Political Science Research and Methods (2026), 1–19
doi:10.1017/psrm.2025.10086

ORIGINAL ARTICLE

Fine-tuned large language models can replicate expert coding better than trained coders: a study on informative signals sent by interest groups

Dahyun Choi[1], Denis Peskoff[2] and Brandon M. Stewart[2]

[1]Department of Politics, Princeton University, Princeton, NJ, USA and [2]Department of Sociology and Office of Population Research, Princeton University, Princeton, NJ, USA
Corresponding author: Dahyun Choi; Email: dahyunc@princeton.edu

(Received 24 February 2024; revised 2 July 2025; accepted 24 September 2025)

Abstract
Understanding how political information is transmitted requires tools that can reliably and scalably capture complex signals in text. While existing studies highlight interest groups as strategic information providers, empirical analysis has been constrained by reliance on expert annotation. Using policy documents released by interest groups, this study shows that fine-tuned large language models (LLMs) outperform lightly trained workers, crowdworkers, and zero-shot LLMs in distinguishing two difficult-to-separate categories: informative signals that help improve political decision-making and associative signals that shape preferences but lack substantive relevance. We further demonstrate that the classifier generalizes out of distribution across two applications. Although the empirical setting is domain-specific, the approach offers a scalable method for expert-driven text coding applicable to other areas of political inquiry.

Keywords: crowd-sourcing; expert-coding; interest groups; large language model

1. Introduction

Measuring latent concepts occupies a significant place in political science (Adcock and Collier, 2001). While scholars increasingly benefit from the development of automated analysis and statistical tools to measure concepts of interest, the quantification of concepts that are not directly observable often leads to imperfect estimations of the underlying theoretical constructs (Grimmer *et al.*, 2022; Knox *et al.*, 2022; Park and Montgomery, 2025). This is particularly true for inherently technical and ambiguous traits that require a more sophisticated understanding of context, such as deliberation (Franchino, 2004) or agency ideologies (Clinton and Lewis, 2008). Although classic methods address this issue by relying on expert readings or annotations for case-specific measurement models, these approaches do not scale. The high costs and limited availability of experts—particularly compared to crowdworkers or other hourly labor—have made expert-driven models less feasible for widespread use by researchers.

Recent work has demonstrated that *zero-shot*[1] large language model (LLM) performance can match that of crowdworkers for annotation of relatively *simple* tasks (Gilardi *et al.*, 2023;

[1]The zero-shot setting is one where the LLM is not given any examples of correctly annotated output and must rely on simply the name of the label or a description.

© The Author(s), 2026. Published by Cambridge University Press on behalf of EPS Academic Ltd. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

Ziems *et al.*, 2024). We build on this finding to demonstrate that *fine-tuned* LLMs can replicate expert coding for *complex* tasks. Our substantive focus is on the informational signals sent by interest groups in the United States. While concerns persist that interest groups strategically use information to shape the political process in their favor, most empirical studies on interest group politics tend to focus on financial resources; empirically demonstrating the informational aspect of interest group behavior has been challenging (Bombardini and Trebbi, 2020) due to the inherent complexity of measuring information. While this manuscript focuses on interest groups' information supply as an example, our approach is broadly applicable across contexts to other expert-driven annotations such as subtle gradations of ideology or democracy.

Political scientists have long been interested in informative signals, which are crucial because they help recipients make more informed political decisions. These signals are often described using various terminologies, including "expertise" (Fong, 2020), "analytical," or "falsifiable" information (Esterling, 2007). To quantify informative signals, Esterling (2007) uses nonexpert human coding and counts the number of coded sentences, and the recent work of Ban *et al.* (2023) considers the proportion of keywords potentially associated with analytical information. However, the judgments of nonexpert coders can differ significantly from those of expert coders, as what political scientists or legislators consider informative may not align with the perceptions of crowdsourced workers. Relying on dictionary methods may also lead to coarse characterizations that differ from what political scientists ideally want to capture. The challenge has been that going beyond these more narrow conceptualizations requires technical familiarity and case-specific coding which does not easily scale.

In this study, we introduce a scalable approach that bridges this gap by fine-tuning LLMs to replicate expert-level annotations in complex political text. Unlike zero-shot prompting or dictionary-based methods, fine-tuning enables the model to learn nuanced coding behavior that would otherwise require substantial training and domain expertise. We focus on a task that exemplifies these challenges: distinguishing between informative and associative signals in political communication. Informative signals convey policy-relevant evidence or reasoning, while associative signals appeal to identity or partisan alignment without advancing substantive arguments. These categories are theoretically distinct and play different roles in models of interest group behavior and elite decision-making, but they are empirically difficult to differentiate—particularly in domains where rhetorical strategies are subtle, context-dependent, and strategically employed.

To develop and evaluate our method, we construct a high-quality training dataset of paragraph-level expert annotations based on a collection of media interviews, press releases, and public statements. We then fine-tune several LLMs—including GPT-3, GPT-3.5, GPT-4, and Llama-3—and compare their performance against a range of alternative approaches: lightly trained coders, untrained crowdworkers, zero-shot LLMs, a bag-of-words logistic regression model, and a fine-tuned RoBERTa model. All models and coders are evaluated on their ability to recover expert-coded labels in a held-out test set. We then demonstrate through our examples that the classifier generalizes well out of domain, providing *prima facie* evidence that it can perform effectively even when applied off-the-shelf.

The findings demonstrate that fine-tuned LLMs substantially outperform all other methods—including both machine learning baselines and human coders—in replicating expert judgments across multiple dimensions of the task. Notably, the best-performing LLMs match or exceed the accuracy of individual trained annotators and aggregate human votes, even when applied to previously unseen documents. While prior work has shown that LLMs can rival crowdworkers on simple annotation tasks, our study provides evidence that fine-tuned LLMs can successfully scale expert-driven annotations for more complex, theory-rich coding schemes. Our work also suggests that fine-tuning might be necessary as we show—consistent with prior work—that careful zero-shot classification matches the performance of individual lightly trained annotators but does not match groups of coders or experts. This represents a methodological contribution with broad utility for political science research, especially in domains where expert coding is traditionally required but prohibitively expensive or labor-intensive to scale.

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

While our empirical application focuses on interest group signals in U.S. trade policy, our proposed approach is broadly applicable to the study of information. The classifier is not offered as a theoretical claim about interest groups per se, but rather as a practical tool that enables scholars to recover conceptually grounded annotations across large textual datasets, as defined by academic literature. Many areas of political science—from legislative studies to international relations—rely on conceptual distinctions that are difficult to observe directly and often require expert interpretation. By fine-tuning LLMs on high-quality annotations, researchers can extend expert-driven frameworks to much larger datasets while preserving the subtlety and judgment those frameworks require. We hope to rekindle political scientists' desire to construct and validate complex measurements which may require expert judgments by demonstrating that these approaches can be scaled efficiently using fine-tuned LLMs.

2. The stakes of coding with experts

A growing number of scholars construct their own measures and make inferences about unobservable concepts from observed data. Crowdsourced workers play vital roles in this process, from collecting data (Sumner *et al.*, 2020) to constructing gold standards (e.g., Quinn *et al.*, 2010) and validating measures (e.g., Chernykh *et al.*, 2017; Ying *et al.*, 2022). The expectation set for human coding as a benchmark has relied on the assumption that their understanding would exceed that of machines, and that their decisions would be mostly correct and valid if they are properly trained. In particular, the "wisdom of the crowd" is expected to match that of experts (Benoit *et al.*, 2016), producing data whose quality is indistinguishable from classic methods using expert coding.

However, such expectations of crowdsourcing are contingent upon the difficulty of measurement tasks. Several studies document that crowdsourced workers are often poor substitutes for experts when specific contextual or subject knowledge is required (Cheng *et al.*, 2020; Marquardt *et al.*, 2025). Certain concepts inherently involve a high degree of technicality, such as agency ideologies (Clinton and Lewis, 2008), legislative power (Chernykh *et al.*, 2017), democracy (Treier and Jackman, 2008), or party space (Huber and Inglehart, 1995); and classic methods rely on expert coding generated by academics, journalists, or Washington think tanks as a means of assessing these complex quantities of interest.

In this section, we use work by Lowande (2024) and Lowande and Shipan (2022) to demonstrate that experts and crowds diverge. The Lowande studies seek to measure the inherently complex concept of presidential discretion. Their main measure is an expert coding constructed by 173 political scientists with expertise in presidential policymaking. They selected 25 topics from the Comparative Agendas Project and asked each coder to compare 20 pairs of randomly selected terms. The following prompt on presidential discretion was given to both groups.

Presidents can use their legal authority as the head of the executive branch to change policy without Congress. In some cases, they have a great deal of discretion, or freedom, to change existing policies and create new ones. In others, their ability to use executive actions to change or create policy is more limited.

Then a random utility model was estimated to obtain a discretion score, following the approach of Carlson and Montgomery (2017). In an appendix, they present a validation check using 303 screened coders recruited from Lucid.[2]

Figure 1 shows the baseline estimates of presidential discretion, with expert coding on the left and nonexpert coding on the right, with cases of substantial difference shown in red.[3] While expert

[2]Note that by including the definition in the prompt, they (in theory) ensure that everyone has the same working concept of discretion. Lowande and Shipan (Lowande and Shipan, 2022) also note that they conducted pre-tests to ensure coders understood the concept of presidential discretion and excluded low-performing respondents ex-post.

[3]We replicate the model using the rstan package. While there may be minor differences from the original plot in Lowande and Shipan (Lowande and Shipan, 2022) due to updated packages, the ordering of topics remains the same.

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

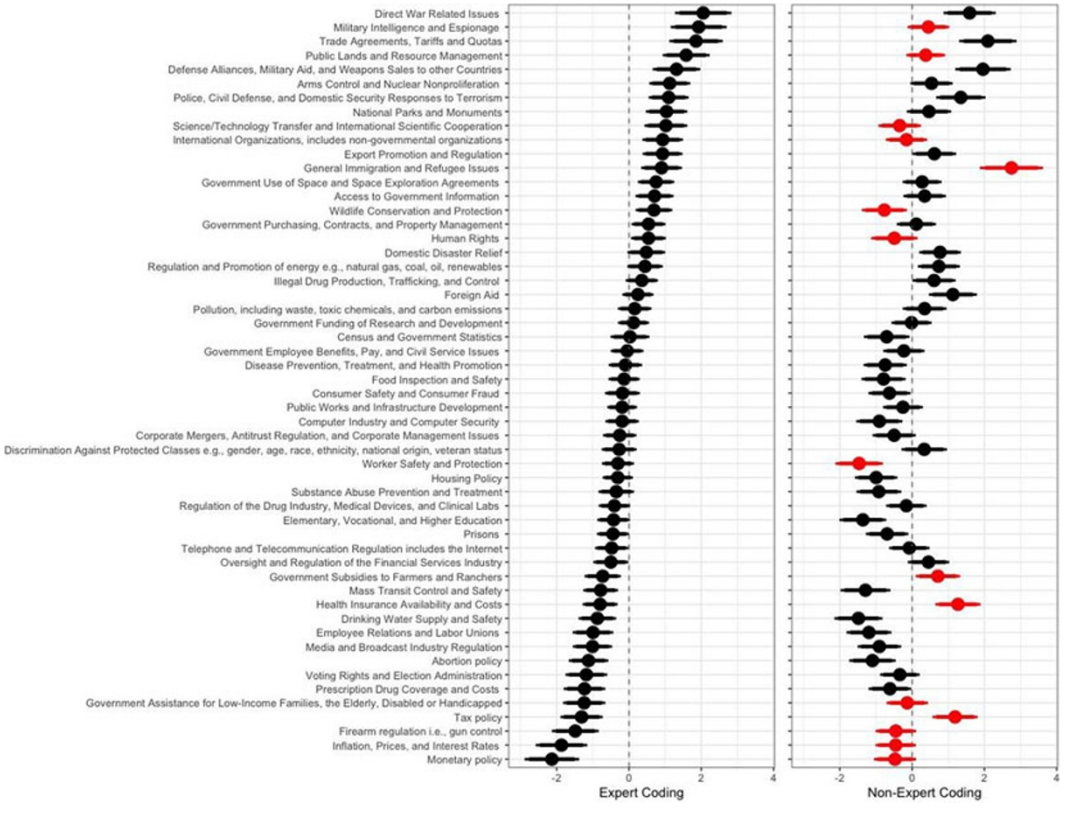


Figure 1. Estimated presidential discretion from Lowande and Shipan (2022). Figure adapted from Lowande and Shipan (2022) using data kindly shared by the original author and regenerated with updated software. Red indicates topics where the estimates show a margin greater than 1 between expert coding and nonexpert coding. Note also that the implied ordering is substantially different between the two sets of estimates.

coding identifies policies like immigration, trade, or defense as more discretionary, nonexperts identify areas such as regulation, health insurance, or taxation as more discretionary, demonstrating the discrepancies between measures constructed by expert coding and those by nonexpert human coding.

The discrepancies in Figure 1 highlight the need to reconsider the application of expert versus nonexpert coding during the design stage of measurement tasks. Lowande (2024) notes that the behavior of nonexpert coders is less likely to show an internally coherent ordering. For example, if the president is perceived to have more power over guns than abortion, and more power over abortion than space, then logically, the president should be perceived to have more power over guns than space. However, this transitive relationship is not consistently observed in data annotated by nonexpert coders. These properties raise some concerns that these deviations may not be purely random, but rather more systematic; these errors are likely to be contaminated by other factors, such as coders' lack of understanding of what presidents can or cannot do, rather than accidental mislabeling. While some might conjecture that such differences could be attributed to the relative level of attention devoted by coders, the authors note that the average response times are quite similar across the two groups, implying that the resulting differences are a function of expertise.

Our case will similarly show that there is a difference between expert and nonexpert judgment. This needn't be the case across all tasks, but many tasks of interest to political scientists will involve technical expertise or judgment. While occasionally large groups of experts can be

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

marshaled (or principal investigators can code large groups of texts themselves), it tends to be more time-consuming and resource-intensive, making it less accessible for researchers in general settings (Benoit *et al.*, 2016). In the following section, we describe the procedures for our own expert-coding task (characterizing the information supply by interest groups) and show how fine-tuned LLMs can achieve expert-level performance.

3. Measuring signals from interest groups

There is a widely held view that interest groups strategically use information to influence the political process in ways that benefit them (e.g., Grossman and Helpman, 2001). As policymakers deal with highly technical and complex issues, they rely not only on their own resources but also on information provided by interest groups (Hall and Deardorff, 2006). In addition to informing policymakers, many interest groups also seek to convince the general public (Kollman, 2021). These efforts are analogous to lobbying activities, as voters—who often lack the expertise and incentive to thoroughly research policy proposals—depend on readily accessible information from sources such as the media. Interest groups are eager to supply this information because it allows them to shape public opinion in a manner that supports their objectives. However, without a clear understanding of what constitutes the information that these interest groups use to influence the process, it is difficult to assess the accuracy of such claims or to consider the normative implications of interest groups' information supply, such as how sophisticated interest groups are in determining what signals to send to whom.

Existing studies using experimental data support the claim that voters consider information from interest groups in order to form their political belief (e.g., Kuklinski *et al.*, 2000) points to two types of signals—(1) *informative* signals that can increase voter's knowledge about a given policy and (2) the rest (that we refer to as *associative* signals) to provide "shortcuts" without direct relevance to the given policy (Mondak, 1993; Lupia, 1994). While scholars have explored the impact of advertising on voting behaviors or turnout (e.g., Lovett and Peress *et al.*, 2015; Kalla and Broockman, 2018; Spenkuch and David, 2018) and discussed frameworks for understanding the persuasive effects of political advertising through experiments,[4] the attempt to link such theoretical distinctions to empirical measurements is rare. In this section, we present a typology for organizing these cues and a fine-tuned classifier for scaling them.

3.1. *Developing a Coding Scheme for* Informative *and* Associative *Signals in Documents about Trade Policy*

An increasing number of studies in political science have utilized automated content analysis to analyze what political actors are saying or writing (Grimmer *et al.*, 2022). In particular, the supervised machine learning approach has dealt with a wide range of textual analysis, from party manifestos to online advertising or social media posts (e.g., King *et al.*, 2013; Barbera´ *et al.*, 2021; Fowler *et al.*, 2021; Park and Montgomery, 2025). The classification tasks for the use of machine learning algorithms have been centered on inferring party political leaders' positions (Laver *et al.*, 2003; Slapin and Proksch, 2008; Catalinac, 2014), measuring populist rhetoric (Di Cocco and Monechi, 2021), or focused on identifying the ideology score from the political texts (Laver *et al.*, 2003; Diermeier *et al.*, 2012).

Unlike conventional classification tasks, measuring *informative* and *associative* signals in texts is a particularly challenging case for content analysis for at least three reasons. First, the type of signal is rarely the main source of variation in the words and thus is difficult for a machine to learn. Second, any given document can exhibit more than one type of signal requiring a representation that can allow for

[4]Coppock (Coppock, 2023) argues that while *informative* signals tend to persuade groups in parallel towards supporting a policy, it can have negative effects on affect towards the sender of the signal. By contrast, the effects of group cues (an example of an *associative* signal) can have effects that are specifically dependent on the audience. We cover additional literature in Appendix B.

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

nonexclusive coding. Third, the concept of the informative signal itself encompasses inherent technical difficulties that require deeper contextual knowledge and specialized expertise in measurement tasks. These intrinsic challenges make this task an ideal setting for expert-based coding.

We begin by developing a codebook to guide annotators in identifying informative signals and linking them to observable patterns in text. Building on typologies of political information developed by Lupia (1994) and Martin (2020), we construct a macro-level coding scheme that distinguishes between two overarching categories of signals: *informative* and *associative*. These categories are further divided into eight subtypes designed to capture rhetorical strategies commonly used by interest groups. While the framework could, in principle, be adapted to other policy areas, we focus on the domain of trade policy for two reasons. First, trade is a context where interest group competition is especially intense and well-documented. Second, unlike many other policy domains, trade policymaking has been the subject of a well-developed literature in international political economy that theorizes the types of information interest groups provide to influence both public opinion and elite decision-making. This literature provided a strong foundation for constructing a theoretically grounded and empirically precise annotation framework. While the primary quantity of interest to scholars is the dichotomy between *informative* and *associative* signals, we break these larger groups into smaller categories so that each component provides more specific instructions for coders, minimizing the potential errors that might arise, even for expert coders.

We do not claim to provide a full account of information tactics observed in the political texts, but we do pinpoint two particular mechanisms of information by interest groups *informative* versus *associative. Informative* signals comprise any traditional economic arguments that primarily focus on expectations of the wage and welfare effects of policies. These arguments help voters to identify whether they are losers or winners in the face of greater economic openness. We annotate three nonexclusive types of informative signals: prediction—the expected outcome of a policy, status-quo—information about the contents of a policy and specifics—jargon, numeric estimates, or witness by experts. Prediction and status-quo have been argued to be primary com-ponents of information that drive deliberation regarding the merit of policies on individual welfare (Chaiken, 1980). Specifics provide the detailed and technical signals that can increase the validity of an argument and exert a stronger impact on political persuasion (Mondak, 1993).

Associative signals have weaker relevance to the details of the specific policy. Scholarly research on trade preference has increasingly provided empirical findings that discuss the importance of noneconomic explanations in the formation of policy preferences (e.g., Bechtel *et al.*, 2014; Guisinger, 2017). Building on the growing number of studies that highlight the influence of other factors not directly related to economic consequences of trade policies, the route of associative signals is identified via five categories. First, sociotropic appeals, the perceptions of how trade affects the country and societies as a whole, have been regarded as the most important influence on trade preference, regardless of income effects of economic openness and material self-interests of individuals (e.g., Mutz, 1992; Mutz and Mondak, 1997; Mansfield and Mutz, 2009). We classify sociotropic appeals as one of associative signals because sociotropic appeals centered on group-based perceptions are identified with groups based on general and broad closeness, rather than on the group people felt closest to economically (Mutz and Mondak, 1997).

Additionally, we include categories of environmental issues and human rights. Extensive experimental evidence (e.g., Bechtel *et al.*, 2012; Brutger and Guisinger, 2022) suggest that concerns over environmental issues or labor rights play a noteworthy role in shaping protectionist attitudes. Interest groups frequently portray themselves as environmentally friendly or respectful of human rights to enhance their image, even though these issues are not concerns in a given policy.[5] Next, ideology—liberal or conservative—provides cues to voters in their preference formation (e.g., Sniderman

[5]For associative signals, we focused on paragraphs that featured emotional and reputational appeals related to these issues. We didn't annotate paragraphs discussing specific details of environmental or labor rights provisions within Free Trade

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

et al., 1993; Milner and Tingley, 2011). Finally, interest groups strategically use their ratings or endorsements to influence public preference (e.g., Arceneaux and Kolodny 2009; Weber *et al.*, 2012; Druckman *et al.*, 2020). Dalton (2013) suggests that many citizens shape their policy preferences based on signals that interest groups provide, such as group ratings, and endorsements from a labor union or business associations.

We note that each category of signals is defined based on scholarly discussions in political science, so that the empirical reality this paper seeks to capture corresponds to the scholarly concepts introduced by the seminal works in the discipline. For this reason, the codebook may look difficult, but it provides the sophisticated nuances of academic concepts and sample paragraphs. A simplified version of the codebook is given in Table 1 with the complete codebook in Appendix A. These types of signals are not intended to be exhaustive, and in earlier iterations we had additional signals in our coding scheme (e.g., tax and security) that appeared too infrequently to be useful. The categories are also not exclusive, allowing us to represent the complex reality that a single message can send several types of signals.

Across different categories, fine-tuned LLMs convincingly outperform all other options. Consistent with the "wisdom of crowds" hypothesis the majority classifier of both Upwork and MTurk coders outperforms any individual coder but almost 5 percentage points worse than the weakest fine-tuned LLM (Table D.4 in the Appendix provides additional details).[7]

Several additional patterns need further discussion. First, one might expect that more recent LLMs would show stronger results given the continued progress demonstrated by these models on zero-shot benchmarks. Yet, our results show that fine-tuned GPT-3, Llama-3, GPT-3.5, and GPT-4 perform comparably across the examined categories. Second, we observe no remarkable differences between open-source and closed-source language models; both consistently outperform majority vote classifiers derived from both trained and untrained human coders. This finding highlights the robustness of fine-tuned automated approaches relative to traditional human coding methods to retrieve expert coding. Third, the zero-shot performance of GPT-4—the primary focus of the literature—falls between that of trained nonexpert coders. This intermediate performance underscores the necessity of fine-tuning to capture the nuanced demands of our coding scheme, aligning with recent work that suggests fine-tuning enhances performance compared to zero-shot methods (e.g., Halterman and Keith, 2025).

While LLMs offer comparatively high accuracy, we do not deny that this comes at a cost of less interpretability and replicability. Moreover, depending on the LLM, there are variations in customizable fine-tuning frameworks. For example, OpenAI's fine-tuning platform currently provides limited control over certain training hyperparameters, such as epochs and batch size. The training process is managed automatically by the system, which means scholars might not have the same level of fine-grained control as they might on other platforms. Despite these limitations, fine-tuned language models effectively recover expert annotations, consistently outperforming human coders, and are easy to produce.

3.2. Constructing expert team-based coding

Our training and test data are formed from press releases, media coverage, and public statements about trade policies released from 2010 to 2020 and collected via the MapLight Bill Position Database, with some retrieved from the Wayback Machine (Lorenz *et al.*, 2020). The dataset includes 130,000 policy positions taken by interest groups on thousands of bills between the 109th and 115th Congresses (2005–2018). The dataset contains 1,347 interest groups which have issued press releases, public statements, or have expressed their policy preferences toward trade policies on social media

Agreements as associative. For instance, if a paragraph mentions environmental issues as part of the provisions included in the trade deals, it is categorized as informative rather than associative.

[7]Note the strong performance is also consistent with a difficult coding task but highly-accurate gold standard data.

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

Table 1. Schematic description of codebook

Category	Definition	Example paragraph
Prediction	The expected consequence of a policy (e.g., price, wage, and welfare effects of trade)	"CTPA will remedy the unfairness of today's U.S.-Colombia trade relationship. In addition, the agreement will open services markets, secure the intellectual property of U.S. inventors, researchers, and creative artists, and introduce enforceable protections for worker rights and the environment." by *the U.S. Chamber of Commerce*
Status-quo	Contents of a given free trade agreement and descriptions of how a trade policy connects to the current state of economic/social structure	"OFTA also extends extraordinary foreign investor protections and rights beyond those provided in CAFTA (which in itself expanded on the NAFTA rules that resulted in 41 foreign investor challenges to U.S. health, environmental and other policies with over $35 million in government funds paid in damages to date) by *AFL-CIO*"
Specifics	Numeric estimates, expert witnesses or jargon (e.g., countervailing duty, tariff base)	"The U.S.-Oman FTA will build upon the Trade and Investment Framework Agreement signed between our two countries on July 7, 2004, and will spur continued growth of U.S. foreign direct investment, which in 2003 was $358 million, a substantial increase from $193 million in 2002."
Sociotropics	Statements spurring sociotropic perceptions of how peace, open societies or the U.S. economy as a whole are affected	"A comprehensive U.S. strategy to counter terrorism should include economic policies that encourage development, more open societies, and opportunities for people to improve the lives of their families and to enhance prospects for their children's future." by *Coca Cola Company.*
Environment	Appeals or generic information concerning environmental issues	"Peruvians are already losing significant tax revenue from illegal natural resource extraction as companies ravage the environment with impunity. This law weakens the government's ability to sanction businesses that do serious environmental damage." by *Environmental Investigation Agency.*
Human/Labor Rights	Statements about labor/human right issues that are not directly related to a policy	"The U.S.-Colombia Free Trade Agreement puts in jeopardy America's moral leadership by sacrificing the lives and livelihoods; the worker and human rights of the Colombian people at the altar of free trade." by *United Steelworkers (USW)*
Ideology	Political words (e.g., republican, democrats), recognizable politicians (e.g., Obama, Bush, etc.)	"We congratulate President Obama and President Lee on reaching an agreement which brings this important trade agreement closer to ratification in the U.S. Congress" by *Chevron*
Interest Groups	Endorsement by specific interest groups or industrial policy preference	"Chevron is a founding member of the U.S.-Korea Free Trade Agreement Business Coalition and worked closely with the U.S. negotiating team to ensure that strong investment protections were maintained. These protections create a stable environment for long term, high risk investments associated with the energy industry." by *Chevron*

platform between 2010 and 2020. We annotate documents at the paragraph level. For our expert coding, we adopt a team-based system. Using an initial codebook, the lead author and two paid, extensively trained, Princeton undergraduate annotators independently annotated 100 paragraphs a week, meeting weekly to discuss disagreements and develop consensus. The slow pace here is deliberate. The 3,192 paragraphs were coded in 390 hours, and the total time cost (including training) was 2.7 minutes per paragraph, per person. We split the ground-truth labels into a training set composed of 1,692 paragraphs, a validation set of 500 paragraphs, and a testing set of 1,000 paragraphs. All results are reported on the test set.

Separating these signals is inherently difficult, and domain expertise in public policy and interest groups turned out to be essential to capturing signals from political texts. For example, the word

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

"environment" may not mean environment-related signals that belong to *associative* category as trade agreements have included labor rights provisions or environmental provisions since NAFTA in 1994. If the term "environment" was used in the context of discussing the particular clauses in free trade agreements, it should be classified as a "status-quo" category under the scheme of informative signals. On the contrary, it should be annotated as an environment category if a paragraph is about the government's inability in partner countries to sanction businesses that do serious environmental damage. The codebook attempts to delineate subtle nuanced differences across categories, but it might still be a challenging task for those who are unfamiliar with this topic.

Another challenge in this task is identifying unique names of politicians and interest groups. Interest group advertising is much more nuanced as unions or associations frequently describe themselves using abbreviations. For instance, United Steelworkers, a general trade union with members across North America, represents itself as "USW," and the National Association of Manufacturers, a business association composed of 14,000 companies, has been described as "NAM" in most of their advertisements. Given that people have varying degrees of interest and backgrounds in politics, crowdsourced annotations may not be able to recognize politicians' names successfully.

The existence of such borderline cases adds complexity to measurement decisions. These instances demonstrate that this measurement task of quantifying *informative* signals requires additional domain-specific expertise. In the next section, we describe our supervised learning approach, compare it to nonexpert human coders, and provide evaluations demonstrating that expert-coding-based models can be made more accessible to researchers through LLMs.

3.3. Fine-tuning classifiers using transformer-based language models

In the following sections, we fine-tune GPT-3, GPT-3.5, GPT-4, RoBERTa, and Llama 3 (8B) to produce a binary classifier for each of the eight categories, describe our evaluation strategy, and compare our results to those of nonexpert human coders, including both lightly trained coders hired from Upwork, and untrained coders hired from Mechanical Turk (MTurk). We note that our selection of language models for this analysis includes both open-weight language models, which are free and available for anyone to access and distribute, such as Llama 3, and closed-source language models, such as those from OpenAI.[6]

LLMs are autoregressive transformer-based language models that can be used for classification. The models can detect subtle meanings that rely on context-specific meaning of words or real-world knowledge (Laurer *et al.*, 2022). The fine-tuning process uses the language generation function of the model to generate the binary label for the presence or absence of each category. Extensive testing showed that bag-of-words and approaches on fine-tuning RoBERTa (Liu *et al.*, 2019)—a smaller, open-source transformer-based language model that generates a contextualized representation of words—demonstrated that we could not achieve comparable performance. By conditioning on our 1,692 training paragraphs, the fine-tuning learns the criterion for selecting a label. The cost to fine-tune the model was $0.03 per 1,000 tokens, which amounts to approximately $17–$18 per category for GPT-3. For GPT-3.5, however, the cost has decreased to around $4–$7 per category with every reason to expect that costs will continue to fall quickly.

While natural language processing models have been consistently moving towards larger "foundation models," the social sciences have (until recently) been focused on comparatively simple bag-of-words models (Barbera´ *et al.*, 2021). This may be because more complex methods have not consistently outperformed the basics in social science tasks (although see Ornstein *et al.*, 2025; Ziems

[6]While models such as o1 or o3-mini have been recently released, GPT-4 remains the most recent version available from OpenAI for fine-tuning. We specifically used gpt-4o-2024-08-06 for GPT-4, gpt-3.5-turbo-0125 for GPT-3.5, and davinci for GPT-3. See Barrie, Palmer, and Spirling (Barrie *et al.*, 2024) and Palmer, Smith, and Spirling (Palmer *et al.*, 2024) for more on the trade-offs of open-weight vs. closed LLMs. We take no position on this debate here and simply demonstrate that both work well.

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

et al., 2024; Gilardi *et al.*, 2023, which are suggestive of nonexpert-human-comparable performance). While more and more social scientists are using LLMs in their research, it seems to primarily be in the zero-shot or few-shot setting that, as we will show, did not work for our task.

3.4. Nonexpert human coders

Our approach to gold-standard annotation using expert-led team coding is costly in terms of time and money, raising the question of whether it is feasible to use more nonexpert coders—a proven strategy for many other tasks (e.g., Snow *et al.*, 2008; Benoit *et al.*, 2016; Carlson and Montgomery, 2017; Ying *et al.*, 2022). We consider two strong baselines—a lightly trained and screened set of hourly workers recruited from Upwork and a screened set of MTurk crowdworkers. In each case, we tried to exceed the typical level of due-diligence for using nonexpert workers.

For the "lightly-trained" group, we used the Upwork platform to interview and hire eight people, providing them with the codebook and an initial test. We also provided feedback on their annotations during the first three weeks of the project to ensure they understood the codebook and the key quantity we expected them to annotate. Their hourly rate of pay was $15. We hired the top five performers of these eight and gave them detailed feedback on their coding. They then annotated the test set of 1,000 paragraphs. Because all five coders annotated every paragraph in the test set, we are also able to construct a majority annotation representing the label for each category that at least three coders agreed on.

The far more common strategy in the literature is the use of crowdworkers who are distinctive because they cannot reliably be trained, only screened. To provide the strongest results, we developed a custom interface that presented each paragraph, eight possible annotation choices, and abbreviated instructions. We restricted participation to MTurk Master workers located in the United States and emphasized in the instructions that results would be verified for accuracy and that multiple categories could apply to a given paragraph. Workers were eligible to complete a task only after answering a five-question screening test. We then selected the top 10 out of 25 participants from this test to label the full dataset. These workers averaged 58 seconds per row at a rate of $0.31 per row, yielding an effective hourly wage of around $19.

3.5. Evaluation

We evaluate the performance of our supervised fine-tuned GPT classifiers, Llama, RoBERTa, and the Upwork/MTurk human coders against our expert team-coded gold-standard data. Figure 2 shows average accuracy over all categories by the different classification approaches including the fine-tuned LLMs; bag-of-words, zero-shot GPT-4, and fine-tuned RoBERTa (Liu *et al.*, 2019) baselines; the five lightly trained Upwork coders; the majority Upwork worker choice and majority MTurk worker choice; and finally, simply guessing the modal class (i.e., 0). A summary of all methods included in the evaluation is provided in Table 2.

Because classes are imbalanced, in Figure 3, we present the best-performing fine-tuned GPT-3 model's results aggregated to the informative and associative categories, showing that precision and recall are both high and approximately comparable for both aggregate classes. Figure D.3 in the Appendix provides a comprehensive breakdown by category.

4. Applications: information supply by the U.S. chamber of commerce and United States trade representative

Next, we provide two illustrative examples of our information coding scheme in use. This serves two distinct purposes. First, it demonstrates how scholars of interest group politics might make use of our coding scheme and our specific trained classifier. Second, the documents that are in related domains

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

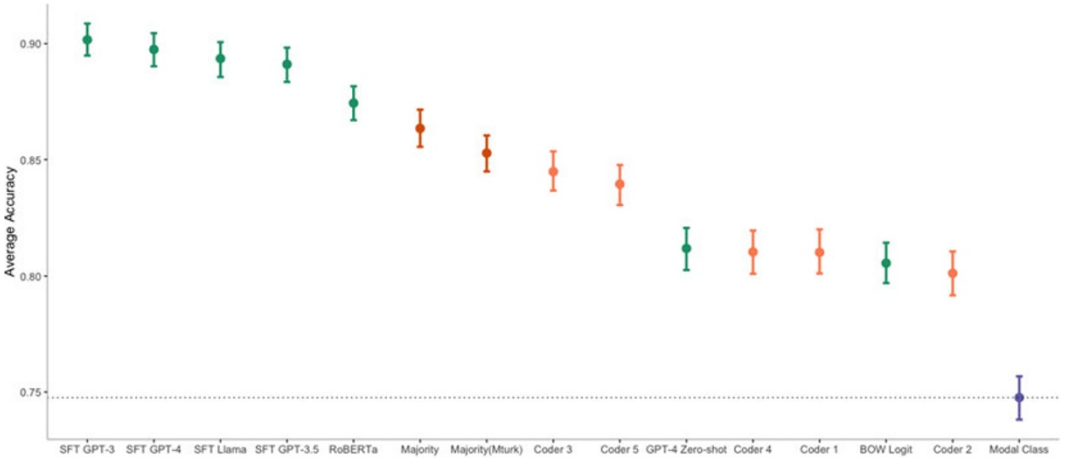


Figure 2. Average accuracy over the eight categories in our coding scheme by method. Green indicates machine performance, while red and orange indicate human coder performance. See Table 2 for a description of the methods compared.

Table 2. Summary of classification methods compared in Figure 2

Method	Description	Category
SFT GPT-3	Supervised fine-tuned GPT - 3	Machine
SFT GPT-3.5	Supervised fine-tuned GPT-3.5	Machine
SFT GPT-4	Supervised fine-tuned GPT-4	Machine
SFT LLaMA	Supervised fine-tuned LLaMA 3	Machine
RoBERTa	Supervised fine-tuned transformer-based open-source model Machine	
BOW Logit	Bag-of-words logistic regression model	Machine
GPT-4 zero-shot	GPT-4 with no fine-tuning (zero-shot inference)	Machine
Coder 1–5	Individual trained nonexpert Upwork workers	Human (trained)
Majority	Majority vote across Coders 1–5 (Upwork)	Human (trained)
Majority (MTurk)	Majority vote of untrained crowdworkers from MTurk	Human (untrained)
Modal class	Always predicts the most frequent class	Baseline

but are *not from the same distribution* as our training/test sets, demonstrating that our classifier can generalize (at least somewhat) out of distribution.

We focus on (1) a comparison between press releases and letters to Congress from the U.S. Chamber of Commerce, and (2) a comparison of testimony at congressional hearings and press releases distributed by the United States trade representative (USTR). We deliberately focus on political advertising disseminated within the same timeframe to make sure the comparability of signal compositions across different institutional contexts.[8] The relatively modest size of these corpora allows us to conduct manual validation of results. These manual evaluations also allow us to provide confidence intervals that account for annotation error using design-based supervised learning (Egami *et al.*, 2024).

[8]For the USTR, we concentrate on testimony and press releases produced between 2013 and 2014, while for the U.S. Chamber of Commerce, our sample comprises press releases and congressional letters from 2012 to 2014.

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

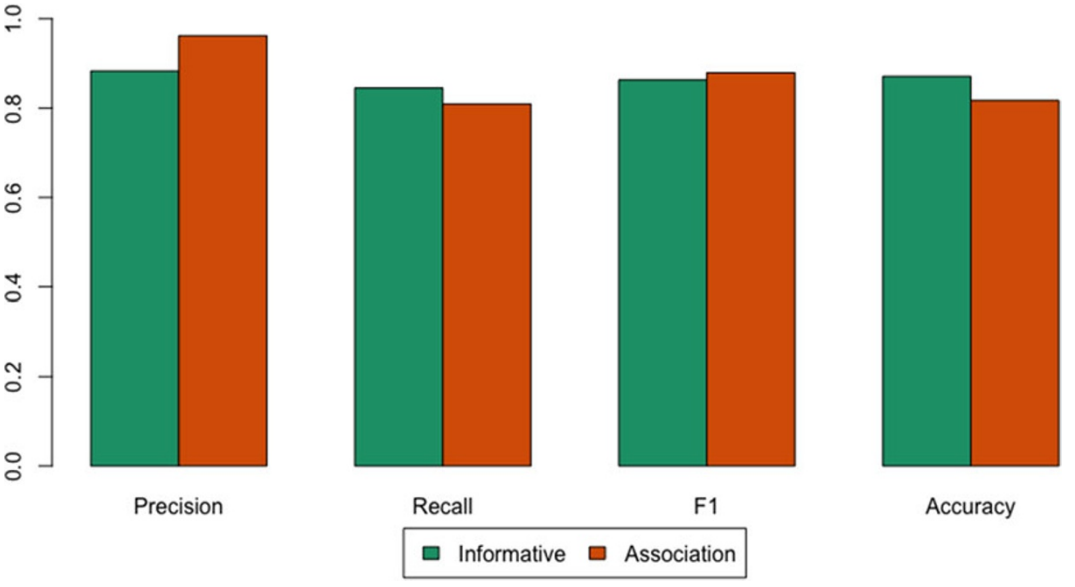


Figure 3. Measures of fine-tuned GPT-3 precision, recall, F1 and accuracy on whether any signal of informative/associative type is present.

4.1. Press releases and congressional letters from the U.S. chamber of commerce

Members of Congress make complex policy decisions with limited time and letters to Congress help them decide which policy proposals are worth supporting (Hall and Deardorff, 2006; Box-Steffensmeier *et al.*, 2019). Therefore, in addition to issuing press releases aimed at the general public, interest groups frequently send letters to Congress to share targeted information. Since letters to Congress are less frequent compared to press releases, we sample a similar-sized corpus from the latter to facilitate a better comparison of two different setups of political advertising. Subsequently, we generate predictions for eight categories using the model constructed in the previous section. By documenting how signal composition varies based on the purpose of in-formation provision, our application highlights differences in the strategic use of signals in shaping public opinion and the legislative process. Figure 4 illustrates the overall composition of signals across document types and shows that the proportion of informative signals in the two types of documents—press releases and letters to Congress—is comparable.

Among associative signals, sociotropic appeals appear more frequently in letters to Congress than in other categories. One explanation for this pattern is that framing trade gains as benefits to the U.S. economy at large offers a persuasive shortcut, given that politicians generally want to view themselves as serving the public interest. As a result, sociotropic cues become particularly effective signals. While the Chamber uses other associative signals related to the environment or human and labor rights in press releases, these signals are not recognizable in letters to Congress. Another interesting pattern to note is that the Chamber emphasizes its endorsements more in letters to Congress compared to the relatively less noticeable interest group cues in press releases. This empirical pattern resonates with previous work showing that members of Congress rely on endorsements from interest groups as important sources of information (e.g., McDermott, 2006; Arceneaux and Kolodny, 2009; Dalton, 2013; Dowling and Wichowsky, 2013), while citizens might be unfamiliar with these same interest groups (Druckman, 2005; Broockman *et al.*, 2021) leaving the cues comparatively ineffective.

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

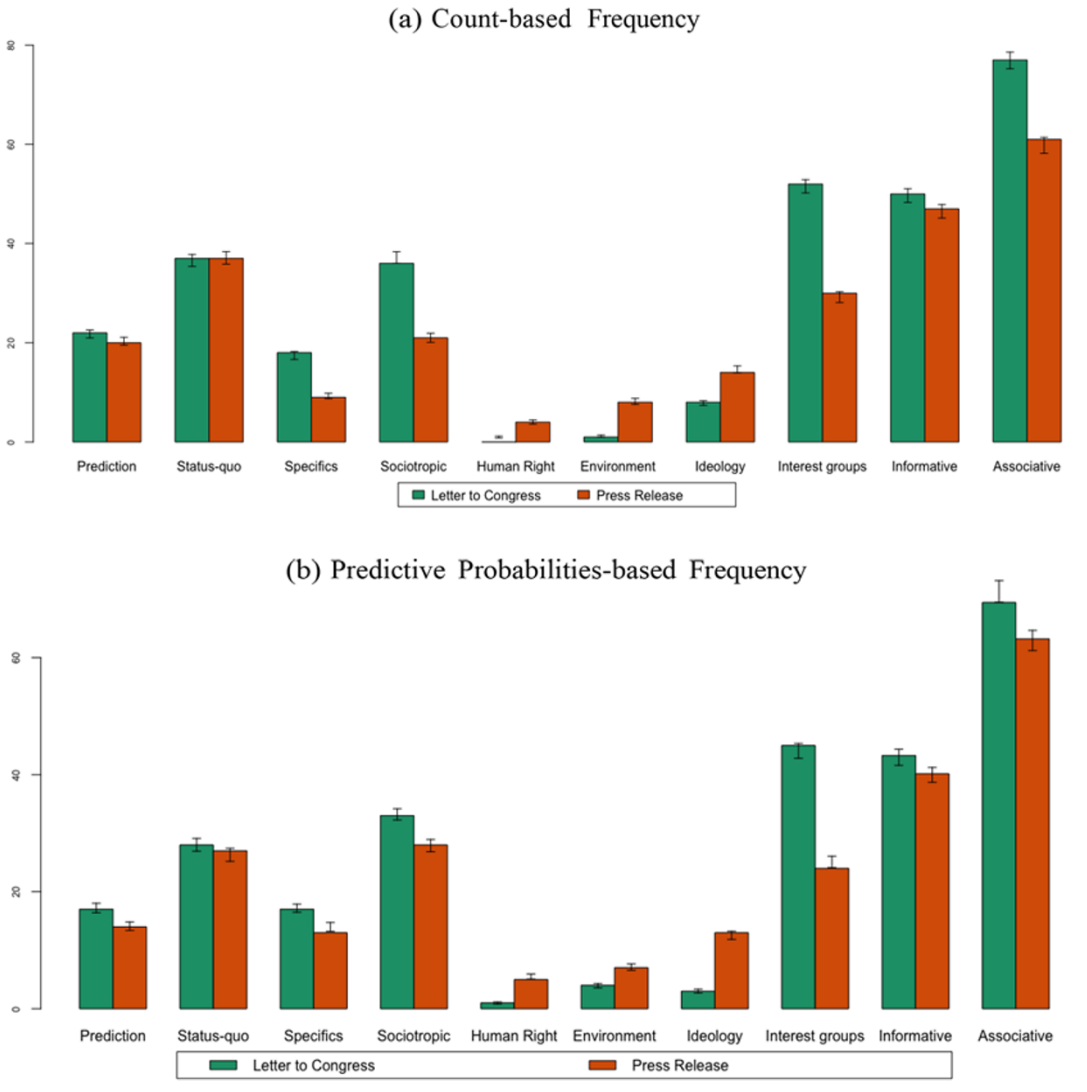


Figure 4. Signal Composition (U.S. Chamber of Commerce). Error bars indicate the confidence intervals estimated from the doubly robust estimation, which integrates expert coding and surrogate labels, calculated using the DSL package (Egami *et al.*, 2024) on February 15, 2025.

4.2. USTR congressional hearings, testimony, and press releases

In our second example, we use a corpus of USTR testimony at congressional hearings and press releases. The USTR is a government agency responsible for coordinating trade policy and advising the President on trade-related issues, and Congressional committee hearings have served as a primary forum for communication and information flow among legislators, interest groups, and bureaucrats (e.g., Quirk and Bendix, 2011; McGrath, 2013). Prior work and case studies have demonstrated that policy outcome has been affected by the information that is aired and discussed at congressional hearings (Burstein, 1998). We focus on the finalized transcripts of USTR's activities in congressional hearings between 2013 and 2014 and explore how the proposed approaches can be applied to other domains—information supply by bureaucrats. Since invitations to congressional hearings are not very frequent, we randomly sampled a similar-sized corpus from the press releases distributed by

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

the USTR between 2013 and 2014 for comparability.[9] By comparing USTR congressional hearing testimony and press releases, we evaluate how bureaucrats tailor their signals for politicians versus those intended for the broader public.

Figure 5 illustrates the composition of signals in these documents. First, informative signals—including prediction, status quo, and specifics categories—appear at levels comparable to sociotropic appeals. Second, we observe a significant increase in the number of ideological cues in the testimony. This trend likely reflects that Congress aims to gather political information in addition to substantive issue knowledge. Speaking to the broader literature on the role of information in legislative organization (Bradley, 1980; Esterling, 2007, 2009), the patterns we observe lend support to Ban *et al.* (2023)'s recent finding that members of Congress rely less on informative signals when policy disagreements exist between the legislative and executive branches. In Appendix D, we provide an additional validation of these results using Upwork coders. See Table D.6 for details.

Beyond this application, our method also offers a foundation for a range of broader empirical inquiries into political science. By producing paragraph-level classifications of rhetorical strategies at scale, the approach opens up new possibilities for analyzing how interest groups communicate under varying institutional contexts. For example, scholars could examine whether groups shift from informative to associative signaling depending on partisan control, bureaucratic capacity, or stages of the policymaking process (Balla, 1998; Balla and Wright, 2001; Potter, 2019). Researchers might also explore whether certain signal types are more common in public comments that receive agency responses or in rules that are ultimately adopted (Yackee, 2006). Additionally, it can be applied to document sets attributed to different kinds of organizations, enabling researchers to assess how signaling strategies vary across group types—such as business versus advocacy groups—or institutional venues, such as legislative versus regulatory settings (Hojnacki and Kimball, 1998; Baumgartner and Leech, 2001). In short, while the present paper is primarily methodological, we see this framework as providing a valuable tool for studying lobbying strategies and information flows in institutional environments characterized by asymmetric information (Lupia and McCubbins, 1998; Hall and Deardorff, 2006).

5. Discussion and limitations

We tackle the persistent and difficult problem in the study of political science that involves the measurement of information. This is unfortunate because a measure of information would be useful for testing and developing theories in interest group politics, especially given that existing studies typically emphasize the strategic provision of information by these groups within institutional contexts. This work helps clarify the underlying dynamics of information supply by interest groups addressing this challenge. This effort is important for crafting general theories of interest group influence on politics, which requires further sophistication of what types of signals are preferred for whom, and how strategically interest groups adjust their behavior in response to different institutional settings.

To address the inherent difficulties associated with measuring the notoriously complex concept of *informative* signals, we propose a coding scheme guided by the literature, drawing on the distinctions outlined in Martin (2020) and Lupia (1994) to connect theoretical discussions to empirical realities. Signals were selected as they represent significant areas of association that interest groups invoke through political appeals. By its nature, this is an intrinsically challenging measurement task, requiring context-specific knowledge and specialized expertise to build case-specific models. We first provide expert coding constructed by coders with domain expertise in political science and economics. We then provide a classifier for quantifying the informative signals from political texts. We demonstrate that the task is too difficult for lightly-trained nonexpert human coders and for

[9] In total, these are 23 documents, each with about 15 paragraphs on average.

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

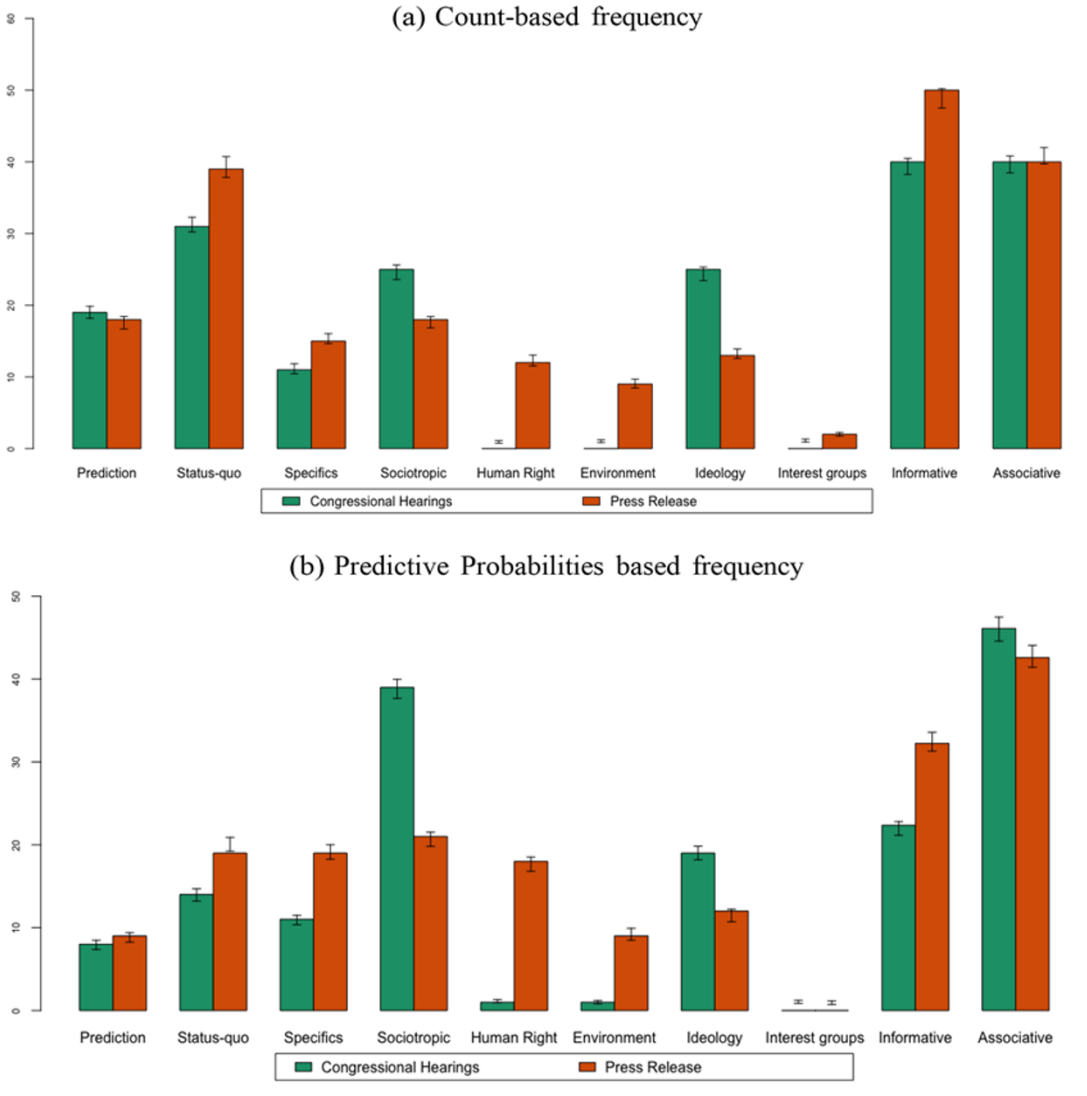


Figure 5. Signal composition (USTR). Error bars indicate the confidence intervals estimated from the doubly robust estimation, which integrates expert coding and surrogate labels, calculated using the DSL package (Egami *et al.*, 2024).

a benchmark BERT-based fine-tuning approach. Despite the difficulty of this task, we show that fine-tuned LLMs are able to yield strong performance comparable to expert coding. We then demonstrate on two small out-of-distribution illustrative examples that our classifier can extrapolate to the domains of real problems in the social sciences.

Our demonstration that LLMs can successfully mimic expert coding provides a compelling case for researchers to adopt expert-based measurement models. This approach addresses concerns regarding the accessibility of domain experts and mitigates resource constraints, particularly in terms of time and costs. While promising, the proposed methods also have a few limitations. First is stability—we access the model through an API that could change at any time (Barrie *et al.*, 2024). Although extensive fine-tuning should keep the result stable, exact computational reproducibility will likely not be possible with these results. Yet, we take these methods seriously because the fine-tuned

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

models maintain consistently higher performance with low variance. Additionally, we fine-tuned the open-source LLM, Llama 3, which also demonstrated strong and consistent results and could be used to achieve exact reproducibility for users with sufficient compute resources. Second, while the accuracy is relatively high (and consistent with results in the social science literature) (Benoit *et al.*, 2016; Barbera´ *et al.*, 2021; Jerzak *et al.*, 2023), working with machine-learned variables poses serious challenges (Fong and Tyler, 2021; Knox *et al.*, 2022). This is especially true when applying the classifier out-of-distribution, as researchers using our model would need to do (and as we did in our examples). The recently developed design-based supervised learning of Egami *et al.* (2024) provides us with a way to perform bias correction in a new domain with a relatively small validation set.

Finally, one might also be concerned that the expert coding here can also be subjective. We address this concern through two complementary strategies. First, we mitigate subjectivity by employing a detailed codebook and a team-based annotation procedure designed to ensure consistency. As additional suggestive evidence, we present in the Appendix that while the Upwork coders don't always concentrate on the correct answer, the majority usually favors the correct answer, providing some evidence that the 'wisdom of crowds" is finding the relevant signal (although at a substantially higher cost). We have investigated cases where the Upwork coders disagree with the ground-truth coding and consistently have sided with the ground-truth. Appendix D provides an error analysis of disagreements between Upwork coders and our team-based coding and is consistent with our expert coding being accurate.[10] Second, we conducted a separate expert annotation exercise using a new team of three expert coders, each holding postgraduate degrees in political science and currently employed at R1 research institutions. These coders followed the same annotation procedure as in the original coding process, using the same codebook to annotate a randomly sampled subset of 202 paragraphs from the testing set. When we evaluated the performance of the fine-tuned LLM, GPT-4, lightly trained coders, and the majority-vote Upwork annotations against this independently generated test set, we found that the fine-tuned LLM consistently outperformed all alternatives. Full details are provided in Appendix E. This cross-validation with an entirely separate expert-annotated dataset offers further evidence of the robustness of our approach and the reliability of the model's performance.

Our approach invites a range of substantive extensions. By applying this framework to different types of documents—such as interest group statements, agency reports, or legislative testimony—researchers can systematically identify what kinds of signals political actors deploy and how these signals might shape the beliefs and behavior of target audiences. While our approach provides a practical way to classify and scale expert-level judgments about political information, further work is needed to adapt this framework to other domains and refine typologies suited to different policy areas. We see this study as an initial step toward improving both the measurement and accessibility of expert-coding in political text.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2025.10086. To obtain replication material for this article, please visit https://doi.org/10.7910/DVN/G40KT8

Acknowledgements. We thank Kenneth Lowande for sharing the data on expert and nonexpert coding. We are grateful to Arthur Lupia and Erin Rossiter for their discussions about this project. We also thank Marilena Zigka, Jongnam Park, and five workers on Upwork for their research assistance as coders. We benefited from the helpful comments from audiences at the 39th Annual Meeting of the Society for Political Methodology at Washington University in St. Louis and the American Political Science Association Conference in Los Angeles. We are also grateful for excellent feedback from the editor and anonymous reviewers.

Funding. This work was supported by the Data-Driven Social Science Initiative at Princeton University.

Conflict of interest. The authors declare no ethical issues or conflicts of interest in this research.

[10]Figures D.1 and D.2 show that the majority of Upworkers consistently agreed with most of our ground-truth labels, while individual Upworkers displayed varying degrees of accuracy.

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

References

Adcock R and Collier D (2001) Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review* **95**(3), 529–546.

Arceneaux K and Kolodny R (2009) Educating the least informed: group endorsements in a grassroots campaign. *American Journal of Political Science* **53**(4), 755–770.

Balla SJ (1998) Administrative procedures and political control of the bureaucracy. *American Political Science Review* **92**(3), 663–673.

Balla SJ and Wright JR (2001) Interest groups, advisory committees, and congressional control of the bureaucracy. *American Journal of Political Science* **45**(4), 799–812.

Ban P, Park JY and You HY (2023) How are politicians informed? wit-nesses and information provision in congress. *American Political Science Review* **117**(1), 122–139.

Barbera´ P, Boydstun AE, Suzanna L, Ryan M and Jonathan N (2021) Automated text classification of news articles: a practical guide. *Political Analysis* **29**(1), 19–42.

Barrie C, Palmer A and Spirling A. 2024. Replication for language models problems, principles, and best practice for political science. Retrieved November 4, 2025. https://arthurspirling.org/documents/BarriePalmerSpirlingTrustMeBro.pdf

Baumgartner FR and Leech BL (2001) Interest niches and policy bandwagons: patterns of interest group involvement in national politics. *The Journal of Politics* **63**(4), 1191–1213.

Bechtel MM, Bernauer T and Meyer R (2012) The green side of protectionism: environmental concerns and three facets of trade policy preferences. *Review of International Political Economy* **19**(5), 837–866.

Bechtel MM, Hainmueller J and Margalit Y (2014) Preferences for international redistribution: the divide over the Eurozone bailouts. *American Journal of Political Science* **58**(4), 835–856.

Benoit K, Conway D, Lauderdale BE, Laver M and Mikhaylov S (2016) Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Review* **110**(2), 278–295.

Bombardini M and Trebbi F (2020) Empirical models of lobbying. *Annual Review of Economics* **12**(1), 391–413.

Box-Steffensmeier JM, Christenson DP and Craig AW (2019) Cue-taking in congress: interest group signals from dear colleague letters. *American Journal of Political Science* **63**(1), 163–180.

Bradley RB (1980) Motivations in legislative information use. *Legislative Studies Quarterly* **5**(3), 393–406.

Broockman DE, Kaufman AR and Lenz GS. 2021. *Heuristic projection: How interest group cues can harm voters' judgements*. Technical report. Working Paper. https://osf.io/6yskq.

Brutger R and Guisinger A (2022) Labor market volatility, gender, and trade preferences. *Journal of Experimental Political Science* **9**(2), 189–202.

Burstein P (1998) *Discrimination, Jobs, and Politics: The Struggle for Equal Employment Opportunity in the United States since the New Deal*. Chicago, IL: University of Chicago Press.

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

Catalinac A (2014) Quantitative text analysis with Asian languages: some problems and solutions. *Polimetrics* **I**(1), 14–17.

Chaiken S (1980) Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology* **39**(5), 752.

Cheng C, Barcelo J, Hartnett AS, Robert K and Luca M-S (2020) Covid-19 government response event dataset (coronanet v. 1.0). *Nature Human Behaviour* **4**(7), 756–768.

Chernykh S, Doyle D and Power TJ (2017) Measuring legislative power: an expert reweighting of the fish-kroenig parliamentary powers index. *Legislative Studies Quarterly* **42**(2), 295–320.

Clinton JD and Lewis DE (2008) Expert opinion, agency characteristics, and agency preferences. *Political Analysis* **16**(1), 3–20.

Coppock A (2023) *Persuasion in Parallel: How Information Changes Minds about Politics*. Chicago, IL: University of Chicago Press.

Dalton RJ (2013) *Citizen Politics: Public Opinion and Political Parties in Advanced Industrial Democracies*. Washington, DC: CQ Press.

Di Cocco J and Monechi B (2021) How populist are parties? measuring degrees of populism in party manifestos using supervised machine learning. *Political Analysis* **30**(3), 311–327.

Diermeier D, Godbout J-FC, Yu B and Kaufmann S (2012) Language and ideology in congress. *British Journal of Political Science* **42**(1), 31–55.

Dowling CM and Wichowsky A (2013) Does it matter who's behind the curtain? anonymity in political advertising and the effects of campaign finance disclosure. *American Politics Research* **41**(6), 965–996.

Druckman JN (2005) Does political information matter? *Political Communication* **22**(4), 515–519.

Druckman JN, Kifer MJ and Parkin M (2020) Campaign rhetoric and the incumbency advantage. *American Politics Research* **48**(1), 22–43.

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

Egami N, Hinck M, Stewart B and Wei H (2024) Using imperfect surrogates for downstream inference: design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems* **36**, 1–14.
Esterling K (2007) Buying expertise: campaign contributions and attention to policy analysis in congressional committees. *American Political Science Review* **101**(1), 93–109.
Esterling K (2009) *The Political Economy of Expertise: Information and Efficiency in American National Politics*. Ann Arbor, MI: University of Michigan Press.
Fong C (2020) Expertise, networks, and interpersonal influence in congress. *The Journal of Politics* **82**(1), 269–284.
Fong C and Tyler M (2021) Machine learning predictions as regression co-variates. *Political Analysis* **29**(4), 467–484.
Fowler EF, Franz MM, Martin GJ, Peskowitz Z and Ridout TN (2021) Political advertising online and offline. *American Political Science Review* **115**(1), 130–149.
Franchino F (2004) Delegating powers in the European community. *British Journal of Political Science* **34**(2), 269–293.
Gilardi F, Alizadeh M and Kubli M (2023) Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* **120**(30), e2305016120.
Grimmer J, Roberts ME and Stewart BM (2022) *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton, NJ: Princeton University Press.
Grossman GM and Helpman E (2001) *Special Interest Politics*. Cambridge, MA: MIT press.
Guisinger A (2017) *American Opinion on Trade: Preferences without Politics*. New York, NY: Oxford University Press.
Hall RL and Deardorff AV (2006) Lobbying as legislative subsidy. *American Political Science Review* **100**(1), 69–84.
Halterman A and Keith KA (2025) Codebook llms: evaluating llms as measurement tools for political science concepts. *Political Analysis* 1–17. https://doi.org/10.1017/pan.2025.10017
Hojnacki M and Kimball DC (1998) Organized interests and the decision of whom to lobby in congress. *American Political Science Review* **92**(4), 775–790.
Huber J and Inglehart R (1995) Expert interpretations of party space and party locations in 42 societies. *Party Politics* **1**(1), 73–111.
Jerzak CT, King G and Strezhnev A (2023) An improved method of automated nonparametric content analysis for social science. *Political Analysis* **31**(1), 42–58.
Kalla JL and Broockman DE (2018) The minimal persuasive effects of campaign contact in general elections: evidence from 49 field experiments. *American Political Science Review* **112**(1), 148–166.
King G, Pan J and Roberts ME (2013) How censorship in China allows government criticism but silences collective expression. *American Political Science Review* **107**(2), 326–343.
Knox D, Lucas C and Cho WKT (2022) Testing causal theories with learned proxies. *Annual Review of Political Science* **25**, 419–441.
Kollman K (2021) *Outside Lobbying*. Princeton, NJ: Princeton University Press.
Kuklinski JH, Quirk PJ, Jerit J, Schwieder D and Rich RF (2000) Misinformation and the currency of democratic citizenship. *Journal of Politics* **62**(3), 790–816.
Laurer M, van Atteveldt W, Casas A and Welbers K (2022) Less annotating, more classifying: addressing the data scarcity issue of supervised machine learning with deep transfer learning and bertnli. *Political Analysis* **32**(1), 84–100.
Laver M, Benoit K and Garry J (2003) Extracting policy positions from political texts using words as data. *American Political Science Review* **97**(2), 311–331.
Liu Y, Ott M, Goyal N, Jingfei Du J, et al. 2019. Roberta: a robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*
Lorenz GM, Furnas AC and Crosson JM (2020) Large-n bill positions data from maplight. org: what can we learn from interest groups' publicly observable legislative positions? *Interest Groups & Advocacy* **9**, 342–360.
Lovett MJ, Peress M, et al. (2015) Targeting political advertising on television. *Quarterly Journal of Political Science* **10**(3), 391–432.
Lowande K (2024) *False Front: The Failed Promise of Presidential Power in a Polarized Age*. Chicago, IL: University of Chicago Press.
Lowande K and Shipan CR (2022) Where is presidential power? Measuring presidential discretion using experts. *British Journal of Political Science* **52**(4), 1876–1890.
Lupia A (1994) Shortcuts versus encyclopedias: information and voting behavior in California insurance reform elections. *American Political Science Review* **88**(1), 63–76.
Lupia A and McCubbins MD (1998) *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge, UK: Cambridge University Press.
Mansfield ED and Mutz DC (2009) Support for free trade: self-interest, sociotropic politics, and out-group anxiety. *International Organization* **63**(3), 425–457.
Marquardt KL, Pemstein D, Petrarca CS, Seim B, Wilson SL, Bernhard M, Coppedge M and Lindberg SI (2025) Experts, coders and crowds: an analysis of substitutability. *International Political Science Review* **46**(4), 622–638.
Martin GJ 2020. The informational content of campaign advertising.

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press

McDermott ML (2006) Not for members only: group endorsements as electoral information cues. *Political Research Quarterly* **59**(2), 249–257.
McGrath RJ (2013) Congressional oversight hearings and policy control. *Legislative Studies Quarterly* **38**(3), 349–376.
Milner HV and Tingley DH (2011) Who supports global economic engagement? the sources of preferences in American foreign economic policy. *International Organization* **65**(1), 37–68.
Mondak JJ (1993) Public opinion and heuristic processing of source cues. *Political Behavior* **15**(2), 167–192.
Mutz DC (1992) Mass media and the depoliticization of personal experience. *American Journal of Political Science* **36**(2), 483–508.
Mutz DC and Mondak JJ (1997) Dimensions of sociotropic behavior: group-based judgements of fairness and well-being. *American Journal of Political Science* **41**(1), 284–308.
Ornstein JT, Blasingame EN and Truscott JS (2025) How to train your stochastic parrot: large language models for political texts. *Political Science Research and Methods* **13**(2), 264–281.
Palmer A, Smith NA and Spirling A (2024) Using proprietary language models in academic research requires explicit justification. *Nature Computational Science* **4**(1), 2–3.
Park JY and Montgomery JM 2025. Towards a framework for creating trustworthy measures with supervised machine learning. *Working Paper*.
Potter RA (2019) *Bending the Rules: Procedural Politicking in the Bureaucracy*. Chicago, IL: University of Chicago Press.
Quinn KM, Monroe BL, Colaresi M, Crespin MH and Radev DR (2010) How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* **54**(1), 209–228.
Quirk PJ and Bendix W (2011) *Deliberation in Congress*. Schickler E and Lee FE Edited by New York, NY: Oxford University Press, 550–574.
Slapin JB and Proksch S-O (2008) A scaling model for estimating time-series party positions from texts. *American Journal of Political Science* **52**(3), 705–722.
Sniderman PM, Brody RA and Tetlock PE (1993) *Reasoning and Choice: Explorations in Political Psychology*. Cambridge, UK: Cambridge University Press.
Snow R, O'connor B, Jurafsky D and Ng AY (2008) Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. *In Proceedings of the 2008 conference on empirical methods in natural language processing*, 254–263.
Spenkuch JL and David T (2018) Political advertising and election results. *The Quarterly Journal of Economics* **133**(4), 1981–2036.
Sumner JL, Farris EM and Holman MR (2020) Crowdsourcing reliable local data. *Political Analysis* **28**(2), 244–262.
Treier S and Jackman S (2008) Democracy as a latent variable. *American Journal of Political Science* **52**(1), 201–217.
Weber C, Dunaway J and Johnson T (2012) It's all in the name: source cue ambiguity and the persuasive appeal of campaign ads. *Political Behavior* **34**, 561–584.
Yackee SW (2006) Sweet-talking the fourth branch: the influence of interest group comments on federal agency rulemaking. *Journal of Public Administration Research and Theory* **16**(1), 103–124.
Ying L, Montgomery JM and Stewart BM (2022) Topics, concepts, and measurement: a crowdsourced procedure for validating topics as measures. *Political Analysis* **30**(4), 570–589.
Ziems C, Held W, Shaikh O, Chen J, Zhang Z and Yang D (2024) Can large language models transform computational social science? *Computational Linguistics* **50**(1), 237–291.

Cite this article: Choi D, Peskoff D and M Stewart B (2026) Fine-tuned large language models can replicate expert coding better than trained coders: a study on informative signals sent by interest groups. *Political Science Research and Methods*, 1–19. https://doi.org/10.1017/psrm.2025.10086

https://doi.org/10.1017/psrm.2025.10086 Published online by Cambridge University Press