Measuring Delegation and Discretion in Democratic Political Institutions with Artificial Intelligence*

Jason Anastasopoulos
University of Georgia†

Jie (Jason) Lian
University of Georgia
Harvard Kennedy School‡

December 3, 2024

Abstract

We introduce a novel framework for measuring delegation and discretion in democratic political institutions using large language models (LLMs). Delegation, defined as the transfer of policy-making authority from principals to agents, requires identification of both the agents receiving authority and the constraints imposed on their actions. Building on existing literature, we combine natural language processing and machine learning techniques to analyze legal texts more effectively than traditional methods. The proposed LLM-based pipeline enhances delegation measurement by processing entire documents, visualizing decision-making processes, and accommodating diverse legal contexts. Empirical applications focus on the U.S. and the EU, highlighting the model's adaptability.

Introduction

Delegation of powers entails the transfer of authority from politicians, who constitutionally hold policy-making power, to an agent or a group of agents, with the scope of their authority defined by the provisions in enabling legislation. This concept involves two critical components that any quantitative approach

*This draft is incomplete. Please do not cite without permission.

†Associate Professor of Public Administration and Policy and Statistics

‡Democracy Fellow, Harvard Kennedy School of Government. PhD Candidate in Political Science and International Affairs, MS candidate in Artificial Intelligence, MS candidate in Statistics.

to measuring delegation must address: (1) identifying the agent or agents who gain policy-making authority through the delegation and (2) outlining the constraints or conditions that govern the exercise of this authority. For example, a law might delegate authority to an environmental agency to establish rules and regulations for wildlife protection but require public consultation as a condition. While delegation is a significant aspect of legislation, it is not the only type of law passed by governments. Consequently, researchers have often employed strict selection criteria to focus on a limited subset of laws, facilitating the application of resource-intensive coding frameworks.

Earlier work on delegation was predominantly focused on theoretical models (see eg. [Epstein and Ohalloran, 1999, Bendor et al., 2001, Volden, 2002, Huber and McCarty, 2004]) while recent research has shifted towards an empirical understanding of delegation by employing methods such as natural language processing, computational linguistics, and machine learning. [Ash et al., 2020, Anastasopoulos and Bertelli, 2020].

As the result of context–specific data availability issues, research on measuring delegation in the United States has employed computational linguistics Ash et al. [2020] while research on measuring delegation in the European Union has focused on the use of machine learning methods [Anastasopoulos and Bertelli, 2020]. Each of these existing approaches has strengths and weaknesses that limit their ability to measure delegation effectively in specific contexts, such as the US or EU. In this paper, we leverage large language models (LLMs), which use generative artificial intelligence (AI), to develop a unified framework for measuring delegation across various democratic political institutions. This framework combines natural language processing and machine learning methods to enhance accuracy and applicability.

The paper is structured as follows: First, we review the current approaches for measuring delegation. We then demonstrate how large language models (LLMs) can integrate the strengths of these approaches to develop a more general and interpretable method for measuring delegation and discretion. This new method can be applied to a wider range of delegation measurement problems.

Next, we propose a LLM-based pipeline to measure delegation from legal documents. By combining fine-tuned LLMs and attention visualization techniques, our pipeline offers four significant advantages in delegation measurement from text data. First, a fine-tuned LLM forms the backbone of our pipeline. This model leverages the general LLM's strong natural language comprehension while being specifically tailored to process legal documents, enhancing its ability to detect key signals of delegation. Second, compared with existing computational methods in measuring delegation from legal documents, the proposed method is not restrained by the structure or length of the input texts in measurement.

For instance, the computational linguistics method proposed by Ash et al. [2020] largely relies on sentence-level text in delegation measurement. In contrast, our method can process sentences, paragraphs, or entire documents, allowing for a more comprehensive capture of subtle delegation signals. Third, our pipeline enables visualization of the key information the fine-tuned LLM uses to detect delegation in legal documents. This has two benefits. It allows researchers to monitor the LLM's decision-making process, aiding in performance evaluation and further model development. Additionally, the identified key phases provide valuable samples, allowing further research on questions like how legal texts indicate different levels of delegation. Finally, our pipeline's flexible structure allows integration with various pre-trained LLMs, enabling researchers to utilize the most advanced language models available.

Measuring Delegation, Authority and Constraint in Legal Documents

Of primary concern to scholars of political institutions is estimating the amount of discretion granted to an agent by a principal. Epstein and O'Halloran (1999) were among the first to systematically measure discretion in the American context and practically all scholarship, at least in political science, have used their definitions as a guide to measure discretion, delegation and constraint [Franchino, 2004, Anastasopoulos and Bertelli, 2020, Ash et al., 2020]. As a result, we also adopt their conventions to measure delegation of authority to an agent, constraints on that delegated authority, and the overall discretion granted from principals to agents which combines delegation and constraint into a discretion index.

In Epstein and OHallorans framework, *delegation and constraint ratios* are quantitative measures used to analyze how legislative bodies delegate authority to executive agencies and the conditions or constraints placed on this delegation. These ratios provide insight into the level of discretion granted to agents (e.g., executive agencies) versus the control retained by principals (e.g., the legislature).

The *delegation ratio* measures the extent to which authority is delegated to an agent. It reflects the proportion of a legislative act that transfers policy-making or implementation authority to another entity. This is calculated as:

$$\Delta_i = \frac{D_i}{P_i}, \tag{1}$$

where Δ_i is the delegation ratio for law i, D_i is the number of provisions in the law that explicitly delegate authority to an agent, and P_i is the total number of provisions in the law. A higher delegation ratio indicates that a larger share of the law delegates authority, signifying greater reliance on the agent to carry out policy.

The *constraint ratio* measures the extent to which the delegated authority is subject to oversight or limitations, as specified in the law. It quantifies the presence of rules or mechanisms designed to restrict the discretion of the agent. This is calculated as:

$$C_i = \frac{C_i}{R_i}, \tag{2}$$

where C_i is the number of categories of restraints used in law i, and R_i is the total possible categories of constraints considered. Constraints might include requirements for mandatory reporting to the legislature, public consultation, limits on budgetary authority, or detailed procedural rules. A higher constraint ratio indicates that the agents authority is more restricted and subject to checks, thereby limiting its discretion.

To combine these two concepts, Epstein and OHalloran introduce a *discretion index*, which accounts for both the level of delegation and the constraints placed on that delegation. The discretion index adjusts the delegation ratio by factoring in the constraint ratio, as follows:

$$\delta_i = \Delta_i - [C_i \times \Delta_i], \tag{3}$$

where δ_i is the discretion index for law i, Δ_i is the delegation ratio, and C_i is the constraint ratio. The discretion index measures the remaining latitude or freedom the agent has to act after accounting for the constraints.

From a measurement perspective, estimating these quantities across contexts poses a number of empirical challenges. We will specifically discuss challenges within the American and the EU perspective since these areas have been the primary research focus within political science.

Delegation in the American context is the most straightforward case since delegation tends to flow only from the Congress (principal) to executive agencies (agent). This, in the American case we estimate delegation of authority from

Congress to executive agencies.

In the European Union, however, the process is more complicated since the European Commission (EC) drafts legislation which is reviewed and amended by the European Parliament and the Council of the EU. Thus while the European Parliament and the Council of the EU can be viewed, ultimately, as the principals with the EC and national administrations as agents, in reality the European Commission serves a dual role as both as principal and agent during the legislative process. For the EU case, we follow Franchino (2004) who estimates delegation ratios as a function of delegation to the European Commission and the national administrations separately.

Estimation of constraints are somewhat more straightforward since Epstein and O'Halloran (1999) identified 7 categories of constraints that can be readily identified in both the US and EU cases. These categories include procedural requirements, reporting requirements, legislative oversight, time limits, budgetary constraints, substantive policy constraints and judicial review. From a measurement perspective, the constraint ratio is calculated by identifying the presence of any of these restraint categories present in the legislation and dividing by the total categories of restraints. For instance, if time limits and budgetary restraints were present in a law, the constraint ratio would be calculated as $2/7 = 0.286$.

Algorithm Design & Experiments

In this research, we aim to develop an effective framework for training a Large Language Model (LLM)-based classifier to accurately detect instances of authority delegation, constraint and discretion in legal documents. Before delving into the methodology, it is crucial to address the unique challenges associated with leveraging LLMs for social science tasks, particularly for training classifiers that address nuanced conceptual frameworks. First, the concepts commonly studied by social scientistssuch as authority delegation, the focus of this researchoften describe complex social interactions. These interactions involve dynamic and potentially ambiguous relationships between multiple entities. This complexity contrasts sharply with tasks such as sentiment analysis or entity recognition, where LLMs have demonstrated substantial success. Unlike these more straightforward tasks, the identification of social science constructs frequently requires an understanding of intricate and contextualized interactions.

Second, the interpretation of social science concepts is rarely static. The meaning of specific constructs often evolves over time as societal norms and scholarly interpretations shift . This temporal variability adds another layer of complexity to designing classifiers capable of handling these fluid definitions effectively. Third, social science research frequently involves imbalanced datasets. For example, in the task of detecting instances of power delegation, positive cases (indicating power delegation) are typically vastly outnumbered by negative cases (indicating no power delegation). This imbalance makes classifiers particularly vulnerable to Type I errors, which pose a more significant concern in this context than Type II errors. Given these constraints, we argue that the conventional recipe to LLM training commonly employed in computer science may not be directly applicable to social science tasks. Each social science task likely requires iterative experimentation and refinement to identify an optimal training framework. Addressing these challenges demands methodological flexibility and domain-specific adjustments to ensure robust and meaningful results.

To date, the most commonly employed methods for developing efficient LLM-based classifiers include in-context learning, transfer learning, and fine-tuning. Additionally, different foundation models exhibit varying levels of effectiveness in performing classification tasks, highlighting the importance of model selection in achieving optimal results. In this research, we conduct a systematic evaluation of the efficiency of various combinations of these methods in the context of the delegation classification task. Based on the experimental outcomes, we identify the most effective approach and use it as the final framework to train a delegation classifier on a larger dataset.

To begin, we incorporated three foundation models into our experiments: BERT, Mistral 7B v0.1, and LLaMa 3.2 3B. For prompt construction, we experimented with few-shot prompts and dynamic in-context learning in the reference stage, utilizing the Retrieval-Augmented Generation (RAG) algorithm to identify the most similar cases. For fine-tuning, given the limitations of computational resources, we focused on two parameter-efficient fine-tuning methods: LoRA (Low-Rank Adaptation) and DoRA (Decoupled Rank Adaptation).

In the experiments, we randomly sampled 5,000 observations from the original dataset in Anastasopoulos and Bertelli [2020] which was compiled from information provided by Franchino [2004]. The data was divided into three subsets: 60% for training, 16% for testing, and 24% for evaluation. To address potential bias caused by dataset imbalance, we performed random upsampling of the positive cases in the training set. All experiments were conducted on a single Nvidia RTX A5000 GPU machine. The results of the experiments are presented in Table 1.

In these examples, we first focus on replicating results from Anastasopoulos and Bertelli [2020] by predicting delegation of authority to the European Commission and national administrations using LLMs. We plan to extend these analyses using trained models to predict delegation of authority in the American context and will also extend these analyses to predicting constraint and discretion in both contexts as well.

Table 1: Experiment Results

Foundation Model	Model Scheme	Transfer Learning	Prompt Engineering	Fine Tuning	Knowledge-based Inference	Accuracy	F1 Score	Positive Precision
RoBERTa	Bidirection Model			Full Parameter Fine Tuning		0.46	.57	0.18
BERT	Bidirection Model	Domain Adaptation		Full Parameter Fine Tuning		0.76	0.85	0.34
GPT 4o (Only on eval set)	Causal Inference Model		Chain of Thoughts		RAG-based Inference	0.84	0.91	0.29
LLAMA 3.2 3B Instruct	Sequential Classification Model			LoRA		0.77	0.86	0.33
LLAMA 3.2 3B Instruct	Sequential Classification Model			LoRA	RAG-based Inference	0.82	0.90	0.20
LLAMA 3.2 3B Instruct	Sequential Classification Model		In Context Learning	LoRA		0.78	0.87	0.33
LLAMA 3.2 3B Instruct	Causal Inference Model			LoRA		0.78	0.87	0.17
LLAMA 3.2 3B Instruct	Causal Inference Model			LoRA	RAG-based Inference	0.81	0.90	0.06
LLAMA 3.2 3B Instruct	Causal Inference Model			DoRA		0.77	0.86	0.32
Mistral 7B v0.1	Sequential Classification Model			LoRA		0.83	0.85	

{

```
# Prompt
def generate_prompt(row):
        line = f"""
### Task:
Classify the Target Text below, which is extracted from Congress
                                 bills, to determine whether it
                                 implies a delegation of powers
                                 based on the Definition. Make
                                 sure to analysis the Target Text
                                 step by step. Return "1" if power
                                  delegation happens and "0"
                                 otherwise.

### Definition:
Power delegation is defined as any major provision that gives
                                 another governmental body the
                                 authority to move policy away
                                 from the status quo.

### Target Text:
{row['Text']}
    """.strip()
        return line
```

}
{

```
# In Context Learning Prompt
def generate_fewshots_prompt(row):
        line = f"""
### Task:
Classify the Target Text below, which is extracted from Congress
                                 bills, to determine whether it
                                 implies a delegation of powers
                                 based on the Definition and
                                 Examples. Make sure to analysis
                                 the Target Text step by step.
                                 Return "1" if power delegation
                                 happens and "0" otherwise.

### Definition:
Power delegation is defined as any major provision that gives
                                 another governmental body the
                                 authority to move policy away
                                 from the status quo.

### Examples of what delegation is:
    The authorization of a new program with some discretionary
                                 powers;
    Discretion to make or modify decision-making criteria;
    Extension of discretionary authority that would otherwise
                                 expire;
```

```
    The creation of a new commission, board, or agency;
    Demonstration projects;
    Grants and loans where the agency determines the size of the
                             award and/or the recipients;
    The right to issue subpoenas;
    The right to bring suit or intervene in an existing suit;
    The right to issue waivers;
    The ability to enter into contracts.

### Examples of what delegation is not:
    Authorizing appropriations or funds for a program;
    Requiring reports, studies, or publication of information;
    The hiring of staff or personnel;
    Transferring delegated authority from one executive branch
                             actor to another without
                             increasing the scope of that
                             authority;
    Evaluations, recommendations, and assessments that do not
                             directly alter policy;
    Audits, which are considered constraints and not delegation to
                             , for instance, the GAO.

### Target Text:
{row['Text']}
    """.strip()
        return line
```

}

Experiments Discussions

The experimental results shown in above table revealed several notable findings. First, when considering common performance metrics such as accuracy and F1 score, larger foundation models like GPT-4, LLaMa, and Mistral outperformed BERT in the classification task. Second, incorporating few-shot prompts during the training phase (in-context learning) yielded a slight improvement in model performance. Meanwhile, leveraging knowledge-based inference with the help of RAG resulted in a significant performance boost. These findings align with empirical results from the computer science literature.

Interestingly, when evaluating precision for positive predictions, the BERT-based classifier with domain adaptation steps demonstrated superior performance compared to the larger models. As discussed earlier, in most text classification tasks within social science research, Type I errors pose a far greater concern than Type II errors. In this context, the precision of the classifierits ability to minimize false positivesis paramount. Thus, for the specific research objectives of this study, the domain-adapted BERT model demonstrates clear superiority over larger LLMs, making it the more suitable choice for this task.

According to existing literature, the superior performance of the BERT-based classifier can be attributed to the structural differences between BERT and most current LLMs, which primarily utilize a decoder-only architecture Anastasopoulos and Bertelli [2020], Ash et al. [2020]. In the decoder-only framework, predictions are made based on the embedding of the last token, which is expected to encapsulate all preceding contextual information. While this design is effective for tasks such as next-token prediction, it may struggle to capture task-specific patterns in the text that are crucial for semantic interpretation, particularly in tasks like the classification of legal provisions Li et al. [2023]. In contrast, BERT's label-supervised architecture, which incorporates bidirectional context and is specifically designed for tasks requiring deep semantic understanding, is better suited for identifying such patterns.

It is important to note that while BERT models demonstrate strong performance on classification tasks, directly fine-tuning a BERT model from a publicly available checkpoint does not yield optimal results. As illustrated in Figure 1, when the BERT checkpoint is fine-tuned directly on the provided dataset, the training loss fails to decrease, indicating that the model is not effectively learning. This issue is further evidenced by the erratic behavior of the evaluation curves, which suggest the model alternates between predicting all positive or all negative labels for the given provisions. In contrast, domain adaptation proves crucial in this scenario. For this research, we utilized a domain-specific checkpoint published by Chalkidis et al. [2020], in which the BERT base model was retrained on a substantial corpus of legal texts. As shown in Figure 2, the training loss demonstrates that the domain-adapted model actively learns during fine-tuning, and the evaluation curves clearly indicate improved performance.

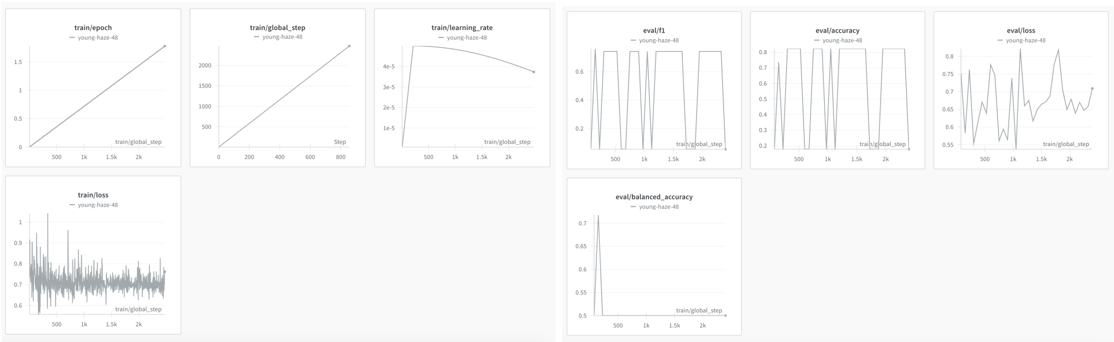

Figure 1: Training & Evaluation Curves for Fine Tuning BERT

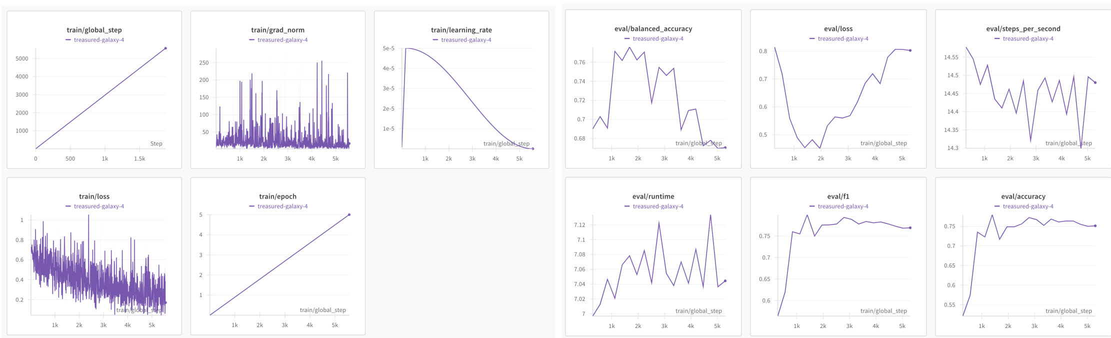

Figure 2: Training & Evaluation Curves for Fine Tuning Domain-Adapted BERT

References

L Jason Anastasopoulos and Anthony M Bertelli. Understanding delegation through machine learning: A method and application to the european union. *American Political Science Review*, 114(1):291–301, 2020.

Elliott Ash, Massimo Morelli, and Matteo Vannoni. Measuring discretion and delegation in legislative texts: methods and application to us states. In *Political Analysis*. Cambridge University Press, 2020. doi: 10.7910/DVN/C13Z0X.

Jonathan Bendor, Ami Glazer, and Thomas Hammond. *Theories of Delegation in Political Science*. Stanford Graduate School of Business, 2001. Working Paper No. 1655.

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

David Epstein and Sharyn Ohalloran. *A transaction cost politics approach to policy making under separate powers*. Cambridge: Cambridge university press, 1999.

Fabio Franchino. Delegating powers in the european community. *British Journal of Political Science*, 34(2):269–293, 2004.

John D. Huber and Nolan McCarty. Bureaucratic capacity, delegation, and political reform. *American Political Science Review*, 98(3):481–494, 2004. doi: 10.1017/S0003055404001283.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. Label supervised llama finetuning. *arXiv preprint arXiv:2310.01208*, 2023.

Craig Volden. A formal model of the politics of delegation in a separation of powers system. *American Journal of Political Science*, pages 111–133, 2002.