Journal of Political Institutions and Political Economy, 2025, 6: 407–426

Can We Measure Legislative Complexity with LLMs?

Austin Bussing[1], Nicholas O. Howard[2], and Joshua Y. Lerner[3*]

[1] *Department of Political Science, Trinity University, San Antonio, TX 78212-7200, USA*
[2] *Department of Political Science, Concordia College, Moorhead, MN 56562, USA; nhoward1@cord.edu*
[3] *Senior Research Methodologist, NORC at the University of Chicago, Chicago, IL 60603, USA*

ABSTRACT

The complexity of legislative language is of theoretical importance to many substantive questions about legislative politics. However, most existing measures of bill complexity are either generated at the broad issue level and applied to individual bills, or they are reliant on a simple metric like length. In this paper, we apply a pairwise comparison framework to the measurement of complexity in legislative texts. We compare the results of a Bradley-Terry model (Bradley and Terry, 1952) fit on pairwise comparisons made by human coders with the results of the same model fit on comparisons made by a large language models (LLMs). There is a moderately high level of agreement between human coders and the LLMs, and the relationships between observable text features and the underlying trait of complexity are similar in comparisons made by human coders and by the LLMs. Our work demonstrates that, with researcher-selected bridging texts and carefully designed prompts, LLMs can be used to measure complexity in legislative texts.

Keywords: Congress; text analysis; LLM; complexity; legislatures

*We thank Jeff Jenkins and participants at the Artificial Intelligence and the Study of Political Institutions Conference at the Sol Price School at USC, as well as participants at the 2025 Southern Political Science Association Conference. We especially thank Jason Anastasopoulos and Doug Rice for their helpful comments and advice.

Online Appendix available from:
http://dx.doi.org/10.1561/113.00000130_app
ISSN 2689-4823; DOI 10.1561/113.00000130

Introduction

Measuring the complexity of language used to communicate policy-relevant information is central to answering important questions spanning the realms of law, policy, and politics. The complexity of a given policy affects its diffusion across jurisdictions (Makse and Volden, 2011), and issue complexity is also a relevant consideration in legislative decisions to delegate policymaking authority to the executive branch (Epstein and O'Halloran, 1999). Scholars of direct democracy at the state and local level are interested in understanding the complexity of language used for ballot initiatives (e.g., Reilly and Richey, 2011). Additionally, conceptualizing and measuring the complexity of written policy will become especially important to analyzing judicial decision-making in a post-*Chevron* era in which courts may take a more active role in resolving statutory ambiguities.

For scholars, policy complexity can be seen as an independent variable that helps shape institutions and, alternatively, as a dependent variable that is an output of institutional dynamics. On the one hand, for Krehbiel (1992), the complexity of policy issues explains the existence and operation of the specialized legislative committee system. Relatedly, committee jurisdictions evolve in part as a response to changes in the complexity of issue areas (Baumgartner *et al.*, 2000), and committee chairs retain power and influence in an increasingly leadership-dominated chamber because of their specialized policy expertise (Curry, 2019). Each of these perspectives treats the inherent complexity of policy as a determinant of institutional dynamics. Alternatively, the complexity of policy—as reflected in written legislation—can be understood as a function of the legislative institutions that produce it. Variations in the complexity of legislative language arise from the institutional features of Congress itself and from strategic actors pursuing policy and electoral goals within that institutional context.

Regardless of whether political scientists are interested in policy complexity as a dependent or an independent variable, a consistent text-based measure of complexity would help answer many important substantive questions in the field. However, existing measures of policy complexity are built upon work which centers human labor. Benoit *et al.* (2019) build a latent model of complexity based around the work of crowd-sourced human coders. Similarly, Senninger (2023) uses comparisons of texts by human coders to build Bradley-Terry comparison models of legislative complexity. These approaches, however, come with all of the shortfalls of human labor, including the inability to parallel process (Jones, 2001) and provide long-term attention to specific tasks (Simon, 1985).

Our work in this paper builds upon previous work on complexity in political texts by Senninger (2023) and Benoit *et al.* (2019) in two ways. First, we focus specifically on legislative language produced by the U.S. Congress. This

requires directed human cognition, as samples are written in legalistic language. Second, we introduce large language models (LLMs) as a potential way to overcome human limitations. This pursuit leads us to our core question of whether LLMs can understand legislative complexity and what limits this new technology faces.

In what follows, we first address how scholars in the applied field of legislative politics understand complexity in both substance and effect. We then discuss how complexity has been measured in extant research before moving to our research design and findings. These findings demonstrate that, while LLMs can mirror human behavior in understanding legislative complexity, such models face similar limitations to humans and existing scaling exercises. We finish with a discussion of what the method and findings mean for the current understanding of legislative design and delegation, as well as ideas for future exploration.

Legislatures, Proposals, and Legislative Complexity

While the study of complexity in various forms of political communications—speeches, ballot initiatives, etc.—forms a helpful foundation for the study of complexity in legislative language, the distinctiveness of legislative language must also be acknowledged. Legislative language, or the actual language that constitutes policy proposals in a legislative body, requires "a degree of precision and internal coherence rarely met outside the language of formal logic or mathematics" (Dickerson, 1986, p. 4). Legislative language is meant to bring about certain policy outcomes in the real world, and in order to do so effectively, it must be legally enforceable, administrable, and in accordance with both the existing body of statutory law and the Constitution. It must also attempt to foresee any possible contingency that may arise in its interpretation or application (Strokoff and Filson, 2007, p. 97). In many cases, these stringent requirements militate against the goal of "readability," and may lead almost inevitably to a certain degree of complexity.

However, to conceptualize the complexity of legislative language as an inevitable consequence of the requirements of its form is to punt on any number of important substantive questions about *variations* in the complexity of policy language produced by legislatures. Scholars of legislative politics have generally understood this variation either as a reflection of variation in the inherent complexity of the underlying issue, or as a byproduct of strategic legislative actors pursuing their goals.

On the one hand, complexity in a written policy can be thought of as a reflection of the complexity of the underlying issue the policy is meant to address. Complex problems, this logic goes, require complex solutions, and therefore the complexity of the language used in a written policy is increasing

in the complexity of the issue the underlying policy.[1] Some empirical work on legislative politics draws on this tradition by generating issue-level measures of complexity and assigning those issue-level measures to individual bills (Epstein and O'Halloran, 1999; Canes-Wrone and De Marchi, 2002).

Another possibility is that policy complexity is strategically produced by purposive political actors pursuing some particular goal. Curry (2015, pp. 102–106), for example, argues that party leaders intentionally craft complex legislative language as a way of enhancing their informational advantage over rank-and-file legislators about the actual content of large legislative packages. The practical realization of this logic is suggested by political scientist and former Appropriations staffer Glassman (2024), who states "[...] [F]rom a drafting POV this conforms to a maxim I was taught: 'stuff you want publicized, put in plain English; stuff you want buried, do by reference."' In the rulemaking context, there is evidence that complex rules attract less attention during notice and comment rulemaking (Pagliari and Young, 2016), and that bureaucrats may strategically write proposed rules in complex language in order to avoid scrutiny by political principals or affected interests (Potter, 2019).

Considerations about the relevant audiences for legislative text may also help explain variation in the complexity of legislative language. Legislative preferences over the administrative specifics of implementation may be written into bill text (McCubbins *et al.*, 1987; McCubbins *et al.*, 1989), and complex policy detail may be used to constrain executive branch actors in their exercise of delegated policy authority (Epstein and O'Halloran, 1999; Huber and Shipan, 2002; Vannoni *et al.*, 2021). Acknowledging that variation in the complexity of legislative text may stem from strategic motivations and institutional dynamics calls for measurement strategies that go beyond broad issue-level complexity, and are able to generate bill-specific scores derived from text characteristics.

Understanding and Measuring Complexity

The discussion above illustrates that complexity may stem from a variety of sources. Whether through features of the underlying policy area (Jochim and Jones, 2013) or as a product of strategic processes (Curry, 2015), the complexity produced is reflected in the lexical and semantic structure of documents. Therefore, we are purposefully broad and inclusive in our definition of complexity rather than narrowly focused on structural or textual aspects to

[1]Whereas professional legislative drafters always strive for readability and clarity in their products, they acknowledge that there are cases in which "the substantive problems involved are so complex or esoteric that nothing could make their solution readable" (Strokoff and Filson, 2007, p. 99).

the exclusion of other aspects. We understand complexity as the product of structural inputs, political decisions, and language usage together.

This approach is consistent with existing scholarship on the subject of textual complexity. Work by Benoit *et al.* (2019) is helpful in terms of generating text complexity measures from political texts (specifically, snippets from U.S. presidential State of the Union addresses), and Senninger (2023) extends that work to European Union policy language—a step closer to our substantive focus on legislative language produced by the U.S. Congress.[2] Senninger (2023) discusses two aspects of policy complexity in text—one based on the length and detail of a policy (Ehrlich, 2011; Hurka and Haag, 2020), and the other based on the relational network of different policy elements referenced within a policy (Krehbiel, 1992; Adam *et al.*, 2019). Long, detailed policy language with many nested references to other policies would be considered very complex, whereas shorter policy language that is sparse on details and does not reference other policies would be considered less complex.

The relationship between these bill text characteristics and complexity are fairly intuitive to human readers. However, it is unclear whether an LLM would make the same connections between these observable characteristics and the underlying latent trait of complexity. An illustrative example is provided by the following section of legislative text:

- SEC. 7. ESTABLISHMENT OF NATIONAL DATABASE FOR RECORDS OF SERVITUDE, EMANCIPATION, AND POST-CIVIL WAR RECONSTRUCTION. (a) In General. The Archivist of the United States may preserve relevant records and establish, as part of the National Archives and Records Administration, an electronically searchable national database consisting of historic records of servitude, emancipation, and post-Civil War reconstruction, including the Refugees, Freedman, and Abandoned Land Records, Southern Claims Commission Records, Records of the Freedmen's Bank, Slave Impressments Records, Slave Payroll Records, Slave Manifest, and others, contained within the agencies and departments of the Federal Government to assist African Americans and others in conducting genealogical and historical research. (b) Maintenance. Any database established under this section shall be maintained by the National Archives and Records Administration or an entity within the National Archives and Records Administration designated by the Archivist of the United States.

[2]The texts that Senninger uses are recitals, which are essentially summaries of articles of legislation written and adopted by the European Union. Recitals, according to Senninger (2023, Supplementary Information, Section G), describe "the reasons, principles, and assumptions of legislation,' in language that is "more similar to text that citizens usually read in news reports."

This section is fairly detailed. It includes specific names of relevant records (i.e., the Refugees, Freedman, and Abandoned Land Records), and it specifies that the Archivist of the United States is able to designate some sub-entity of the National Archives and Records Administration to maintain the resulting database. However, a human coder would not necessarily conflate this detail with complexity, as unfamiliarity with the specific details does not hinder understanding of the section itself. As long as the reader picks up on the fact that the middle part of the section is simply a list of records with which the Archivist of the United States is likely familiar, the specificity does not add to the complexity or detract from the ability to understand. It is unclear however, whether an LLM would arrive at the same conclusion.

Of course one of the benefits of a pairwise comparison framework is that neither human coders nor the LLM needs to generate a raw complexity score on an arbitrary scale for each text. The relevant question about the section above, then, is whether human coders and the LLM would make the same judgment about the *relative* complexity of that section compared to some other section. We seek to answer that question below.[3] We test the relationship between text characteristics meant to tap the latent trait of complexity—word count, sentence length, word rarity, number of U.S. Code references, etc.—and the outcomes of pairwise comparisons. Of particular interest is whether the relationships between these text characteristics and pairwise comparison outcomes are similar when the comparisons are made by human coders versus when the comparisons are made by the LLM.

Data and Methods

Given our central question about the capacity of LLMs to capture complexity, we proceed in several steps. First, we obtained human evaluations of the relative complexity of legislative texts, in a fashion similar to Senninger (2023). This approach involved carefully setting parameters for which texts were included. We selected sections of bills that became law during the 110$^{\text{th}}$ and 111$^{\text{th}}$ Congresses which were between 1000 and 1200 characters. This was due to our desire to make comparisons equivalent between the two samples

[3]We provide our instructions to human coders in the Online Appendix. We also replicated these instructions to the LLM, and the text of the prompt is also provided in the Online Appendix. We engaged in prompt engineering for these prompts, as the goal is to solicit clear and direct understandings of complexity and have both sets of coders complete the direct task. Given our interest in the comparison between human coders and an LLM, we avoid using a fine-tuned model or moving to a Retrieval-Augmented Generation (RAG) framework. While either option would potentially improve the output generated by the LLM, the inability to engage with a similar task for our human coders makes the prompt-engineered zero-shot model we use the most appropriate structure. The primary point of this exercise is a direct one-to-one comparison between humans and off-the-shelf LLMs.

(Carlson and Montgomery, 2017) and have the comparison not depend on differences in length for human coders (Senninger, 2023). We then randomly sampled 200 observations meeting these length requirements.

As discussed in Eldes *et al.* (2024, pp. 238–239), having a carefully chosen comparison set for any pairwise comparison exercise can help in making fine-grained distinctions in the latent characteristic of interest. A well-chosen comparison set should encompass the full spectrum of the latent trait—in our case, the complexity of the text. We chose five sections that, in our judgment, range from very complex to very simple. Every pairwise comparison in our data includes one of these five sections, which can be found in our Online Appendix. We randomly generated pairings in which each pairing had a randomly drawn text from this comparison set and a randomly drawn text from the pool of 200 text sections described above. Human coders were then asked to compare the relative complexity of the two selected texts, and repeat this for twenty randomly selected pairs of observations.[4]

We next estimated the underlying complexity of a given document using a model for pairwise comparisons from Bradley and Terry (1952). The Bradley-Terry model is a probabilistic framework used to model pairwise comparisons between items as a contest between two items. Think of these items, i and j, as engaged in a competition where we can think of the odds that i beats j as α_i/α_j, where α is a "skill" parameter for both i and j respectively. If we define α_i as eqivalent to $\exp(\lambda_i)$, the odds of i beating j is $\log \frac{Pr(ibeatsj)}{Pr(jbeatsi)} \equiv \lambda_i - \lambda_j$. Thus, the larger the value of λ_i with respect to λ_j, the more likely it is that i will beat j.[5] For our purposes, we can think of this as a competition between two documents for which is the less complex document.

These models are used to infer a latent construct, such as quality, preference, or sophistication. They assume that the probability of one item being preferred over another depends on their relative attributes, which are represented by parameters estimated from the comparison data in the manner described above. For our purposes, if two texts are compared for sophistication, the model estimates a latent measure for each text based on the observed outcomes of all pairwise comparisons. Done iteratively across multiple comparisons for each document, these latent measures are then used to rank or position items along the construct of interest. Carlson and Montgomery (2017) demonstrate that the model is particularly valuable for measuring constructs that are difficult to observe directly, as it relies on relative judgments rather than absolute measures. This allows researchers to derive meaningful insights even in the absence of explicit ratings or objective benchmarks. Relying on explicit rankings rather

[4]See the Online Appendix for the instructions given to respondents as well as a sample comparison. For coding responses, the authors and students from two of their universities were used as coders.

[5]For a similar discussion in an alternative setting, see Wu *et al.* (2023).

than these competitive, iterative pair-wise comparisons could introduce bias due to placing an *ex-ante* measure onto the documents that would constrain coders in a way that more direct comparisons of documents do not. This Bradley-Terry framework as performed with human-coded comparisons provides us a benchmark against which to judge the capability of LLM models to understand complexity in legislation.

In order to give an LLM the same pairwise comparisons made by human coders, we interface with the OpenAI API using the "promptr" package in R (Ornstein, 2024). We also follow the practices of using LLMs for text classification tools described in Ornstein *et al.* (2024), which describes LLMs as "stochastic parrots" that can be effectively adopted in traditional NLP applications. We used GPT-4.o with the temperature set at 0.0 to reduce the variability in response and make our result replicable. We chose GPT-4.o as it is the most affordable and best performing LLM for broad spectrum classification tasks – or was when we did our initial diagnostics (Shi *et al.*, 2024). For the LLM performing the comparison task, we decided to use the exact same directions that we provided for the human coders as we did for the LLM. While this is likely not the most optimal long-term prompt engineering solution, we believe that this approach gives us the most direct apples-to-apples comparison for our question. We set the system message, which governs the overall logic of the LLM architecture, to instruct it to act as a coder who is versed in text complexity and has studied American politics.[6] To generate scores for each, we fit a Bradley-Terry model on the data generated from the pairwise comparisons between each document. This makes the tasks between the LLM and the human coders completely analogous.

We next explored having the LLM perform a much larger set of comparisons that were not the same sets that our human coders analyzed, creating a comparison point to evaluate how the model performs outside of its regular context. For this set of comparisons, we selected all sections of legislation that became law from the 110th and 111th Congresses between 750 and 2500 characters. This allows the comparison made by the LLM to include variation in length of sections not given to our human coders or original LLM comparison. We also did not provide the comparison bridging set structure in order to more clearly match general conceptions of complexity outside of Bradley and Terry (1952). The logic behind this choice is to fully evaluate the constraints of an LLM in making these pairwise comparisons effectively.

LLMs as a Measurement Tool

The integration of Large Language Models into social science research has opened significant opportunities for text classification and measurement tasks.

[6]The text of our prompt and system message are included in the Online Appendix.

These models, such as GPT-3 and GPT-4, demonstrate strong performance in analyzing and classifying text with minimal task-specific training data. Ornstein *et al.* (2024) provide a succinct overview for how to use LLMs in text classification tasks more traditionally suited for NLP methods/models. For example, Wu *et al.* (2023) utilized ChatGPT to estimate U.S. senators' ideological leanings through pairwise comparisons analyzed using the Bradley-Terry model. Their "Ideology LaMP scores" showed high correlation with the first dimension of DW-NOMINATE while also providing unique insights into ideological distinctions (Wu *et al.*, 2023). Indeed, the approach Wu *et al.* (2023) take to using LLMs relies on a similar adversarial pairwise comparison logic and Bradley Terry model that we do. Burnham (2024) extended this line of inquiry by introducing "Semantic Scaling," a method that combines LLM-generated classifications with item response theory to measure ideological dimensions in both mass and elite political texts (Burnham, 2024). These applications highlight how LLMs can replicate or extend existing measurement frameworks in political science.

More broadly, LLMs have demonstrated their utility in computational social science by classifying and interpreting social phenomena, such as political ideology and persuasiveness, offering nuanced analyses of social behavior (Ziems *et al.*, 2023). They have also been employed to simulate responses to social science experiments, with GPT-4 accurately predicting outcomes that align closely with empirical results (Argyle *et al.*, 2023). Despite these successes, the use of LLMs is not without challenges. Algorithmic bias, ethical considerations, and the need for effective prompt engineering remain critical concerns (Ziems *et al.*, 2023). Additionally, while LLMs offer scalability and versatility, their outputs require careful validation to ensure reliability and accuracy (Egami *et al.*, 2024). There are also concerns that LLMs collapse the complexity of measurement found in human-generated data when generating synthetic data, a concern highlighted by Bisbee et al. when they examined synthetic survey response data and found that LLM-generated responses lack the noise and messiness inherent in real response data (Bisbee *et al.*, 2023). This is a concern for us, given that we are using an LLM as a replacement for human coders.

Operationalizing Complexity Metrics

Given our interest in whether LLMs can measure legislative complexity, we utilize the Bradley-Terry method to evaluate our pairwise comparisons, which produces a probabilistic estimate of whether a given text selection is the less complex of two in a given comparison. This probabilistic estimate forms our dependent variable in all models. For what drives this selection, we begin with the metrics developed by Benoit *et al.* (2019) (hereafter BMS) to assess textual sophistication. These metrics provide a systematic approach to measuring the

complexity and sophistication of texts by focusing on linguistic and structural features.

The first measure is a composite metric that integrates sentence length, word rarity, and syntactic structure. This score is derived from a similar procedure using Bradley Terry models of pairwise comparisons of textual snippets that we use. Increasing values in this variable increase the relative "simplicity" of a given text within a given comparison. We also utilize a decomposed version of this score through essentially the same variables as Benoit *et al.* (2019) and Senninger (2023). The first variable in this separated measure, the *Google Min Score*, calculates word rarity using the least frequent word's relative frequency from the Google Books N-gram corpus, where less frequent words contribute to higher scores. The second measure, the *Proportion of Nouns*, evaluates the ratio of nouns to total words in the text, with a higher proportion of nouns indicating greater complexity and abstractness.

Additionally, we include two supplementary measures: *Mean Sentence Length*, which captures the average number of words per sentence and reflects syntactic complexity, and *Mean Word Syllables*, which measures the average syllables per word as an indicator of vocabulary sophistication. These metrics together provide a robust framework for evaluating textual sophistication, allowing for nuanced comparisons of complexity across different texts.

In addition to the previously mentioned metrics, we incorporate three additional measures to enhance our assessment of textual sophistication. First, we utilize the *Flesch-Kincaid Readability Grade Level*, a widely recognized metric that evaluates text readability by considering average sentence length and average syllables per word. This formula assigns a U.S. school grade level, indicating the minimum education required to comprehend the text Flesch (1948). Second, we count the number of references to the U.S. Code within each section. This quantifies the extent to which a bill section is interconnected with existing legislation, reflecting its integration into the broader corpus of American law. This is comparable to Senninger (2023) using references in his study, though applied for the American context. For the same conceptual purpose, we include a count of the references in each section to *other* sections, and a count of the references to other laws.[7] Finally, we include a binary indicator denoting whether a section delegates authority to an administrative agency. This measure is derived from the methodology outlined by Lerner and Spell (2020) who employed deep and active learning classifiers to identify instances of congressional delegation to administrative agencies.

These additional metrics provide a comprehensive framework for evaluating textual sophistication, capturing various dimensions of complexity. Each captures an underlying dimension of textual sophistication or a separate

[7]References to law can be done through citations to the U.S. Code, but can also be done by referencing a specific Public Law number. We create a separate variable for each of type of reference.

component of this, with the possibility that the contents of a section capture separate factors.

Results

We present the results of our first set of comparisons in Table 1. This table contains the human classifications of the comparisons underlying the Bradley-Terry models. Analysis of the LLM-human coder treatments for our comparisons reveals that the LLM agreed with human coders on roughly 71.5% of the 1,716 comparisons.[8] Each model uses different textual measures of complexity to assess both the effectiveness of human coders in identifying complex texts and the strength of the relationship between the textual measures and the latent trait of complexity. For each of these results, the model is predicting the simplicity, or the ease with which a human would be able to read and understand a given document. Coefficients capture the relationship between each covariate and the likelihood that a given text is selected as easier to understand in any given pairwise comparison.

We see a very consistent story with the human classification models. For the BMS score, which is the aggregate textual ease score, we observe a strong relationship in predicting which text will be selected in the binary, parallelized comparisons. This is evident in Model BT1. For Model BT2 we examine the components of the BMS Score index broken out, finding that *Google Min Score* and *Proportion Nouns* are the most significant predictors. In Model BT3, the *Flesch-Kincaid Score* is negatively associated with text simplicity, aligning with expectations. Similarly, in Models BT4 and BT5, the counts of references to the *U.S. Code*, references to other *Sections* of legislation, and references to other *Public Laws* are all negatively associated with section simplicity, although the relationship is only statistically significant for *U.S. Code* references. As the number of references to the U.S. Code increase, the text becomes more complex, a result consistent even when controlling for the aggregate measure of text complexity in Model BT5. Models BT6 and BT7 demonstrate that, while conditioning on the aggregate text complexity metric (BT6) and its components (BT7), predicted delegation in a section has a positive and statistically significant association with the simplicity of a section. These results in Table 1 provide the benchmark for comparison with our LLM structure, giving us an understanding of how human coders with detailed instructions and some training understand relative complexity.

Next, we move on to the set of results replicating the same procedure using GPT 4.o instead of human coders.

[8]See the Online Appendix for more discussion and a breakdown of agreement for each comparison section.

Table 1: Bradley terry models: Human classification benchmark.

	BT 1	BT 2	BT 3	BT 4	BT 5	BT 6	BT 7
BMS Score	0.199*				0.155*	0.412*	
	(0.044)				(0.046)	(0.054)	
Mean Sentence Length		−0.001					−0.002*
		(0.001)					(0.001)
Mean Word Syllables		0.542					−0.122
		(0.291)					(0.328)
Google Min Score		−5224.159*					−1915.011
		(2428.094)					(2877.415)
Proportion Nouns		12.043*					13.068*
		(1.013)					(1.154)
Flesch Kincaid			−0.005*				
			(0.001)				
U.S.C. count				−0.138*	−0.117*	−0.123*	−0.056
				(0.035)	(0.034)	(0.039)	(0.038)
SEC count				−0.010	−0.002	0.003	−0.027
				(0.024)	(0.025)	(0.027)	(0.029)
Public Law Count				−0.025	−0.043	0.287*	0.142
				(0.095)	(0.098)	(0.107)	(0.115)
Delegation?						1.108*	0.443*
						(0.100)	(0.117)
Num.Obs.	1716	1716	1716	1716	1716	1716	1716
AIC	2359.8	2066.7	2361.9	2355.2	2345.6	2178.4	1988.0
BIC	2365.2	2088.5	2367.3	2371.6	2367.4	2260.1	2086.0
RMSE	0.50	0.45	0.50	0.50	0.49	0.47	0.44

Note: $^{*}p < 0.05$
Dependent variable is the outcome of each pairwise comparison, where higher values indicate a less complex text. Observations are comparisons between one of five bridging comparison texts and a randomly selected comparison text.

Table 2: Bradley terry models: LLM classification.

	BT 1	BT 2	BT 3	BT 4	BT 5	BT 6	BT 7
BMS Score	0.031				0.135*	0.078	
	(0.043)				(0.047)	(0.052)	
Mean Sentence Length		−0.001*					0.000
		(0.000)					(0.001)
Mean Word Syllables		1.582*					1.013*
		(0.284)					(0.316)
Google Min Score		6423.158*					3519.834
		(2312.091)					(2730.835)
Proportion Nouns		5.266*					4.382*
		(0.908)					(1.035)
Flesch Kincaid			−0.001				
			(0.001)				
U.S.C. count				−0.158*	−0.183*	−0.168*	−0.076
				(0.037)	(0.039)	(0.043)	(0.042)
SEC count				−0.062*	−0.069*	−0.069*	−0.078*
				(0.025)	(0.025)	(0.027)	(0.028)
Public Law Count				−0.223	−0.191	0.048	0.002
				(0.115)	(0.112)	(0.125)	(0.126)
Delegation?						1.010*	0.651*
						(0.099)	(0.112)
Num.Obs.	1716	1716	1716	1716	1716	1716	1716
AIC	2380.4	2221.500	2379.900	2316.100	2309.500	2169.800	2119.800
BIC	2385.800	2243.300	2385.400	2332.400	2331.300	2251.500	02217.800
RMSE	0.50	0.47	0.50	0.49	0.49	0.47	0.46

Note: $^{*}p < 0.05$
Dependent variable is the outcome of each pairwise comparison, where higher values indicate a less complex text. Observations are comparisons between one of five bridging comparison texts and a randomly selected comparison text.

In this set of results, seen in Table 2, we observe roughly similar patterns to those in Table 1, though there are some differences. For all models in which it is included, the BMS score, which is the aggregate index of textual ease, is positively associated with section simplicity, although that relationship is only statistically significant in Model BT5. Human coders in Table 1 and the LLM in Table 2 appear to understand complexity in broadly similar terms, though the decomposed results in columns BT2 and BT7 reveal some differences for the components of the *BMS Score*. An interesting and related result is that *Flesch Kincaid* is also not statistically significantly related to our simplicity outcome. Another difference worth noting between the results in Table 1 and those in Table 2 is that, in the comparisons made by the LLM, the number of references to different *Sections* of legislation are negatively and significantly associated with section simplicity in all models in which the variable is included. Just as in the models using human comparisons, the coefficients on the variable for predicted delegation are positive and statistically significant in Models BT6 and BT7, demonstrating that delegating sections use less complex language than non-delegating sections, all else equal. Cumulatively, one major takeaway from comparing the results in Tables 1 and 2 is that text-based variables drawn from the extant literature on complexity help to explain variation in both human and LLM attempts to pick the less complex section in a set of pairwise comparisons. In other words, human coders and the LLM seem to "think about" complexity in similar ways.

For the final set of models, shown in Table 3, we look now at the unstructured comparisons between all bill sections evaluated by the LLM. We do not use our fixed bridging texts here and just use fully random comparisons – but also use a much larger set of bills. We find few significant relationships between any of our substantive variables and the modeled pairwise comparisons. Most of the significant relationships—such as the negative and statistically significant coefficients on the *BMS Score* variable in BT1, BT5, and BT6 and *Flesch Kincaid* in column BT2 —are in the opposite of the expected direction. We find the general lack of substantively-meaningful relationships in these models to be indicative that the structuring of the pairwise comparisons matters significantly for the identification of the model.

This insight suggests that relative anchoring points are critical for the validity of our measurement model. As Eldes *et al.* (2024) point out, without a carefully selected set of "bridging items," respondents may be able to distinguish between items near opposite extremes of the latent trait of interest (in our case, complexity), but they will not be able to make fine-grained distinctions between observations between those poles. This is a well-known feature of many issues in unidimensional scaling and latent trait modeling. It is encouraging to observe that the same logic applies to LLM-based coding. This insight, of course, is prominently discussed in the literature on modeling ideal points in Congress, particularly in the works of Poole and Rosenthal, 2000, as well as Clinton

Table 3: Bradley terry models: LLM classification with all documents.

	BT 1	BT 2	BT 3	BT 4	BT 5	BT 6	BT 7
BMS Score	−0.033*				−0.040*	−0.039*	
	(0.013)				(0.013)	(0.014)	
Mean Sentence Length		0.000*					0.000*
		(0.000)					(0.000)
Mean Word Syllables		0.014					−0.038
		(0.056)					(0.059)
Google Min Score		−8.228					−59.647
		(500.916)					(509.029)
Proportion Nouns		−0.011					−0.104
		(0.218)					(0.225)
Flesch Kincaid			0.001*				
			(0.000)				
U.S.C. count				−0.002	−0.001	−0.002	−0.004
				(0.009)	(0.009)	(0.010)	(0.010)
SEC count				−0.008	−0.011*	−0.012*	−0.013*
				(0.005)	(0.005)	(0.005)	(0.005)
Public Law Count				−0.023	−0.024	−0.025	−0.027*
				(0.013)	(0.013)	(0.013)	(0.014)
Delegation?						−0.018	−0.015
						(0.019)	(0.020)
Num.Obs.	22425	22425	22425	22425	22425	22425	22425
AIC	31083.3	31089.3	31083.3	31085.0	31078.1	31057.7	31062.3
BIC	31091.4	31121.4	31091.3	31109.0	31110.2	31370.4	31399.1
RMSE	0.50	0.50	0.50	0.50	0.50	0.50	0.50

Note: $^{*}p < 0.05$
Dependent variable is the outcome of each pairwise comparison, where higher values indicate a less complex text. Each observations is a comparison between two randomly selected texts.

et al. (2004) among many others. These studies emphasize the importance of anchoring and comparative structure in developing robust scaling models, providing a theoretical foundation that aligns well with our findings.

Discussion

In this paper, we set out to answer the question of whether or not LLMs are capable of measuring legislative complexity. Our answer is a qualified "yes." As our analyses demonstrate, when given identical instructions and fed the same set of pairwise comparisons, an LLM performed comparably to human coders, agreeing on roughly 71.5% of 1,716 comparisons. The results of the Bradley Terry models in Tables 1 and 2 demonstrate that the latent trait of complexity produced by these comparisons has relatively consistent relationships with variables drawn from the existing literature.

However, we offer some caveats about the performance of the LLM for our specific task. As is discussed in more detail in the Online Appendix, the human coders and the LLM did not agree perfectly on the ordering of our five selected bridging texts from most complex to least complex. In individual cases in which the human coders and the LLM disagree on a pairwise comparison, we are agnostic about which is the "correct" answer, but our goal here was to answer the question of how well an LLM could replicate human coding. As we demonstrate, the results are positive, at least in a structured setting. In a setting in which the LLM is asked to make comparisons between randomly generated sections of text without a consistent set of bridging texts, however, we have serious concerns about its performance. The results in Table 3 are a testament to the importance of adequately structuring these pairwise comparisons with a carefully selected set of bridging observations.

Careful and informed use of LLMs in the measurement of text-based policy complexity opens up a number of avenues for interesting and important research. The results from the Bradley Terry model allow us to make inferences about the relationships between easily-observable and measurable text features and the latent trait of complexity in policy language. By extracting the coefficients from any of the models we present and applying them to text-based data on any legislative text, other scholars are able to create complexity scores characterized by uncertainty. Alternatively, if scholars prefer to run their own sets of pairwise comparisons on a different set of legislative texts, we have demonstrated that LLMs are capable of performing that task, given proper instructions and constraints. The scalability of this solution beyond what is feasible for human coders represents a significant advancement in the measurement of complexity in legislative texts.

A fine-grained text-based measure of complexity at the legislative section level could allow scholars leverage on a number of persistent research questions

about political institutions and policymaking. The discipline, for the most part, has relied on two methods of accounting for the complexity of legislative texts—either relying on length of the text, or making assumptions about the complexity of certain *policy areas*, and characterizing the complexity of entire bills based on their classification within a given policy area. The method that we demonstrate in this paper allows for the derivation of complexity scores at the section level based on the actual characteristics of the text. Such a measure would afford the opportunity to answer questions about how the complexity of policy language is associated with delegation (Anastasopoulos and Bertelli, 2020), voting on legislative proposals (Canes-Wrone and De Marchi, 2002), policy diffusion (Makse and Volden, 2011), and the duration of the legislative process (Hurka and Haag, 2020), for example. While these questions take complexity as a given and estimate its effect on other outcomes, a fine-grained text-based measure of complexity would also allow for the testing of propositions that conceptualize complexity as a dependent variable. For example, such a measure would allow scholars to answer questions about how variations in the centralization of the legislative process or the necessity of pre-floor negotiations affect the complexity of legislative proposals (Curry, 2015). These potential applications highlight just a fraction of what LLM-derived complexity measures can achieve. With these tools, researchers have a versatile foundation to investigate a broad range of legislative phenomena.

References

Adam, C., S. Hurka, C. Knill, and Y. Steinebach. 2019. *Policy accumulation and the democratic responsiveness trap*. Cambridge University Press.

Anastasopoulos, L. J. and A. M. Bertelli. 2020. "Understanding delegation through machine learning: A method and application to the European Union". *American Political Science Review*. 114(1): 291–301.

Argyle, L. P., E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples". *Political Analysis*. 31(3): 337–351. DOI: 10.1017/pan.2023.2.

Baumgartner, F. R., B. D. Jones, and M. C. MacLeod. 2000. "The evolution of legislative jurisdictions". *Journal of Politics*. 62(2): 321–349.

Benoit, K., K. Munger, and A. Spirling. 2019. "Measuring and explaining political sophistication through textual complexity". *American Journal of Political Science*. 63(2): 491–508.

Bisbee, J., J. D. Clinton, C. Dorff, B. Kenkel, and J. M. Larson. 2023. "Synthetic replacements for human survey data? The perils of large language models". *Political Analysis*: 1–16.

Bradley, R. A. and M. E. Terry. 1952. "Rank analysis of incomplete block designs: I. The method of paired comparisons". *Biometrika*. 39(3/4): 324–345.

Burnham, M. 2024. "Semantic Scaling: Bayesian Ideal Point Estimates with Large Language Models". *arXiv preprint arXiv:2405.02472*. URL: https://arxiv.org/abs/2405.02472.

Canes-Wrone, B. and S. De Marchi. 2002. "Presidential approval and legislative success". *Journal of Politics*. 64(2): 491–509.

Carlson, D. and J. M. Montgomery. 2017. "A pairwise comparison framework for fast, flexible, and reliable human coding of political texts". *American Political Science Review*. 111(4): 835–843.

Clinton, J., S. Jackman, and D. Rivers. 2004. "The statistical analysis of roll call data". *American Political Science Review*: 355–370.

Curry, J. M. 2019. "Knowledge, Expertise, and Committee Power in the Contemporary Congress". *Legislative Studies Quarterly*. 44(2): 203–237.

Curry, J. M. 2015. *Legislating in the Dark: Information and Power in the House of Representatives*. Chicago, IL: University of Chicago Press.

Dickerson, R. 1986. *Fundamentals of Legal Drafting*. Little, Brown and Company.

Egami, N., M. Hinck, B. M. Stewart, and H. Wei. 2024. "Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses". *Working paper*.

Ehrlich, S. D. 2011. *Access points: an institutional theory of policy bias and policy complexity*. Oxford University Press.

Eldes, A., C. Fong, and K. Lowande. 2024. "Information and Confrontation in Legislative Oversight". *Legislative Studies Quarterly*. 49(2): 227–256.

Epstein, D. and S. O'Halloran. 1999. *A transaction cost politics approach to policy making under separate powers*. Cambridge: Cambridge university press.

Flesch, R. 1948. "A new readability yardstick". *Journal of Applied Psychology*. 32(3): 221–233. DOI: 10.1037/h0057532.

Glassman, M. 2024. "Setting aside substantive/political debate over a pay raise, from a drafting POV this conforms to a maxim I was taught: "stuff you want publicized, put in plain English; stuff you want buried, do by reference." This was never staying secret, but you still don't spell it out." https://x.com/MattGlassman312/status/1869391578420433169. Tweet, posted at 8:37 AM on December 18, 2024.

Huber, J. D. and C. R. Shipan. 2002. *Deliberate discretion?: The institutional foundations of bureaucratic autonomy*. Cambridge University Press.

Hurka, S. and M. Haag. 2020. "Policy complexity and legislative duration in the European Union". *European Union Politics*. 21(1): 87–108.

Jochim, A. E. and B. D. Jones. 2013. "Issue politics in a polarized congress". *Political Research Quarterly*. 66(2): 352–369.

Jones, B. D. 2001. *Politics and the architecture of choice: Bounded rationality and governance*. University of Chicago Press.

Krehbiel, K. 1992. *Information and legislative organization*. University of Michigan Press.

Lerner, J. Y. and G. Spell. 2020. "Using Deep and Active Learning Classifiers to Identify Congressional Delegation to Administrative Agencies". *Working Paper*. George Mason Law: Center for the Study of the Administrative State Working Paper series. URL: https://sites.google.com/view/joshuaylerner/research.

Makse, T. and C. Volden. 2011. "The role of policy attributes in the diffusion of innovations". *The Journal of Politics*. 73(1): 108–124.

McCubbins, M. D., R. G. Noll, and B. R. Weingast. 1987. "Administrative procedures as instruments of political control". *The Journal of Law, Economics, and Organization*. 3(2): 243–277.

McCubbins, M. D., R. G. Noll, and B. R. Weingast. 1989. "Structure and process, politics and policy: Administrative arrangements and the political control of agencies". *Va. L. Rev.* 75: 431.

Ornstein, J. 2024. *promptr: Format and Complete Few-Shot LLM Prompts*. R package version 1.0.0. URL: https://cran.r-project.org/web/packages/promptr/index.html.

Ornstein, J. T., E. N. Blasingame, and J. S. Truscott. 2024. "How to Train Your Stochastic Parrot: Large Language Models for Political Texts". *Political Science Research and Methods*. Forthcoming. URL: https://joeornstein.github.io/publications/ornstein-blasingame-truscott.pdf.

Pagliari, S. and K. Young. 2016. "The interest ecology of financial regulation: interest group plurality in the design of financial regulatory policies". *Socio-economic review*. 14(2): 309–337.

Poole, K. T. and H. Rosenthal. 2000. *Congress: A political-economic history of roll call voting*. Oxford University Press, USA.

Potter, R. A. 2019. *Bending the rules: Procedural politicking in the bureaucracy*. University of Chicago Press.

Reilly, S. and S. Richey. 2011. "Ballot question readability and roll-off: The impact of language complexity". *Political Research Quarterly*. 64(1): 59–67.

Senninger, R. 2023. "What makes policy complex?" *Political Science Research and Methods*. 11(4): 913–920.

Shi, H., S. Page, J. Lerner, H. Tran, and B. Sepulvado. 2024. "Assessing the Accuracy of and Bias with Zero-Shot Text Classification using GPT: A Case Study with Social Media and Survey Data". In: *MAPOR 2024*. Conference presentation. Chicago, IL.

Simon, H. A. 1985. "Human nature in politics: The dialogue of psychology with political science". *American political science review*. 79(2): 293–304.

Strokoff, S. L. and L. E. Filson. 2007. *Legislative Drafter's Desk Reference*. Cq Press.

Vannoni, M., E. Ash, and M. Morelli. 2021. "Measuring discretion and delegation in legislative texts: methods and application to US states". *Political Analysis*. 29(1): 43–57.

Wu, P. Y., J. Nagler, J. A. Tucker, and S. Messing. 2023. "Large Language Models Can Be Used to Estimate the Latent Positions of Legislators". *arXiv preprint arXiv:2303.12057*. URL: https://arxiv.org/abs/2303.12057.

Ziems, C., W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang. 2023. "Can Large Language Models Transform Computational Social Science?" *arXiv preprint arXiv:2305.03514*. URL: https://arxiv.org/abs/2305.03514.