

Advancing Ideological Measurement: A Text-Grounded Framework Using LLM-Powered Summarization and Scaling for Diverse Political Arenas

Mitchell Bosley

June 4, 2025

Abstract

This project proposes a novel framework to measure legislator ideology directly from political texts, addressing critical limitations of existing methodologies. Current approaches often rely on specific behavioral data, such as roll-call votes, which are unavailable or uninformative in many political systems characterized by strong party discipline, for instance, Westminster systems. Other methods depend on Large Language Models' (LLMs) pre-existing knowledge, restricting applicability to well-known public figures and established political arenas. There is a significant need for robust, text-grounded, and broadly applicable measures to enhance our understanding of political representation, legislative behavior, and ideological dynamics across diverse contexts.

Our framework introduces an innovative multi-stage, text-grounded process. First, we leverage the advanced natural language understanding capabilities of LLMs to generate issue-specific "structured summaries" from collections of legislators' political communications, such as parliamentary speeches and debate contributions. These summaries are designed to distill complex textual data into concise, comparable representations of legislators' stances and arguments on defined policy dimensions. Second, these structured summaries are then subjected to pairwise comparisons, again facilitated by LLMs, to determine relative ideological positioning. Finally, these pairwise judgments are scaled using the Bradley-Terry model to derive continuous ideological scores with associated uncertainty measures.

The core innovation lies in its data-driven nature—grounding ideological measurement in the specific textual evidence provided, rather than relying on an LLM's opaque prior knowledge or restrictive behavioral datasets. This makes the framework applicable to diverse political arenas and actors, including those less prominent or operating in systems where traditional scaling methods falter. Furthermore, it enables granular, issue-specific ideological scaling directly from the content of political discourse.

The project aims to develop and refine a robust LLM-based structured summarization module specifically tailored for political texts across different issue areas. It will also implement and rigorously validate the subsequent pairwise comparison and Bradley-Terry scaling pipeline. The framework's validity and comparative performance will be demonstrated by applying it to U.S. Congressional debates, comparing outputs to established ideological benchmarks like DW-NOMINATE and CFscores. Its unique utility in overcoming existing methodological gaps will be showcased by applying it to the UK House of Commons, a Westminster system where individual legislator scaling is notoriously challenging.

Expected outcomes include a validated, open-source methodological toolkit, novel datasets of text-derived ideological scores for U.S. and UK legislators, and significant contributions to the understanding of legislative communication, intra-party ideological heterogeneity, and comparative legislative behavior. This research will advance both political science methodology and the responsible application of LLMs in computational social science.

Contents

1	Introduction	1
2	Proposed Research: Methodology and Approach	1
2.1	Conceptual Framework	1
2.2	Phase 1: Development of the LLM-Powered Structured Summarization Module	2
2.2.1	Text Data Acquisition and Preparation	2
2.2.2	Design of Structured Summaries	2
2.2.3	LLM Selection and Prompt Engineering for Summarization	2
2.3	Phase 2: Implementation of Pairwise Comparison and Bradley-Terry Scaling	2
2.3.1	Pairwise Comparison of Summaries	2
2.3.2	Bradley-Terry Scaling	3
2.4	Technical Implementation Plan	3
3	Research Plan, Timeline, and Deliverables	3
3.1	Detailed Research Activities and Timeline	3
3.2	Case Study 1: Validation in the U.S. Congress	4
3.3	Case Study 2: Application in a Westminster System (e.g., UK House of Commons)	4
3.4	Evaluation Plan	4
3.5	Deliverables	5
4	Expected Outcomes, Significance, and Broader Impact	5
4.1	Intellectual Merit & Novelty	5
4.2	Contribution to Political Science	5
4.3	Contribution to Computational Social Science & Text-as-Data	5
4.4	Broader Impacts	5
4.5	Dissemination Plan	6
A	Illustrative Draft Prompts for Structured Summarization	8
B	Illustrative Draft Prompts for Pairwise Comparison	8

1 Introduction

The measurement of legislator ideology is a cornerstone of political science, fundamental to understanding political representation, legislative behavior, policy-making processes, and democratic accountability (Poole and Rosenthal, 1997). However, the diverse and evolving nature of political systems and communication presents ongoing challenges for existing methodologies.

Traditional methods, while valuable, often fall short in specific contexts. For instance, roll-call vote analysis, exemplified by DW-NOMINATE (Poole, 2005), is powerful in systems like the U.S. Congress but depends heavily on the nature of the legislative agenda and the frequency of non-unanimous votes. In parliamentary systems with strong party discipline, such as the UK, roll-call votes offer little insight into individual legislator ideology beyond party affiliation. Similarly, measures based on campaign finance (Bonica, 2014) reflect donor perceptions and are not universally available. This creates a significant gap in our ability to empirically assess individual ideological positioning in many important political arenas.

Beyond behavioral data, text-as-data methods have offered new avenues, but traditional approaches like Wordscores or Wordfish often require large, comparable corpora and can be challenging to interpret for specific ideological dimensions. The advent of Large Language Models (LLMs) presents new opportunities, yet current applications also have limitations. Direct scaling by LLMs can suffer from anchoring biases and unreliability (DiGiuseppe and Flynn, 2025). Pairwise comparison of *individuals* based on an LLM’s pre-existing knowledge, as demonstrated by Wu et al. (2023b), shows promise for well-known figures such as U.S. Senators but is inherently limited to contexts where the LLM possesses extensive prior information about the actors. This dependency restricts broader applicability to less-documented political systems, newly elected officials, or historical figures with limited digitized records. Other work has focused on pairwise comparisons of short texts or augmented short texts (DiGiuseppe and Flynn, 2025; Wu et al., 2023a), which differs from our focus on distilling long-form primary source documents like speeches.

There is a clear and pressing need for a methodology that can derive nuanced, issue-specific ideological measures directly from legislators’ textual communications. Such a method should be applicable across diverse political contexts, including those with strong party discipline where voting records are uninformative for individual scaling. Crucially, it must operate effectively even when LLMs lack extensive prior knowledge about the specific legislators being analyzed. This project proposes such a framework, leveraging the strengths of LLMs for text processing while grounding the analysis firmly in the provided textual data.

2 Proposed Research: Methodology and Approach

Our proposed framework integrates LLM-driven text summarization and pairwise comparison with statistical scaling via the Bradley-Terry model to measure legislator ideology. The core idea is to transform complex and potentially lengthy textual data, such as speeches, into structured, comparable summaries focused on specific issue areas. These summaries then form the basis for robust ideological measurement.

2.1 Conceptual Framework

We conceptualize ideology as a latent disposition that legislators express through their verbal and written communications on various policy issues. Our method operationalizes this by first distilling these expressions into structured summaries that capture key ideological indicators such as stance, arguments, and rhetoric. Pairwise comparisons of these summaries then allow for a relative ordering of legislators along an ideological continuum, which is subsequently quantified using the Bradley-Terry model.

2.2 Phase 1: Development of the LLM-Powered Structured Summarization Module

2.2.1 Text Data Acquisition and Preparation

We will collect publicly available debate transcripts from two primary sources. First, for the U.S. Congress, transcripts from the Congressional Record for selected sessions of the U.S. House of Representatives and Senate (e.g., 115th-117th Congresses) will be gathered, focusing on debates on key legislation. Second, for the UK Parliament, transcripts from Hansard for selected parliamentary sessions of the UK House of Commons (e.g., post-2015 general elections) will be collected, focusing on debates on major government bills and opposition day debates. All texts will be preprocessed, including cleaning and speaker identification, and segmented into manageable units for summarization. Segmentation will primarily be by individual legislator’s speaking turns within a specific debate or on a particular bill or topic.

2.2.2 Design of Structured Summaries

Issue areas, for example, economic policy, healthcare, immigration, and environmental policy, will be defined based on salient policy debates within each legislative context and existing political science typologies. For each issue area, a template for the structured summary will be developed. This template will guide the LLM to extract several key components: the legislator’s primary stance on the core issue or bill; the key arguments or justifications presented; specific policies or solutions advocated or opposed; the rhetorical style or framing used, such as appeals to specific values or targeting of groups; and the expressed sentiment towards the issue or opposing viewpoints.

2.2.3 LLM Selection and Prompt Engineering for Summarization

We will explore state-of-the-art LLMs, for instance, models from the GPT series, Claude series, and potentially leading open-source models like Llama if their performance is comparable for this task, for their summarization capabilities. Prompts will be iteratively developed and refined using a pilot set of diverse debate excerpts. This process will involve clear instructions outlining the task and the desired structured output format, potentially incorporating few-shot examples to guide the LLM, and techniques to encourage focus on relevant ideological content while discouraging irrelevant information. An example draft prompt for summarizing a stance on a hypothetical healthcare bill is:

```
"You are an expert political analyst. Given the following debate excerpt by [Legislator Name] on [Date] regarding the [Specific Healthcare Bill Name]: [TEXT EXCERPT OF LEGISLATOR’S SPEECH]
```

```
Please provide a structured summary covering the following points regarding their expressed views on this healthcare bill:
```

1. Overall Stance (e.g., Support, Oppose, Critical Support, Amend):
2. Main Arguments For/Against (List 2-3 key points):
3. Proposed Policies/Alternatives Mentioned (if any):
4. Key Concerns Highlighted:
5. Target Beneficiaries/Groups Harmed (as per their argument):"

Quality control will involve manual review of a subset of summaries for accuracy, completeness, and consistency.

2.3 Phase 2: Implementation of Pairwise Comparison and Bradley-Terry Scaling

2.3.1 Pairwise Comparison of Summaries

For each issue area, the structured summaries generated for each legislator will be compared pairwise. We will use a capable LLM, potentially the same model used for summarization or one optimized for

comparative reasoning, for this task. Prompts will be carefully designed. An example prompt is:

```
"Consider the following two structured summaries on
[Specific Issue Area, e.g., environmental regulation policy]:
```

```
Legislator A's Summary:
```

```
Stance: [Stance A]
```

```
Arguments: [Arguments A]
```

```
Policies: [Policies A]
```

```
Concerns: [Concerns A]
```

```
Legislator B's Summary:
```

```
Stance: [Stance B]
```

```
Arguments: [Arguments B]
```

```
Policies: [Policies B]
```

```
Concerns: [Concerns B]
```

```
Based on these summaries, which legislator expresses a more
[ideological pole, e.g., 'pro-environmental regulation'] stance?
Respond with only 'Legislator A', 'Legislator B',
or 'Indistinguishable/Tie'."
```

Ties will be handled either by assigning 0.5 wins to each in a standard Bradley-Terry model or by employing a model variant like the Bradley-Terry-Davidson model (Davidson, 1970) that explicitly accommodates ties. We will generate a sufficient number of comparisons for robust scaling, ensuring each legislator's summary is part of at least N comparisons.

2.3.2 Bradley-Terry Scaling

The outcomes of the pairwise comparisons will be used to fit a Bradley-Terry model. We plan to use a Bayesian implementation, likely via the `brms` package in R (Bürkner, 2017), to estimate latent ideological scores for each legislator on each issue area. This approach allows for the natural incorporation of uncertainty through credible intervals. Priors will be weakly informative. Model convergence will be assessed using standard diagnostics such as \hat{R} and effective sample size. Scores will be normalized, for example, to a mean of 0 and standard deviation of 1, or to a [0,1] range, for interpretability and comparison across issue areas.

2.4 Technical Implementation Plan

The project will primarily utilize Python for LLM API interaction and text preprocessing. R will be used for statistical modeling, specifically `brms` for Bayesian Bradley-Terry models, and for generating visualizations. Version control will be managed using Git and GitHub. Access to institutional High-Performance Computing (HPC) resources will be leveraged for large-scale summary generation and potentially for model fitting if required.

3 Research Plan, Timeline, and Deliverables

3.1 Detailed Research Activities and Timeline

The project is planned for a duration of approximately 24 months. The initial phase, Months 1-3, will involve finalizing the literature review, addressing ethical considerations, detailing the U.S. data acquisition plan, securing all necessary LLM API access and computational environment setup, and conducting initial pilot summarization prompt engineering on a small U.S. dataset. Phase 1a, spanning Months 4-7,

will focus on the full-scale development and rigorous refinement of the structured summarization module using U.S. Congressional debate transcripts, culminating in the generation of all U.S. legislator-issue summaries and a quality control assessment.

Phase 2a, from Months 8-10, includes the implementation of the pairwise comparison module for U.S. summaries, subsequent Bradley-Terry scaling, rigorous validation against DW-NOMINATE, CF-scores, and other relevant benchmarks, and an initial analysis of U.S. results. A mid-project review and preparation for the UK case study will occur in Months 11-12 (Phase 3), involving an internal review of U.S. findings, methodology refinement if needed, detailed UK data acquisition from Hansard, and adaptation of summarization prompts for the UK context.

Phase 1b, covering Months 13-16, will see the full-scale generation of UK MP-issue summaries, along with a quality control assessment. Phase 2b, from Months 17-19, will involve pairwise comparison of UK summaries, Bradley-Terry scaling, and analysis of UK results, with a focus on intra-party dynamics and comparison with U.S. findings. The final phase, Months 20-24 (Phase 4), will be dedicated to manuscript preparation for journal submissions, conference presentations, creation of an open-source toolkit and documentation, and final project reporting.

3.2 Case Study 1: Validation in the U.S. Congress

The objectives for this case study are twofold: first, to assess the convergent validity of our text-derived ideological scores by comparing them against established benchmarks such as DW-NOMINATE and CF-scores; and second, to explore whether our method captures additional nuances of ideology expressed in debate text not fully reflected in behavioral measures. We hypothesize that ideological scores derived from debate summaries will show strong, statistically significant positive correlations with first-dimension DW-NOMINATE scores. Furthermore, we expect scores to correlate significantly with CF-scores, reflecting an alignment with donor-perceived ideology. Qualitative analysis is anticipated to reveal instances where our text-based scores diverge meaningfully from benchmarks, attributable to specific rhetorical strategies or issue framings in debates.

3.3 Case Study 2: Application in a Westminster System (e.g., UK House of Commons)

The objectives for this application are to demonstrate the utility of our framework in a system where strong party discipline limits the informativeness of roll-call votes for individual MP scaling, and to uncover and quantify potential intra-party ideological heterogeneity based on MPs' debate contributions. Key research questions include: To what extent can meaningful and coherent ideological scales be derived for individual UK MPs based on their debate summaries across different issue areas? Does the method reveal significant ideological variance within major UK political parties on specific policy issues? And how do these text-derived ideological positions relate to known party factions or public stances of MPs?

3.4 Evaluation Plan

The success of the project will be evaluated through a multi-faceted approach. Quantitative validation, particularly in the U.S. case, will involve Pearson correlations with benchmark scores and regression models predicting benchmarks using our scores, alongside model fit statistics for Bradley-Terry models. Qualitative assessment for both cases will include examining the face validity of generated structured summaries and derived ideological scales through review of legislators placed at ideological extremes and moderates, cross-referenced with expert knowledge and media commentary. Methodological robustness will be tested via sensitivity analyses regarding LLM choice, variations in prompt structure, and the number of pairwise comparisons. For the UK case, comparative utility will be judged by its ability to generate scales where existing methods are insufficient and by identifying substantively interesting patterns like intra-party dissent.

3.5 Deliverables

The project aims to produce several key deliverables. These include a publicly available, documented open-source codebase, for instance on GitHub, for implementing the entire methodological pipeline. Anonymized or aggregated datasets of structured summaries and derived ideological scores for both the U.S. Congressional and UK Parliamentary case studies will be made available via a public data repository such as Dataverse or OSF, where permissible. We anticipate at least one peer-reviewed journal article detailing the full methodology and the U.S. validation study, targeted at journals like *Political Analysis* or the *American Journal of Political Science*. Another peer-reviewed publication will present the application and findings from the UK Westminster system study, with targets such as the *British Journal of Political Science* or *Legislative Studies Quarterly*. Findings will also be disseminated through presentations at major national and international academic conferences, and a project website will provide an overview, updates, and links to resources.

4 Expected Outcomes, Significance, and Broader Impact

4.1 Intellectual Merit & Novelty

The primary intellectual merit of this project lies in the development of a novel, robust, and broadly applicable methodological framework for measuring legislator ideology directly from textual data. Its novelty stems from the explicit two-stage LLM process, first generating structured summaries of long-form text tailored to specific issue areas, and then using these summaries for pairwise comparison. Furthermore, its data-grounded nature enables application to political actors and arenas not extensively covered in LLM training data, thus overcoming a key limitation of some existing LLM-based scaling methods. This framework also offers the capacity for granular, multi-dimensional ideological scaling based on the direct content of political communication.

4.2 Contribution to Political Science

This project will provide political scientists with new tools and empirical insights for a deeper understanding of several areas. It will enhance the study of legislative behavior by moving beyond roll-call votes to analyze ideology as expressed in deliberation and rhetoric. In terms of political representation, it allows for an assessment of how legislators articulate positions on behalf of constituents. For intra-party dynamics, it offers a method to quantify ideological heterogeneity within parties, especially in systems with strong party discipline. Finally, in comparative politics, it enables more consistent ideological comparison across different types of political systems by relying on a common data source: text.

4.3 Contribution to Computational Social Science & Text-as-Data

Methodologically, our work will showcase a sophisticated and principled multi-stage application of LLMs for extracting nuanced, scalable measures from complex, unstructured political text. It contributes by demonstrating a workflow for effectively reducing information from long texts for comparative analysis using LLMs. It also provides a template for combining LLM-based text processing with established psychometric scaling models, and offers insights into the capabilities and challenges of using LLMs for fine-grained analysis of political discourse.

4.4 Broader Impacts

Beyond academia, this research has the potential to inform public discourse by providing more nuanced, text-grounded data for journalists and civic organizations analyzing political positions. It can enhance transparency by offering a way to systematically track how legislators articulate their stances on key issues over time. The open-source tools and datasets developed will also serve as valuable resources for training students and future researchers in cutting-edge computational methods.

4.5 Dissemination Plan

Findings will be actively disseminated through multiple channels. We will submit manuscripts to leading peer-reviewed journals in political science, such as the *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *British Journal of Political Science*, and *Political Analysis*, as well as relevant computational social science outlets. We plan to present our research at major national and international academic conferences, including those organized by the American Political Science Association (APSA), Midwest Political Science Association (MPSA), European Consortium for Political Research (ECPR), European Political Science Association (EPSA), the Society for Political Methodology (PolMeth), International Conference on Computational Social Science (IC2S2), and workshops associated with EMNLP/ACL. Open-source code will be released on GitHub, and datasets will be shared via a public repository. A dedicated project website will provide an overview, updates, links to resources, and publications.

References

- Bonica, A. (2014). Mapping the ideological marketplace. *American Journal of Political Science*, 58(2):367–386.
- Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.
- Davidson, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328.
- DiGiuseppe, M. and Flynn, M. (2025). Scaling open-ended survey responses using LLM-paired comparisons. Working Paper, Version dated January 25, 2025. OCR'd from provided image.
- Poole, K. T. (2005). Spatial models of parliamentary voting.
- Poole, K. T. and Rosenthal, H. (1997). Congress: A political-economic history of roll call voting.
- Wu, P. Y., Nagler, J., Tucker, J. A., and Messing, S. (2023a). Concept-guided chain-of-thought prompting for pairwise comparison scoring of texts with large language models. Version arXiv:2310.12049v2 [cs.CL] 24 Jan 2025.
- Wu, P. Y., Nagler, J., Tucker, J. A., and Messing, S. (2023b). Large language models can be used to estimate the latent positions of politicians.

A Illustrative Draft Prompts for Structured Summarization

An illustrative prompt for economic policy debates might be:

"Given the following speech excerpt by [Legislator] on the [Budget Bill]:
[TEXT EXCERPT]

Summarize their position covering these aspects:

1. Stance on overall government spending levels:
2. View on specific taxation policies mentioned:
3. Arguments regarding the national debt or deficit:
4. Preference for economic intervention versus laissez-faire approaches:"

B Illustrative Draft Prompts for Pairwise Comparison

An illustrative prompt for comparing summaries on social welfare policy could be:

"Comparing the following two summaries on [Social Welfare Policy]:

Summary X from Legislator X: [Summary X details focusing on stance, arguments]

Summary Y from Legislator Y: [Summary Y details focusing on stance, arguments]

Which summary indicates a stronger preference for universal social programs as opposed to means-tested approaches? Respond 'Summary X', 'Summary Y', or 'Tie'."