensing and bias. In our study, we have gathered state-level legislative data primarily from Vote Smart (VoteSmart, 2023). As a non-profit organization, Vote Smart offers free, trustworthy, and unbiased info. about U.S. political figures. They analyze public data from diverse sources, including legislative votes, campaign funding reports, public speeches, interest group ratings, and personal backgrounds. Additionally, VoteSmart actively promotes API usage and supports 3rd party development to empower civic engagement and informed decision-making during elections.

Data Leakage and copy-cat bills. Copycat bills are model bills that are replicated across states, often by interest and lobbying groups (ALEC, 2023). These bills present an intriguing challenge for political science research. However, we believe that they did not significantly impact our study and results. Our dataset, based on Vote Smart, only includes topical and novel legislation, excluding copycat bills (Section 5). Even if such bills were present in our dataset, our evaluation remains fair and consistent, as all models utilize the same data. Looking ahead, we aim to extend our architecture to identify and classify copycat bills as a new task. Achieving this would require either automatic (LLM) or human annotation of state bills.

References

- ALEC. 2023. Model policies. <https://alec.org/model-policy/>.
- Ballotpedia. 2024. Abortion regulations by state. https://ballotpedia.org/Abortion_regulations_by_state.
- Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with gpt-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300.
- Daniel E Chand and William D Schreckhise. 2013. Keeping score on congress: explaining variations in interest group ratings of us senators. *Business and Politics*, 15(3):401–433.
- Maryam Davoodi, Eric Waltenburg, and Dan Goldwasser. 2020. Understanding the language of political agreement and disagreement in legislative texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5358–5368.
- Maryam Davoodi, Eric Waltenburg, and Dan Goldwasser. 2022. Modeling us state-level policies by extracting winners and losers from legislative texts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 270–284.
- DGL. 2023. Deep graph library. <https://www.dgl.ai/>.
- Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.
- Vlad Eidelman, Anastassia Kornilova, and Daniel Argyle. 2018. How predictable is your state? leveraging lexical and contextual information for predicting legislative floor action at the state level. *arXiv preprint arXiv:1806.05284*.
- Shangbin Feng, Zhaoxuan Tan, Zilong Chen, Ningnan Wang, Peisheng Yu, Qinghua Zheng, Xiaojun Chang, and Minnan Luo. 2022. Par: Political actor representation learning with social context and expert knowledge. *arXiv preprint arXiv:2210.08362*.
- FiveThirtyEight. 2023. How red or blue is your state? <https://fivethirtyeight.com/features/how-red-or-blue-is-your-state-your-congressional-district/>.
- Martin Gilens and Benjamin I Page. 2014. Testing theories of american politics: Elites, interest groups, and average citizens. *Perspectives on politics*, 12(3):564–581.
- John E Jackson and John W Kingdon. 1992. Ideology, interest group scores, and legislative votes. *American Journal of Political Science*, pages 805–823.
- The Wallstreet Journal. 2023. Politicians grow wary of survey as internet spreads attack ads. <https://www.wsj.com/articles/SB116173744805102790>.
- Hamid Karimi, Tyler Derr, Aaron Brookhouse, and Jiliang Tang. 2019. Multi-factor congressional vote prediction. *Advances in Social Networks Analysis and Mining (ASONAM)*.
- Anastassia Kornilova, Daniel Argyle, and Vladimir Eidelman. 2018. Party matters: Enhancing legislative embeddings with author attributes for vote prediction. In *Proceedings of ACL*.
- Peter Kraft, Hirsh Jain, and Alexander M Rush. 2016. An embedding model for predicting roll-call votes. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.
- LegiScan. 2023. State-level legislative data. <https://legiscan.com/>.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Microsoft. 2023a. DeepSpeed. <https://www.microsoft.com/en-us/research/blog/deepspeed-zero-a-leap-in-speed-for-llm-and-chat-model-training-with-4x-less-communication/>.
- Microsoft. 2023b. Phi-2: The surprising power of small language models. <https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/>.
- Xinyi Mou, Zhongyu Wei, Lei Chen, Shangyi Ning, Yancheng He, Changjian Jiang, and Xuan-Jing Huang. 2021. Align voting behavior with public statements for legislator representation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1236–1246.
- Xinyi Mou, Zhongyu Wei, Qi Zhang, and Xuan-Jing Huang. 2023. Uppam: A unified pre-training architecture for political actor modeling based on language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11996–12012.
- NCSL. 2024. Ncsl research tools: Legislation databases and the bill information service. <https://www.ncsl.org/about-us/ncsl-research-tools>.
- OpenAI. 2023. Gpt platform and api. <https://platform.openai.com/docs/api-reference>.
- Maria Leonor Pacheco and Dan Goldwasser. 2021. Modeling content and context with deep relational learning. *Transactions of the Association for Computational Linguistics*, 9:100–119.

- Pallavi Patil, Kriti Myer, Ronak Zala, Arpit Singh, Sheshera Mysore, Andrew McCallum, Adrian Benton, and Amanda Stent. 2019. Roll call vote prediction with knowledge augmented models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 574–581.
- Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*.
- Rajkumar Pujari and Dan Goldwasser. 2021. Understanding politics via contextualized discourse processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1353–1367.
- Shamik Roy, María Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958.
- Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer.
- Gregory Spell, Brian Guay, Sunshine Hillygus, and Lawrence Carin. 2020. An embedding model for estimating legislative preferences from the frequency and sentiment of tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–641.
- LM studio. 2024. Lm studio. <https://lmstudio.ai/>.
- VoteSmart. 2023. Vote smart: The voter’s self defense system. <https://justfacts.votesmart.org/>.
- WashingtonMonthly. 2023. Why it’s so hard to find out where the candidates stand. <https://washingtonmonthly.com/2016/11/04/why-its-so-hard-to-find-out-where-the-candidates-stand/>.
- Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short 12 essays on the cefr scale with gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584.
- Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.
- Yiding Yang, Jiayan Qiu, Mingli Song, Dacheng Tao, and Xinchao Wang. 2020. Distilling knowledge from graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7074–7083.
- Tae Yano, Noah A Smith, and John D Wilkerson. 2012. Textual predictors of bill survival in congressional committees. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 793–802. Association for Computational Linguistics.

A Appendix: State Legislative Process

In this appendix, we provide a more detailed discussion on the state-level legislative process and its importance.

A.1 Background

In this paper, we focused on “floor vote” or “roll call vote”, as explained in the Ethics section. A state bill is a proposed law that must go through several stages in both houses of the legislature before becoming a law. The process is similar for the House (upper chamber) and the Senate (lower chamber), but with some variations. The main stages are: (1) Introduction: A legislator proposes a bill and gives it to the clerk of their respective house. The bill gets a number, a title, and a committee. (2) Committee Review: The committee reviews the bill and decides whether to support it, oppose it, or make changes to it. The committee can also ignore the bill and let it die. (3) Floor action: The house votes on the bill after debating and amending it. The bill needs a majority vote to pass and go to the other house. (4) Second floor action: The second house repeats the same steps as the first house. If the second house agrees with the first house, the bill goes to the governor. If the second house disagrees with the first house, they can try to reach a compromise through a conference committee. (5) Governor’s action: The governor can sign the bill into law, veto it, or do nothing. The legislature can override the governor’s veto with a two-thirds vote in both houses.

A.2 National Impact of State Legislation

The impact of state bills and laws is not limited to the state level. They also influence national matters, as federal courts often respect the states’ decisions on constitutional questions (see examples of the federal abstention doctrine in Table 4). Moreover, some state laws have inspired or provoked the enactment of federal laws (See examples in Table 5).

B Appendix: Data & Experimental Details

Here, we include more details about our dataset as well as experiments:

B.1 Categorizing States: Red, Purple, Blue

In Section 7, we categorized each U.S. state as red, purple, or blue. However, there is no definitive or official way to classify states as red, purple, or blue

1 | In *Sprint Communications, Inc. v. Jacobs* (2013), the Supreme Court said that federal courts should not hear a case about a state agency’s regulation of telecom rates, as the state had a specific process for reviewing and appealing such issues.

2 | In *Google LLC v. Oracle America, Inc.* (2021), the Supreme Court did not review a state court decision that said that Google’s use of Oracle’s software code was fair use under federal copyright law, because the state court followed federal law correctly and there was no disagreement with other federal rulings.

Table 4: *Examples of the abstention doctrine, a principle that guides federal courts to avoid deciding constitutional issues if there is a state law issue that could resolve the case.*

1 | The state of California passed a legislation in 2018 that mandated ISPs to not discriminate any web traffic based on its origin or content. This legislation was a response to the FCC’s decision to revoke net neutrality, which had ensured equal access to the internet for all users. The California legislation inspired other states to propose similar laws, and also triggered a discussion in Congress about whether to reinstate net neutrality at the national level.

2 | Several states have legalized the use of marijuana for recreational purposes by adults since 2012, when Colorado and Washington became the first ones to do so. This has created a contrast with the federal ban on cannabis, and has increased the demand for the federal government to change its drug policy. The legal marijuana industry also poses challenges and opportunities for regulation, public health and safety, and states’ rights.

Table 5: *Some examples of how state bills have inspired or influenced the enactment of federal laws.*

based on solely their state legislatures. Different sources may use different criteria or thresholds to determine the color of a state. For example, some sources may consider a state purple if it has a divided government, meaning the governor and at least one chamber of the state legislature belong to different parties. Other sources consider a state purple if it has a close partisan balance in both chambers of the state legislature, meaning neither party has a large majority. We adopt the definition of partisan lean proposed by (FiveThirtyEight, 2023). This metric represents the average margin difference between a state’s voting pattern and the national average in congressional and gubernatorial elections. It incorporates data from both presidential and state-legislative election results:

- **Red states:** States with a partisan lean of R+10 or more

- **Blue states:** States with a partisan lean of D+10 or more.
- **Purple states:** States with a partisan lean between R+10 and D+10.

Using this method, here is the list of red, purple, and blue state. As can be seen from the list, there are more red states than blue or purple states:

- **Red states (22):** Alabama, Alaska, Arkansas, Idaho, Indiana, Kansas, Kentucky, Louisiana, Mississippi, Missouri, Montana, Nebraska, North Dakota, Ohio, Oklahoma, South Carolina, South Dakota, Tennessee, Texas, Utah, West Virginia, Wyoming.
- **Blue states (13):** California, Connecticut, Delaware, Hawaii, Illinois, Maryland, Massachusetts, New Jersey, New York, Oregon, Rhode Island, Vermont, Washington.
- **Purple states (15):** Arizona, Colorado, Florida, Georgia, Iowa, Maine, Michigan, Minnesota, Nevada, New Hampshire, North Carolina, Pennsylvania, New Mexico, Virginia, Wisconsin.

B.2 Calculating Interest Group (IG) Scores

We explained how our model benefits from and uses the interest group (IG) scores of legislators in Sections 3 and 6. Here, we provide additional details on how these IG scores are calculated and processed by Vote Smart. These scores are often publicized by interest groups (IGs) in an attempt to influence the future behavior of legislators. The IG scores we collected from Vote Smart are represented as percentages ranging from 0 to 100. There are several ways in which these scores are adjusted by Vote Smart

- The vast majority of the interest groups use a 0-100 system for rating candidates, so Vote Smart simply consumes the rating already provided to it.
- Some interest groups provide a list of recent legislation that the interest group was for and against and show how each legislator voted on that legislation. Vote Smart easily takes this data and calculates a percentage that represents how often a legislator voted in favor of the interest group's legislative position (most interest groups that actually list a 0-100 score do so through this exact method).

- Some interest groups use letter grades to rate candidates. If this is the case Vote Smart uses a basic linear equation to turn these grades into numbers on a 0-100 scale. Vote Smart assigns the lowest possible letter grade a score of "0", the highest possible grade a score of "100", and then use an equation to assign scores for all other possible grades the candidate could receive.
- Some interest groups have other ways of rating candidates. Vote Smart studies any such rating system carefully and try to develop a fair way to turn the rating into a number on a 0-100 scale. If there is no fair or accurate way to translate the rating to a 0-100 grade, Vote Smart simply does not post it.

B.3 Details of Political Courage Tests (PCT)

In Section 5, we provided an overview of the Political Courage Tests (PCTs) data that we have gathered. Here, we delve into more details on how these tests have evolved over time and why we should infer legislators' response to them (Section 6). The PCT asks state candidates who want to be elected to answer questions about different topics that will be crucial in the upcoming term. The test is made with the help of more than 200 experts from different fields and political views. The test answers are shown on Vote Smart's website, helping voters learn more about their options. The test also encourages candidates be honest and responsible for their views. Overall, these tests are optional and usually available during elections. If candidates take them, their answers are shared with the public. Also, the questions in the PCT could be slightly different across states and election years.

Predicting legislators' responses to these tests are crucial, as the number of candidates who take the PCT has gone down, from 72% in 1996 to 48% in 2008 and to 20% in 2016 ([WashingtonMonthly, 2023](#)). The reason is that candidates from both parties are scared that their rivals will use their answers to attack them in ads. Rep. Anne Gannon, a Democratic leader in the Florida House of Representatives, said: "We tell our candidates not to do it. It sets them up for a hit piece.". To change this, Vote Smart has tried to make motivate candidates to take the test, and allows them to skip up to 30% of the questions ([Journal, 2023](#)). Note that If a candidate does not select a response to any part or all of any question, it does not necessarily indicate that

the candidate is opposed to that particular item.

B.4 Evaluation Details: Need for Normalizing Language of State Bills

State legislative texts often exhibit inconsistent, verbose, and complex language (Sections 1, 3). Here, we present an experiment to underscore the necessity of normalizing the text and language used in these bills. To achieve this, we employed a text embedding and classification architecture (The text model based on Roberta, as discussed in Section 7). Next, we compared various ways of representing state legislation: Official title, Official synopsis, and Bill highlights crafted by a centralized team of political science researchers from Vote Smart. The latter option ensures a consistent language across all bills, regardless of the legislator responsible for drafting them. We then applied our study to two classification tasks: (1) Predicting roll-call outcomes (pass/fail); (2) Identifying sharp splits in votes based on the ideological or demographic orientation of legislators, as in (Davoodi et al., 2020). Our results revealed that bill highlights significantly enhance the performance of both classification tasks when compared to using the official titles and synopses. We observed a rather significant improvement of up to 3.5 points in the F1 score.

C Appendix: Additional Evaluation and Experiments

In this appendix, we detail some additional evaluation on our framework:

C.1 Evaluation of LLM-Generated Bill Highlights

We also conducted a brief study on the problem of hallucination. Specifically, We investigated the impact of applying LLM-based optimization to textual information within our legislative graph. Our focus centered on cases where GPT-generated bill highlights contained factually incorrect or unsupported information. In the Ethics section, we discussed the potential risks associated with these phenomena. To evaluate the accuracy of GPT-generated highlights, we conducted a small-scale measurement. We randomly sampled 150 bills and compared the GPT highlights against two ground-truth sources: (1) Highlights written by the Vote Smart researchers; and (2) The full text of the bills. During this evaluation, two annotators assessed the

factual correctness of the GPT highlights, without considering informativeness, coherence, or fluency. In 95% of the sampled bills (based on the majority vote), we discovered that GPT-generated highlights were accurate—sometimes even surpassing researcher-written highlights. This finding is further supported by the F1 performance results of our classification tasks (such as roll-call prediction) before and after enabling GPT-based optimization on the legislative graph (Section 7). As part of our future work, we aim to better understand any potential biases introduced by GPT-based summaries into our system. We plan to explore more sensitive classification tasks, including predicting gender-charged voting splits (Davoodi et al., 2020). Additionally, we are considering using MTurk to compare LLM-generated highlights with those written by experts for select bills (Similar to our study in Appendix C.2). It’s worth noting that atomic evaluation metrics such as ROUGE and BLEU are not sufficient, as they focus solely on simple n-gram-based matches between the generated summary and the ground truth.

C.2 Evaluation on LLM-based Winner/loser Extraction

In Section 7, we conducted an ablation study on the effect of each additional context within our graph. We showed that adding GPT-generated winners/losers from bill texts to the graph improves our roll-call prediction performance. Here, we use the crowd-sourcing technique, similar to (Davoodi et al., 2022), to further study the effectiveness of our approach:

Questions/Tasks
How many female justices are currently on the US Supreme Court?
Which party currently holds the majority of seats in the US Senate?
What is the main topic of the following bill? <i>Prevention and control of, emergency and involuntary commitment for, and treatment programs and services for drug dependence.</i>
Select the entities that would lose benefits from this bill? <i>Requires Oregon Health Authority to commission independent study of costs and impacts of operating basic health program in Oregon. Specifies parameters of study. Requires a report to Legislative Assembly by November 30, 2014. Appropriates money from General Fund to authority for contract costs to conduct study. Declares emergency, effective on passage.</i>

Table 6: Sample questions from our Political Science qualification test.

Setup and results. We conducted a comparison between the LLM-extracted results and human annotations. We used MTurk to identify the stakeholders who would gain or lose from the bills’ policy, regardless of its legislative outcome. Since annotating state bills was difficult for typical MTurk workers, we created a Political Science Qualification test with different tasks (e.g., Sentiment analysis, winner/loser extraction, political knowledge). Table 6 provides an overview of some of the tasks. Then, we selected annotators (mostly US-based) who passed the test with 80% accuracy and added them to our pool. The test had a 30-minute time limit and no retries. We assigned 3 annotators from our pool to each bill highlight (ground-truth Vote Smart) and asked them to assess the impact of bill policies on the stakeholders (the list of stakeholder for each bill was extracted from our LLM analysis). Then, we determined the winners and losers of each bill based on the majority of annotators’ choices. If there was still no consensus, we break the tie randomly. Finally, we compared the human-based annotation results from MTurk with those from GPT for 150 bills. In 40% instances where there were differences, we found out that the LLM results were accurate 85% of the time. This further confirmed our method is more accurate and less expensive than human annotation.

C.3 Example of GPT-Generated Bill Highlights

Table 8 shows a sample state bill that were summarized by the GPT 3.5 model.

C.4 Open Source Models: A Case Study of Phi-2 and Llama-2

We built our framework using closed-source OpenAI LLMs (Ada and GPT 3.5), but it can work with any language model that satisfies our performance and reliability criteria. We also explored the possibility of using open source models in our framework (e.g., to reduce the inference cost). We specifically selected the 3.7B parameter Phi-2 (Microsoft, 2023b) from Microsoft and the 7B parameter Llama-2 from Meta as potential alternatives.

Choice of models. These models have much fewer parameters (by an order of magnitude) than the GPT 3.5 model that we used in our study, but they have demonstrated remarkable performance in particular benchmarks. According to Microsoft: “Phi-2 outperforms LLMs such as the 7B Mistral, 13B Llama-2, and even the 70B Llama-2 on some

Open Source LLMs	% of summarized test bills			
	0	2	4	8
Llama-2 (13B)	77.8	77.6	75.4	72.1
Phi-2 (2.7B)	77.8	76.2	74.3	70.4

Table 7: Impact on the F1 performance when incorporating open-source models into our framework during inference for bill highlight generation. Our joint text and graph model was trained using the random split and for the roll-call (RC) task in Table 1.

benchmarks. It also matches or surpasses the performance of Google Gemini Nano 2, which is larger in size. The benchmarks include various tasks such as reasoning, language comprehension, mathematics, coding challenges, and more.”

Implementation details. We used LM Studio (studio, 2024) to run these open source models with minimal code modification in our framework. LM studio is a solution designed for running and experimenting with different LLMs locally. We opted for LM Studio due to its chat interface, OpenAI-compatible local web server, and seamless integration with HuggingFace models. We communicate with LM Studio by sending HTTP requests to its OpenAI compatible API endpoint locally.

Experimental setup. We present a limited experiment to investigate the feasibility and performance of these models. Due to some challenges that we will explain shortly, we did not train our joint text and graph architecture with these language models; Instead, we use them at the inference stage for generating bill highlights (but not for augmenting the other data in the legislative graph in Section 4). Specifically, we retrieve our model checkpoint trained for roll-call task (Table 1). Then, for a small fraction of bills in the test dataset, we tasked Phi-2/ Llama-2 with creating bill highlights. We continue to rely on GPT for the remaining bills and all legislators during the roll-call inference, *In essence, bill highlights can be generated using various methods during inference, including manual human annotation.* Our underlying OpenAI Ada and RGCN model continue to generate the embeddings and model score in the roll-call prediction task. Finally, we evaluate our joint model’s performance using the average F1 score.

Analysis. The open source models Phi-2 and Llama-2 exhibit declining F1 scores based on the increasing proportion of test bills they summarize, as indicated in Table 7. Our observations highlight a decline in the F1 performance when we utilize them, even if only partially,

for summarizing a small subset of bills. Upon analyzing the bill highlights generated by these models (through random sampling, similar to our analysis in Appendix C.1), we identified three primary issues impacting the classification performance: *(1) Incomplete capture of highlights:* These models occasionally fail to capture the essential aspects of bills, resulting in superficial or empty highlights. For instance, in Table 9 and Table 10, both Phi-2 and Llama-2 produced inadequate highlights for an abortion bill when compared to GPT and our ground-truth (Table 8). *(2) Incorrect or overly detailed winners/losers:* Another problem arises when these models generate lists of potential winners or losers associated with bills. We observed a highly detailed and exaggerated list of winners/losers in some cases. In the case of the abortion bill, Llama-2 included unborn children as winners. This issue impacts the underlying efficiency in generating embeddings for these bills. *(3) Factual errors and lack of frequent updates:* More importantly, these models struggle with summarizing newer bills, potentially due to infrequent updates or limited training data. Consequently, their summaries may contain factual errors (such as incorrect dates) or irrelevant information (fake highlights not mentioned in the bill). Consider the case of state bill LD51 titled “An Act To Enact the Maine Insurance Data Security Act” in Maine during 2021. Phi-2 failed to identify any winners or losers for this bill. More importantly, it made a year-related mistake in its summary. Llama-2 inaccurately asserted that insurers were required to notify the Maine Office of the Attorney General (OAG), within 72 hours of detecting a security breach involving sensitive information. However, this requirement was absent from the actual full text of the bill. Notably, GPT did not make any error when summarizing the same bill.

Readiness. While these two open-source models could indeed serve as viable alternatives to the OpenAI GPT model in various problem domains, their performance remains uncertain for our political science problem. Factual error in summarizing bills could significantly impact the public or legislative bodies. Currently, using them for training or scaling up during inference would undermine the purpose of bill summarization in our work, compared to using the official abstract. As part of our future work, we intend to explore even larger open-source models, such as the 70B Llama-2, to over-

come these limitations. Additionally, we plan to evaluate the efficiency of these open-source models as an alternative to OpenAI Ada for embedding.

D Appendix: Further Discussion on Related Work

In Sections 1 and 2, we discussed the most recent NLP/ML studies that have developed promising methodologies allowing vast amounts of legislative text to be analyzed and incorporated into models of legislative behavior. However, they have some key limitations: most focus on the centralized federal-level process, miss out on some state-level contextual information (e.g., the influence of local lobbyists), use low-quality and inconsistent state data, have limited support for non-vote prediction tasks, or often rely on traditional NLP/ML architectures. In this section, we delve into additional works that were not fully covered in our related work section 2 due to space limitations.

Incorporating social network contexts. Some recent studies at the congressional level have integrated social context and valuable expert knowledge in the roll-call prediction task. [Feng et al. 2022](#) builds a heterogeneous information network for Congress legislators based on their social network info (Wikipedia information) and expert knowledge (two political think tanks). It uses an RGCN network to embed this network and predict votes. [Mou et al. 2021](#) creates a similar network to predict the voting behavior of Congress legislators using their public statements on Twitter. [Mou et al. 2023](#) employs the same dataset but develops a Pretrained Language Model (PLM) based on legislator’s profile (e.g., ideology) and behavior (e.g., voting pattern) to examine different political science tasks (e.g., roll-call prediction). In addition, [Spell et al. 2020](#) develops an embedding model that predicts the frequency and sentiment of congress legislator tweets based on the text of Trump’s tweets. It claims that ideal points estimated from votes do not capture this aspect of political preferences for members of Congress. [Pujari and Goldwasser 2021](#) create a joint embedding model for politicians, twitter posts and event information to predict legislators’ interest group grades based on their votes. [Roy et al. 2021](#) use both text and twitter network information to predict moral stances by politicians using a neuro-symbolic modeling approach ([Pacheco and Goldwasser, 2021](#))

We use political courage tests to measure the

Title	Prohibits Abortions After 15 Weeks
Synopsis	A vote to pass a bill that prohibits abortions after 15 weeks with no exceptions for rape, incest, or human trafficking.
VoteSmart	<ul style="list-style-type: none"> Prohibits a physician from performing an abortion for individual if they determine that the gestational age of the fetus is more than 15 weeks, without exceptions for rape, incest, or human trafficking (Sec. 4). Requires the director of any medical facility in which abortions are performed to submit a monthly report each month to the state legislature documenting the medical and surgical procedures done including the number of abortions (Sec. 5). Specifies that this act shall take effect July 1, 2022 (Sec. 8).
GPT	<p>House Bill HB 5, titled "Prohibits Abortions After 15 Weeks in FL," aims to restrict access to abortions in the state of Florida by prohibiting the procedure after 15 weeks of pregnancy. Here are some detailed highlights of the bill, as well as potential winners and losers:</p> <p>Highlights:</p> <ol style="list-style-type: none"> Abortion Prohibition: The bill prohibits abortions after 15 weeks of pregnancy, except in cases where the mother's life is at risk or if there are lethal fetal anomalies. Criminal Penalties: It imposes criminal penalties on healthcare providers who perform abortions after 15 weeks, with potential charges ranging from a third-degree felony to a first-degree felony, depending on the circumstances. Informed Consent and Ultrasounds: The bill requires healthcare providers to perform an ultrasound and allow the pregnant person to view it before performing an abortion. They must also provide certain information to the patient about the ultrasound images. Reporting Requirements: The bill imposes additional reporting requirements on healthcare providers who perform abortions, including reporting specified information to the state's Agency for Health Care Administration. <p>Potential Winners:</p> <ol style="list-style-type: none"> Pro-Life Advocates: Supporters of the bill argue that it protects the rights and lives of unborn children, and they view any limitation on abortion as a step in the right direction. Religious and Conservative Groups: The bill aligns with the beliefs and values of religious and conservative groups that oppose abortion. <p>Potential Losers:</p> <ol style="list-style-type: none"> Abortion Rights Advocates: Those who support a woman's right to choose may oppose this bill, as it limits access to abortions beyond the 15-week mark. Healthcare Providers: The bill puts healthcare providers in a difficult position by potentially subjecting them to criminal charges if they perform abortions after 15 weeks, even in cases where the pregnant person's health is at risk.

Table 8: Abortion Bill: Comparing bill title, synopsis, researcher-written highlights and GPT-generated highlights.

Title	Prohibits Abortions After 15 Weeks
Synopsis	A vote to pass a bill that prohibits abortions after 15 weeks with no exceptions for rape, incest, or human trafficking.
Phi-2	<p>House Bill HB 5, titled "Vote to pass a bill that prohibits abortions after 15 weeks with no exceptions for rape, incest, or human trafficking," aims to restrict access to abortion in Florida. Here are the detailed highlights of the bill and potential winners and losers:</p> <p>Highlights:</p> <ol style="list-style-type: none"> Prohibits abortions after 15 weeks of pregnancy, with no exceptions for rape, incest, or human trafficking <p>Potential Winners:</p> <ol style="list-style-type: none"> Pro-life advocates: The passage of this bill is a victory for pro-life advocates who are passionate about protecting the rights of unborn babies. It reflects their efforts to restrict access to abortion and promote the value of human life from conception until birth. Parental rights: This bill supports parental rights by allowing parents or guardians to be notified before an abortion is performed on a minor, ensuring that they have the opportunity to make informed decisions about their child's healthcare. <p>Potential Losers:</p> <ol style="list-style-type: none"> Women seeking abortions: The passage of this bill limits access to safe and legal abortion for women who are 15 weeks pregnant. It may force them to carry a pregnancy to term or seek unsafe alternatives if they cannot travel out of state. Healthcare providers: This bill places additional restrictions on healthcare providers, requiring them to adhere to specific guidelines when performing abortions after 15 weeks. This may impact their ability to provide comprehensive and timely care to patients. Rape and incest victims: The bill does not make any exceptions for victims of rape or incest, which can have serious implications for those who are already vulnerable. Human trafficking victims: Similarly, the bill's lack of exceptions for rape and incest can have serious implications for members of the LGBTQ+ community, who may be more likely to experience sexual violence and its aftermath. Low-income individuals and families: The bill's restrictions on abortion access may disproportionately impact low-income individuals and families, who may have limited access to healthcare resources and support. <p>Overall, the passage of House Bill HB 5 has significant implications for reproductive rights in Florida, and highlights the ongoing debate around access to safe and legal abortion in the United States.</p>

Table 9: Abortion Bill: Phi-2 generated bill highlights.

Title	Prohibits Abortions After 15 Weeks
Synopsis	A vote to pass a bill that prohibits abortions after 15 weeks with no exceptions for rape, incest, or human trafficking.
Llama-2	<p>House Bill HB 5, also known as the "Pro-Life Act," aims to prohibit abortions after 15 weeks of gestation in Florida. Here are some potential winners and losers of this bill:</p> <p>Highlights:</p> <ol style="list-style-type: none"> 1. N/A <p>Potential Winners:</p> <ol style="list-style-type: none"> 1. Anti-abortion groups: The bill would greatly restrict access to abortion, which is a fundamental right guaranteed by the US Constitution. This would be a huge victory for anti-abortion groups, who have long sought to limit or ban abortion in Florida and across the country. 2. Unborn children: By prohibiting abortions after 15 weeks, the bill would save the lives of thousands of unborn children who might have been terminated through abortion. This could be seen as a moral victory for those who believe that every life is precious and should be protected. 3. Pro-Life lawmakers: The passage of HB 5 would be a major achievement for pro-life lawmakers in Florida, who have been working tirelessly to restrict abortion access. It would demonstrate their commitment to protecting the unborn and upholding their pro-life beliefs. 4. Religious groups: Many religious groups, particularly those with strong anti-abortion stances, would likely view the passage of HB 5 as a moral victory. They might see it as a way to uphold their religious beliefs and values, which may prioritize the protection of human life. <p>Potential Losers:</p> <ol style="list-style-type: none"> 1. Women's rights groups: The passage of HB 5 would be a significant blow to women's rights groups in Florida, who have long advocated for reproductive freedom and access to safe and legal abortion. They may see this bill as an attack on their constitutional rights and a violation of the autonomy of women to make decisions about their own bodies. 2. Obstetricians and gynecologists: The ban on abortions after 15 weeks would directly impact obstetricians and gynecologists, who may lose business and face legal challenges if they continue to provide abortion services beyond that gestation limit. 3. Women's health clinics: Women's health clinics that provide abortion services could be forced to close their doors or risk being sued for violating the new law. This would result in the loss of access to reproductive healthcare for thousands of women in Florida. 4. Pro-Choice lawmakers: The passage of HB 5 would be a significant defeat for pro-choice lawmakers who have been working to protect and expand access to safe and legal abortion. They may see this bill as an attack on their constituents' rights and a violation of the US Constitution. <p>It is important to note that HB 5, or any other anti-abortion legislation, is likely to face legal challenges and could be blocked by court orders before it can be implemented. Additionally, there may be political fallout for lawmakers who support such bills, as they may lose votes and face backlash from constituents who value reproductive freedom.</p>

Table 10: Abortion Bill: Llama-2 generated bill highlights.

public positions of state legislators, which are more reliable, official, and indicative of their preferences than other methods, according to our ablation studies. According to our small-scale data collection and analysis, state legislators have distinct patterns of Twitter usage compared to other federal-level political actors. Firstly, state legislators tweet less frequently and tend to focus on personal or non-political topics rather than policy issues or partisan debates. Secondly, they do not interact much with their followers or other politicians on the platform, indicating a low level of engagement and dialogue. Thirdly, our attempt to use Twitter data to complete missing profiles for some state legislators did not yield significant improvement in our results.

Incorporating Interest groups context. Our framework uses interest group (IG) data for analyzing the state legislative process. Our framework extends the traditional political science research that mostly focuses on the impact of interest groups on Federal legislation. For instance, [Chand and Schreckhise 2013](#) investigated the criteria that IGs use to rate US senators according to their policy views and votes. The authors use a simple statistical model to demonstrate that IGs take into account senators' ideology, party, seniority, committee, and state characteristics. [Jackson and Kingdon 1992](#) examine the role of ideology and IG scores in shaping roll-call voting in Congress. The authors estimate the ideological positions of legislators and IGs with a Bayesian model, and find that both factors influence voting behavior, especially for moderate legislators. [Gilens and Page 2014](#) assess how US policies align with different citizens' preferences. The authors compare average citizens, economic elites, and interest groups' opinions and outcomes. They show that elites and groups affect policies, while citizens have little impact.