Can We Measure Legislative Complexity with LLMs?

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

[1]Assistant Professor of Political Science, Trinity University
[2]Assistant Professor of Political Science, Concordia College
[3]Data Scientist and Research Methodologist, NORC at the University of Chicago

December 3, 2024

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 & 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 Model (LLM). There is a moderately high level of agreement between human coders and the LLM, and the relationships between observable text features and the underlying trait of complexity are similar in comparisons made by human coders and by the LLM. Our work demonstrates that, with researcher-selected bridging texts and carefully designed prompts, LLMs can be used to measure complexity in legislative texts.

1 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 & Volden 2011), and issue complexity is also a relevant consideration in legislative decisions to delegate policymaking authority to the executive branch (Epstein & 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 & 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 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 will be 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 ways to understand complexity in published applied work and broad concept 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.

2 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, re-

quires "a degree of precision and internal coherence rarely met outside the language of formal logic or mathematics" (Dickerson 1986, pg. 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 & Filson 2007, pg. 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 merely 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 & O'Halloran 1999; Canes-Wrone & 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

[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 & Filson 2007, pg. 99).

legislative packages. In the rulemaking context, there is evidence that complex rules attract less attention during notice and comment rulemaking (Pagliari & 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). This runs counter to the advice of professional legislative drafters, who urge that "Unless it is absolutely necessary for the accurate expression of an unusual or complex idea, any language that could confuse or bewilder the reader is suspect even though it may be technically correct, and you should seek an acceptable alternative" (Strokoff & Filson 2007, pg. 94).

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, 1989), and complex policy detail may be used to constrain executive branch actors in their exercise of delegated policy authority (Epstein & O'Halloran 1999; Huber & 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.

3 Understanding and Measuring 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

[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."

detail of a policy (Ehrlich 2011; Hurka & 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.

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.[3]

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. 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.

4 Data and Methods

Given our central question of 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^{th} and 111^{th} Congresses which were between 1000 and 1200 characters. This was due to our desire to make comparisons equivalent between the two samples (Carlson & 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

[3]Interestingly, as discussed in our appendix, human coders did find this section considerably easier to understand than did the LLM.

requirements.[4]

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 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.[5]

We next estimated the underlying complexity of a given document using a model for pairwise comparisons from Bradley & Terry (1952). The Bradley-Terry model is a probabilistic framework used to model pairwise comparisons between items to infer a latent construct, such as quality, preference, or sophistication. It assumes that the probability of one item being preferred over another depends on their relative strengths, which are represented by parameters estimated from the comparison data. For example, if two texts are compared for sophistication, the model estimates a latent score for each text based on the observed outcomes of all pairwise comparisons. These latent scores are then used to rank or position items along the construct of interest. The model is particularly valuable for measuring constructs that are difficult to observe directly, as it relies on relative judgments rather than absolute measures, allowing researchers to derive meaningful insights even in the absence of explicit ratings or objective benchmarks (Carlson & Montgomery 2017).

Next, we repeated the same structure given to human coders with an LLM. We interface with the OpenAI API using the "promptr" package in R (Ornstein 2024). We also follow the

[4]We do not have full coverage at this time for the 200 observations in our data due to limitations from the survey instrument.

[5]See the 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.

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-4.0 with the temperature set at 0.1 to replicate the randomness that sometimes occurs with human coders. We used this to see if the LLM would generate similar comparisons and make similar choices in the same sets of comparisons as our human coders would. 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]

For generating 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. In later versions, we will ask the LLM to directly generate text complexity as a standalone metric in order to compare Bradley-Terry to a more traditional metric of complexity, but for now, we thought the pairwise comparison tasks make the comparison more direct to our human coders.

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 110^{th} and 111^{th} 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 & Terry (1952). The logic behind this choice is to fully evaluate the constraints of an LLM in making these pairwise comparisons

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

effectively.

4.1 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. 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.

4.2 Operationalizing Complexity Metrics

Given our interest in whether LLMs can measure legislative complexity, we utilize the Bradley-Terry scores as our dependent variable in all models. Thus, the dependent variable throughout our paper is the choice of the easier text in a comparison setting. 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 the main metric from the 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 "difficulty" of a given comparison. We also utilize a separated version of this score through much the same variables as Benoit *et al.* (2019) and Senninger (2023). The first variable in this separated measure, the *Google Mean Score*, calculates word rarity using average frequencies 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. 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 Bussing *et al.* (2022) 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 and legislative intent. Each captures an underlying dimension of textual sophistication or a separate component of this, with the possibility that the contents and intent of a section capture separate factors.

5 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. As a reminder, we have 625 comparisons.[7] Each model uses different textual measures of complexity to assess both the effectiveness of human coders in identifying complex texts and the strength

[7]While we fielded more than 625 pairwise comparisons, we subset our data to only include sections that showed up in at least 5 comparisons. This process yielded 625 pairwise comparisons.

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.

Table 1: Bradley Terry Models using Human classification

	BT 1	BT 2	BT 3	BT 4	BT 5	BT 6
BMS Score	0.315***				0.213*	0.216**
	(0.090)				(0.095)	(0.098)
Mean Sentence Length		0.000				
		(0.001)				
Mean Word Syllables		4.157***				
		(0.531)				
Google Mean Score		39.573***				
		(3.476)				
Proportion Nouns		2.097				
		(1.893)				
Flesch Kincaid			−0.009***			
			(0.002)			
U.S.C. count				−0.200***	−0.162***	−0.173***
				(0.046)	(0.046)	(0.047)
Delegation?						0.132
						(0.115)
Num.Obs.	625	625	625	625	625	625
AIC	855.6	656.0	852.0	844.4	841.5	847.0
BIC	860.0	673.8	856.4	848.8	850.3	900.3
RMSE	0.49	0.41	0.49	0.49	0.49	0.48

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Dependent variable is the outcome of each pairwise comparison. Observations are comparisons between one of five bridging comparison sets and a randomly selected comparison text.

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 same components broken out, finding that *Mean Word Syllables*—the average number of syllables per word—and the *Google Mean Score*

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, BT5, and BT6, the counts of references to the *U.S. Code* emerge as significant predictors. As the number of references increases, the text becomes more complex, a result consistent even when controlling for the aggregate measure of text complexity. However, we do not find a significant relationship between predicted delegation to administrative agencies and text complexity in the human classification set.

Next, we move on to the set of results replicating the same procedure using GPT 4.0 instead of human coders.

Table 2: Bradley Terry Models using LLM classification

	BT LLM 1	BT LLM 2	BT LLM 3	BT LLM 4	BT LLM 5	BT LLM 6
BMS Score	0.258**				0.160+	0.276**
	(0.089)				(0.094)	(0.099)
Mean Sentence Length		−0.001				
		(0.001)				
Mean Word Syllables		2.601***				
		(0.478)				
Google Mean Score		29.215***				
		(2.952)				
Proportion Nouns		−1.756				
		(1.783)				
Flesch Kincaid			−0.007**			
			(0.002)			
U.S.C. count				−0.183***	−0.155***	−0.164***
				(0.045)	(0.046)	(0.046)
Delegation?						0.591***
						(0.120)
Num.Obs.	625	625	625	625	625	625
AIC	859.7	714.5	857.2	847.5	846.7	806.9
BIC	864.2	732.2	861.6	852.0	855.6	860.1
RMSE	0.50	0.43	0.50	0.49	0.49	0.47

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Dependent variable is the outcome of each pairwise comparison. Observations are comparisons between one of five bridging comparison sets and a randomly selected comparison text.

In this set of results, seen in Table 2, we observe that most relationships are stronger

than those in Table 1. For the BMS score, which is the aggregate textual ease score, the coefficients are similar, though slightly smaller. For Model BT2, the *Mean Word Syllables* and the *Google Mean Score* remain the strongest predictors. In Model BT3, the negative association between the *Flesch-Kincaid Score* and text simplicity is consistent with the first set of results. Similarly, for Models BT4, BT5, and BT6, the counts of references to the *U.S. Code* show a more pronounced effect, with increasing counts corresponding to higher levels of textual complexity, even when controlling for the aggregate measure of text complexity. We also observe a significant relationship between predicted delegation to administrative agencies and text complexity, though inverse – when controlling for observed textual reading ease, sections that delegate are more likely to be chosen by the LLM as easier to understand. This may be a function of what bill sections that are not delegating are likely doing – something worth thinking about some more.

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 reference points here and just use random comparisons – but also use a much larger set of bills. We find no real relationship between any of our substantive variables and the modeled pairwise comparisons. We find the lack of a relationship here 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. 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 seminal works of Poole & Rosenthal (2000), as well as Clinton *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.

Table 3: Bradley Terry Models using LLM classification: all docs

	BT LLM 1	BT LLM 2	BT LLM 3	BT LLM 4	BT LLM 5	BT LLM 6
BMS Score	−0.033*				−0.034*	−0.032*
	(0.013)				(0.013)	(0.013)
Mean Sentence Length		0.000*				
		(0.000)				
Mean Word Syllables		0.020				
		(0.057)				
Google Mean Score		0.110				
		(0.277)				
Proportion Nouns		−0.014				
		(0.217)				
Flesch Kincaid			0.001*			
			(0.000)			
U.S.C. count				−0.009	−0.010	−0.012
				(0.009)	(0.009)	(0.009)
Delegation?						0.011
						(0.019)
Num.Obs.	22 425	22 425	22 425	22 425	22 425	22 425
AIC	31 083.3	31 089.2	31 083.3	31 088.6	31 084.1	31 065.0
BIC	31 091.4	31 121.2	31 091.3	31 096.6	31 100.1	31 361.6
RMSE	0.50	0.50	0.50	0.50	0.50	0.50

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Dependent variable is the outcome of each pairwise comparison. Observations are comparisons between one of five bridging comparison sets and a randomly selected comparison text.

6 Discussion

This paper has investigated the capacity of LLMs to capture and understand legislative complexity. Through a Bradley-Terry model structure, we find that increase in the scale designed by Benoit *et al.* (2019) has a statistically significant effect on the comparison between sections across nearly all models where it is included, only falling under a standard definition of statistical significance in one model in 2. The components of this measure have largely consistent findings in column 2 Tables 1 and 2, especially for Mean Word Syllables and Google Mean Score. Reading Ease in column 3 of those same tables produces nearly identical effects across human and LLM models, as does the U.S.C. count in columns 4 through 6. The largest difference we find between human coders in Table 1 and LLM models in Table

2 is the effect of a section delegating authority, as the LLM model found a statistically and substantively strong effect for this covariate while the human coders found no such relationship. However, we find either reversed effects for our statistically significant covariates in Tables 1 and 2 in Table 3 or no effect at all where we observed a strong relationship in the earlier models.

In total, we find that LLM structures can be designed to function similar to guided human coders. We see nearly identical effects for guided LLMs and human coders with the exception of delegation, lending credence to the power of this tool to undertake complexity as a measurement exercise. However, we find that unguided LLMs cannot independently capture complexity in a more general sense. Thus, our exploration agrees with Kirsten *et al.* (2024) that LLM-based coding methodology has limitations. However, we also agree with them that with human controls and input LLM coding may provide a fruitful way to understand coding complex phenomena.

This project illustrates the potential for LLMs as research devices as well as the potential pitfalls. We see that LLMs can indeed capture complexity in a fashion similar to human coders, but cannot independently develop a sense of complexity in legislative language. Future research should investigate the outer limits of LLM-based approaches, but with these initial findings we offer evidence that with oversight and input training LLMs can indeed capture complexity.

References

Adam, Christian, Hurka, Steffen, Knill, Christoph, & Steinebach, Yves. 2019. *Policy accumulation and the democratic responsiveness trap.* Cambridge University Press.

Argyle, L. P., Busby, E. C., Golshan, B., Lu, J., Reich, J., & Westwood, S. J. 2023. Predicting Results of Social Science Experiments Using Large Language Models. *arXiv preprint arXiv:2308.07857*.

Baumgartner, Frank R, Jones, Bryan D, & MacLeod, Michael C. 2000. The evolution of legislative jurisdictions. *Journal of Politics*, **62**(2), 321–349.

Benoit, Kenneth, Munger, Kevin, & Spirling, Arthur. 2019. Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, **63**(2), 491–508.

Bisbee, James, Clinton, Joshua D, Dorff, Cassy, Kenkel, Brenton, & Larson, Jennifer M. 2023. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 1–16.

Bradley, Ralph Allan, & Terry, Milton E. 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*.

Bussing, Austin, Lerner, Joshua Y., & Spell, Gregory P. 2022. Using Deep and Active Learning Classifiers to Identify Congressional Delegation to Administrative Agencies. *arXiv preprint arXiv:2206.10513*.

Canes-Wrone, Brandice, & De Marchi, Scott. 2002. Presidential approval and legislative success. *Journal of Politics*, **64**(2), 491–509.

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

Clinton, Joshua, Jackman, Simon, & Rivers, Douglas. 2004. The statistical analysis of roll call data. *American Political Science Review*, 355–370.

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

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

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

Egami, Naoki, Hinck, Musashi, Stewart, Brandon M, & Wei, Hanying. 2024. Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses. *working paper*.

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

Eldes, Ayse, Fong, Christian, & Lowande, Kenneth. 2024. Information and Confrontation in Legislative Oversight. *Legislative Studies Quarterly*, **49**(2), 227–256.

Epstein, David, & O'Halloran, Sharyn. 1999. *A transaction cost politics approach to policy making under separate powers*. Cambridge: Cambridge university press.

Flesch, Rudolf. 1948. A new readability yardstick. *Journal of Applied Psychology*, **32**(3), 221–233.

Huber, John D, & Shipan, Charles R. 2002. *Deliberate discretion?: The institutional foundations of bureaucratic autonomy*. Cambridge University Press.

Hurka, Steffen, & Haag, Maximilian. 2020. Policy complexity and legislative duration in the European Union. *European Union Politics*, **21**(1), 87–108.

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

Kirsten, Elisabeth, Buckmann, Annalina, Mhaidli, Abraham, & Becker, Steffen. 2024. Decoding Complexity: Exploring Human-AI Concordance in Qualitative Coding. *arXiv preprint arXiv:2403.06607*.

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

Makse, Todd, & Volden, Craig. 2011. The role of policy attributes in the diffusion of innovations. *The Journal of Politics*, **73**(1), 108–124.

McCubbins, Mathew D, Noll, Roger G, & Weingast, Barry R. 1987. Administrative procedures as instruments of political control. *The Journal of Law, Economics, and Organization*, **3**(2), 243–277.

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

Ornstein, Joe. 2024. *promptr: Format and Complete Few-Shot LLM Prompts*. R package version 1.0.0.

Ornstein, Joseph T., Blasingame, Elise N., & Truscott, Jake S. 2024. How to Train Your Stochastic Parrot: Large Language Models for Political Texts. *Political Science Research and Methods*. Forthcoming.

Pagliari, Stefano, & Young, Kevin. 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, Keith T, & Rosenthal, Howard. 2000. *Congress: A political-economic history of roll call voting*. Oxford University Press, USA.

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

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

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

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

Strokoff, Sandra L, & Filson, Lawrence E. 2007. *Legislative Drafter's Desk Reference*. Cq Press.

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

Wu, Patrick Y., Nagler, Jonathan, Tucker, Joshua A., & Messing, Solomon. 2023. Large Language Models Can Be Used to Estimate the Latent Positions of Legislators. *arXiv preprint arXiv:2303.12057*.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. 2023. Can Large Language Models Transform Computational Social Science? *arXiv preprint arXiv:2305.03514*.