Observational Equivalence of LLM and Human Annotation*

Kentaro Nakamura[†] Jing Ling Tan[‡] George Yean[§]

Last Updated: September 1, 2026

Abstract

In this paper, we show that LLM and human coding are observationally equivalent in terms of annotation quality: recent LLMs agree with expert coders at rates comparable to those observed among experts. We demonstrate this through replications of text-classification tasks from 14 peer-reviewed political science studies, in which ten LLMs, three human experts, and 165 crowdsourced workers independently classify the same texts using identical codebooks. We find that this equivalence is driven by ambiguity in the texts and coding rules. When LLMs disagree with experts, experts are also more likely to disagree with one another, and clarifying coding rules reduces disagreement among both experts and recent LLMs. Thus, there is little empirical basis for preferring human coding on the basis of annotation quality alone, while LLMs offer substantial advantages in speed and cost. We therefore argue that the central challenge of text annotation is no longer choosing between human and machine coders, but developing coding rules that minimize ambiguity and accounting for the ambiguity that remains. To this end, we propose using disagreement across LLMs to identify difficult cases and refine codebooks, and we develop ambiguity-aware bounds for downstream inference when a unique annotation cannot be defined for every text.

Keywords: large language models, automated text analysis, Text-as-Data, AI

*We thank Antonio Camara, Stephen Chaudoin, Danny Ebanks, Valentina González-Rostani, Musashi Hinck, Connor Jerzak, Gary King, Dean Knox, Eddie Yang, and participants of PolMeth2026 for helpful comments. We also gratefully acknowledge financial support from the Institute for Quantitative Social Science (IQSS).

[†]Ph.D. student at John F. Kennedy School of Government, Harvard University, Email: knakamura@g.harvard.edu

[‡]Ph.D. student at Department of Government, Harvard University, Email: jingling_tan@g.harvard.edu

[§]Ph.D. student at Department of Government, Harvard University, Email: gyean@g.harvard.edu

1 Introduction

Large language models (LLMs) are changing quantitative text analysis by applying natural-language instructions directly to raw text. Earlier text-as-data methods typically relied on dictionaries, bag-of-words representations, or restrictive parametric models that abstracted away from how people interpret language (Grimmer and Stewart, 2013; Grimmer et al., 2021, 2022). Because these methods were understood as imperfect approximations to human reading, researchers generally evaluated automated outputs against carefully hand-coded labels and interpreted disagreement as algorithmic error. LLMs call this human-centered validation paradigm into question. A single prompt can provide concept definitions, examples, and contextual information, allowing LLMs to apply a codebook in ways that increasingly resemble human coders. If LLMs can apply the same coding rules about as reliably as humans, however, disagreement with a human label need not indicate model error. This raises two questions: Can LLMs annotate political texts as reliably as careful human coders, and what does their performance imply for how researchers should conduct and validate quantitative text analysis?

In this paper, we evaluate LLM-based annotation through replications of binary text-classification tasks from 14 peer-reviewed political science studies. For each study, ten LLMs and three expert coders independently classify the same texts using identical codebooks, allowing us to compare LLM–expert agreement directly with agreement among experts themselves. We find that, with the exception of an older and weaker model, recent LLMs agree with expert coders at rates comparable to the agreement observed among experts. Across studies, LLM–expert agreement closely tracks expert–expert agreement: when experts agree more with one another, LLMs also agree more with experts, and when expert agreement declines, LLM–expert agreement declines as well. In this sense, LLM and human coding are observationally equivalent in annotation quality.

What explains this observational equivalence? We find that disagreement among different annotators is systematically concentrated in the same texts. When two experts disagree about a text, the other expert is also substantially more likely to disagree, and the same pattern is observed for both LLMs and crowd-sourced workers (although crowd-worker's overall agreement with experts is substantially lower than LLMs). These patterns point to a common source of disagreement: ambiguity in the text or coding rule rather than errors specific to any one type of annotator. Consistent with this interpretation, clarifying an underspecified codebook substantially reduces disagreement among experts and sufficiently capable LLMs.

These findings change how researchers should validate text annotations. When an LLM disagrees with a human coder, the disagreement should not automatically be interpreted as model error, because careful human coders are often divided on the same cases. Thus, the central challenge of text annotation is not simply choosing between human experts and LLMs based on annotation quality, but developing coding rules that minimize ambiguity and accounting for the ambiguity that remains. Importantly, the ambiguity is concentrated in particular texts rather than randomly distributed across a corpus. Consequently, annotation errors can be systematically related to textual content rather than classical measurement error. In some cases, the problem is even more fundamental: when the codebook admits multiple defensible classifications or texts do not provide sufficient information or contexts, a unique target annotation may not exist at all.

Building on this reinterpretation, we propose a practical workflow for text annotation with LLMs. Researchers should first develop a clear codebook and manually apply it to a small random sample of texts to identify common ambiguities and refine the coding rules. They should then verify that candidate LLMs can understand and reliably apply the codebook, and use disagreement across multiple suitable LLMs to identify additional borderline cases for targeted review before annotating the full corpus. Because some ambiguity may remain even after careful codebook development, we also develop ambiguity-aware bounds for downstream inference. These bounds use disagreement across LLMs as a scalable proxy for residual ambiguity and allow researchers to conduct inference without assuming that every text has a uniquely defined ground-truth annotation, an assumption underlying existing methods for correcting annotation error (Angelopoulos

et al., 2023; Egami et al., 2024).

The contribution of this paper is summarized as follows. First, recent studies have questioned the use of human annotations as a gold standard and argued that humans and LLMs should instead be treated as alternative annotators whose reliability can be compared directly (Bisbee and Spirling, 2026; Clark et al., 2021; Hosking et al., 2024; Yang et al., 2025). We extend this literature by examining the source of disagreement and show that disagreement among experts, LLMs, and crowd-sourced workers is systematically concentrated in the same texts and decreases when coding rules are clarified. Second, we develop an ambiguity-aware framework for text annotation and downstream inference. We show that disagreement across LLMs provides an informative and scalable proxy for ambiguity, which researchers can use to identify difficult cases and refine codebooks. For the ambiguity that remains, we develop ambiguity-aware bounds for downstream inference, which allows the target annotation to be non-unique under a given codebook. This framework relaxes the assumption in existing methods that a uniquely defined target annotation exists for every text (Angelopoulos et al., 2023; Egami et al., 2023, 2024).

2 Paradigms of Quantitative Text Analysis: Before and After LLMs

2.1 Goal and Philosophy of Text Annotation

The goal of content analysis is to make "replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use" (Krippendorff 2018, p.24). In this respect, coding text is not different in kind from measurement elsewhere in the social sciences. In the conceptual model of measurement proposed by Adcock and Collier (2001), researchers first have a *background concept*, the broad constellation of meanings associated with a term such as "democracy" or "security", and then formulate them into a *systematized concept*, the specific formulation the researcher adopts for a given study. The operationalization needs to track the systematized concept (validity) and can be reproduced by different experts (reliability).

Text annotation is a special case of exactly this process, but what makes text unique is its unstructured and high-dimensional nature. By unstructured, we mean that the concept of interest is not directly observed. Instead, it must be inferred from language. As Krippendorff (2018) emphasizes, texts have no "reader-independent" qualities: their meaning is not contained in words but is brought to them by readers. Thus, two careful coders can therefore read the same passage differently, not because one has made an error, but because text genuinely admits more than one interpretation. High dimensionality compounds this problem. As a corpus grows, the space of possible textual configurations grows with it, continually producing borderline and outlier cases (Grimmer et al., 2022). The ambiguity that enters at the operationalization stage therefore cannot be fully eliminated in many cases, no matter how carefully the codebook is written.

This also points to an inherent tension between reliability and validity. One could push annotation toward the structured-data ideal by operationalizing a concept as a mechanical, easily replicated procedure, such as counting occurrences of a keyword. Such rules are highly reliable and trivially replicable, but they typically sacrifice validity to the systematized concept (Kracauer, 1952). Moving in the opposite direction, i.e., toward richer, more holistic human judgment, improves validity but lowers reliability, precisely because it leaves more to the reader. Any codebook that is at once usable and valid thus retains some irreducible ambiguity, and that ambiguity is distributed unevenly across a corpus: some texts are read the same way by everyone, while others are genuinely contestable.

Taken together, these features imply that text annotation is a demanding task even for humans, and that there is no ground truth to be recovered for some cases. What researchers can do is narrow the range of defensible readings through operationalization, and validate the measurement repeatedly (Krippendorff, 2018). This is precisely what disciplined content analysis aims to achieve. Yet even the most careful coding rule cannot turn human labels into perfect: on genuinely ambiguous texts, trained coders following identical instructions still fail to converge, and inter-coder agreement is rarely complete (Mikhaylov et al., 2012).

2.2 Text-as-Data before Large Language Models

The central motivation for automated text analysis is a basic problem of scale. Traditional content analysis requires trained coders to apply a codebook document by document, but many modern corpora, such as legislative speeches, news articles, party manifestos, social media posts, and open-ended survey responses, are far larger than any research team can feasibly code by hand. Automated content analysis therefore emerged as a way to extend systematic content analysis to large corpora, translating human-defined concepts into reproducible computational procedures.

Before LLMs, automated text analysis relied on a diverse toolkit of methods, including dictionary approaches, topic models, and supervised classifiers (see Grimmer and Stewart 2013 for a review). Yet these methods do not explicitly model the data-generating processes underlying either text production or human interpretation (Grimmer et al., 2021). For this reason, Grimmer and Stewart (2013) argue that automated text-analysis algorithms "use insightful, but wrong, models of political text to help researchers make inferences from their data" (p. 270) and "will never replace careful and close reading of texts" (p. 268). Automated methods were therefore trusted primarily to the extent that they could reproduce human judgments (Hase, 2022). Researchers constructed validation sets whose labels were supplied by humans and then assessed whether automated method could recover those labels. High agreement with human coders was interpreted as evidence of accuracy, whereas low agreement was attributed to errors or biases in the automated method.

Researchers also sometimes relied on crowdsourced workers to construct validation sets (e.g., Budak et al. 2016; Park 2021; Rheault et al. 2019; Theocharis et al. 2016; Ying et al. 2022). Crowd workers offered a substantially less expensive alternative to expert coders, although their individual annotations were generally noisier because they tended to have less training, expertise, and incentive to attend carefully to the coding task (Benoit et al., 2016; Snow et al., 2008). Nevertheless, even crowdsourced annotations were typically presumed to be more informative about the meaning of a text than the automated methods. This human-centered validation paradigm was reasonable in the pre-LLM era because human annotations were assumed to provide a closer approximation to the true meaning of a text than any existing algorithm could.

2.3 Text-as-Data after Large Language Models

LLMs can now perform tasks that, until recently, were thought to require sophisticated human reasoning: they pass professional and graduate examinations, solve competition-level mathematics and coding problems, and score competitively on broad multitask language-understanding benchmarks (Hendrycks et al., 2021; Rathje et al., 2024). Against this backdrop, it is worth asking a simple question: if these systems can handle such hard tasks, are we still confident that they cannot perform text annotation as human experts do?

This question is crucial because it undermines the earlier validation paradigm. When an LLM disagrees with a human coder, the disagreement need not imply that the LLM is wrong or biased. It may instead reflect ambiguity in the text or codebook. Importantly, these explanations apply not only to LLMs but also to humans. Different human experts may code the same text differently when the text or codebook is ambiguous. Indeed, several studies have already questioned the use of human evaluations as a gold standard (e.g., Bisbee and Spirling 2026; Clark et al. 2021; Hosking et al. 2024). Thus, validating LLM annotations against a single human-generated label may not be the best validation practice.

Yet, methodological practice has not fully adjusted to this change in the relative capabilities of humans and machines, especially in recent work on downstream inference with machine-generated annotations. Social scientists are often interested not in annotations themselves but in quantities estimated using annotated text, such as regression coefficients. To correct bias arising from machine-generated annotations, recent methods combine machine annotations observed for the full corpus with human annotations observed for a random validation subset (e.g., Angelopoulos et al. 2023; Egami et al. 2023, 2024; Carlson and Dell 2025; Fong and Tyler 2021; Nakamura 2025; Zrnic and Candès 2024). Although much of this literature emerged after the rise of LLMs, it largely inherits the validation logic of the pre-LLM era: the human-coded subset

is typically treated as the ground truth (or at least as more accurate than the machine-generated annotations) and is therefore used to anchor corrections for measurement error in LLM annotations.

Because of the importance of the question, recent years have witnessed a large body of validation work on LLM annotation (e.g., Gielens et al. 2026; Gilardi et al. 2023; Goodall et al. 2026; Halterman and Keith 2026; Heseltine and Clemm von Hohenberg 2024; Liu and Sun 2025; Rathje et al. 2024; Törnberg 2025; Yang et al. 2025; Ziems et al. 2024; Zhu et al. 2023), but existing evaluations provide mixed evidence. Several studies find that LLMs outperform trained annotators on tasks such as relevance, topic, and frame classification (Heseltine and Clemm von Hohenberg, 2024; Törnberg, 2025). Other research documents substantial variation across tasks and models (Gong et al., 2026; Ziems et al., 2024). In the context of political science papers, Halterman and Keith (2026) show that LLMs can struggle to follow real-world political-science codebooks that operationalize complex, researcher-defined constructs. On the other hand, Gilardi et al. (2023) found that early LLMs agreed with ground truth generated by expert annotations more often than crowd-sourced workers.

These evaluations, however, are insufficient to answer whether LLMs can perform as well as human coders, since most studies assess model outputs against a single human-coded reference set, treating those labels as ground truth. [1] Such designs can show whether an LLM reproduces human labels, but they cannot distinguish among three possibilities: that the LLM is wrong, that the human label is wrong, or that the text is ambiguous and admits multiple interpretations. This issue is especially important in social science, where the concepts researchers measure are often subtle and complex. What is needed, therefore, is a validation design that compares LLM–human agreement with human–human agreement and examines whether both forms of disagreement are driven by the same underlying ambiguity.

2.4 Testable Hypotheses

Our argument yields four sets of predictions. First, if recent LLMs apply a common coding rule as reliably as trained experts, their agreement with experts should approach the agreement observed among experts. Because both groups receive the same texts and codebooks, this comparison isolates differences in how they apply the codebook.

Hypothesis 1 (Observational Equivalence). *Recent LLMs achieve agreement with expert coders comparable to the agreement observed among expert coders themselves.*

Second, if annotation disagreement reflects ambiguity in the text or codebook, the same observations should generate disagreement across different annotators. LLMs should disagree more with a held-out expert on texts that divide the other experts, and LLM–expert disagreement should increase with disagreement across LLMs.

Hypothesis 2-a (Expert disagreement). *LLM–expert disagreement is greater for texts on which expert coders disagree with one another.*

Hypothesis 2-b (LLM disagreement). *LLM–expert disagreement is greater for texts that generate more disagreement across LLMs.*

Third, if disagreement arises from ambiguities, clarifying the codebook should improve agreement across annotators.

Hypothesis 3 (Clearer codebook). *A clearer and more explicit codebook increases agreement among experts, between LLMs and experts, and across LLM annotations.*

[1] Among the few exceptions is Yang et al. (2025), which collects an additional set of annotations from the authors as a robustness check. Our design differs from theirs in two respects. First, their study focuses primarily on how the choice of LLM affects downstream estimates rather than on the sources of annotation disagreement. Second, their human coders first annotated a common set of 20 texts and reconciled their disagreements before coding the remaining texts. Although this calibration may improve human–human agreement, it also means that the human coders received information unavailable to LLMs.

Finally, the same relationship between textual ambiguity and disagreement should extend to crowd-sourced coders. However, because crowd coders generally have less substantive expertise, we expect their annotations to correspond less closely with expert annotations than LLM annotations do.

Hypothesis 4-a (Crowd disagreement). *Disagreement among crowd-sourced coders is greater for texts on which experts disagree or LLM outputs vary across models.*

Hypothesis 4-b (LLMs versus crowd coders). *LLMs achieve greater agreement with expert coders than crowd-sourced coders do.*

3 Replication Studies by LLMs and Experts

3.1 Replication Procedure

Data Collection: We focus on recent applications of automated text analysis published in leading political science journals. Specifically, we selected 13 articles with journal publication years from 2018 through 2025 in the *American Journal of Political Science*, the *American Political Science Review*, and the *Journal of Politics* that use text classification as part of their empirical analyses. To ensure broad substantive coverage, we selected studies spanning the major subfields of political science, including three in American politics (Feltovich and Giovannoni, 2024; Gilardi et al., 2021; Lacombe, 2019), four in international relations (Clark and Dolan, 2021; Mattingly et al., 2025; Schub, 2022; Thrall, 2025), and six in comparative politics (Blaydes et al., 2018; Blumenau and Lauderdale, 2018; Bush and Clayton, 2023; Jung, 2020; Parthasarathy et al., 2019; Osnabrügge et al., 2021). As a theoretically hard case, we further include Arias (2022). Recent work suggests that LLMs may struggle to classify concepts that are rare or absent from their training data (Halterman and Keith, 2026). Consequently, findings based on general-interest journals may not fully generalize to highly specialized subfields. The study by Arias (2022) provides a demanding substantive and robustness check because it focuses on *climate securitization*, a construct in international relations that is unlikely to appear frequently in LLM training corpora and remains difficult to define even within the international relations scholarship (Buzan et al., 1998; Phillis et al., 2018).

For each paper, we selected one concept to reanalyze in our replication study. In some papers, this choice was straightforward because the analysis focused on a single concept. In others, particularly those using topic models, the authors analyzed multiple concepts simultaneously. In these cases, we selected the concept that was most central to the paper's substantive findings. When multiple concepts were equally relevant, we randomly selected one for analysis. See Appendix S1.1 for the summary of each study and the selected concept we reanalyze.

After selecting the target concept, we drew a simple random sample of 100 text units from the original corpus for each study. All analyses reported below are based on these sampled texts, which were independently annotated by LLM and human coders.

Codebook / Prompt generation: For each paper, we construct a standardized codebook template to ensure consistency across papers. The template consists of seven components: (1) role assignment, (2) topic definition, (3) paper and data description, (4) category definition, (5) classification rules, (6) label wording, and (7) step-by-step reasoning instructions. We design this template based on the prompt engineering literature for large language models (e.g., Atreja et al. 2024; Halterman and Keith 2026; Mu et al. 2024; Weber and Reichardt 2024). Because components (2)–(6) are application-specific, we use ChatGPT 5.5 to generate them for each paper, ensuring that the quality and structure of the codebooks remain as consistent as possible across studies. We then manually review each generated codebook and make minimal revisions when necessary. Finally, to facilitate comparison across studies, we restrict all classification tasks to binary category. See Appendix S1.2 for the detailed procedure and the exact prompt. While our main analysis uses the prompt containing all components, we also check how much the results would change by systematically varying prompts (see Appendix S3.1.2).

Table 1: Language models used in the replication, with release dates.

Model	Developer	Release date
Mistral 7B (Instruct v0.3)	Mistral AI	May 2024
LLaMA 3.1-8B	Meta	July 2024
Qwen2.5-7B	Alibaba	September 2024
Qwen2.5-14B	Alibaba	September 2024
Gemma 3-12B	Google	March 2025
Gemma 3-27B	Google	March 2025
GPT-OSS-20B	OpenAI	August 2025
GPT-5.4 mini	OpenAI	March 2026
GPT-5.4	OpenAI	March 2026
GPT-5.5	OpenAI	April 2026

LLM Annotation: We then annotate each text using ten different LLMs that span a wide range of parameter counts, developers, release dates, and likely training data. Seven are open-weight models: Mistral 7B (Jiang et al., 2023), LLaMA 3.1-8B (Grattafiori et al., 2024), Qwen2.5-7B and Qwen2.5-14B (Qwen Team, 2024), Gemma 3-12B and Gemma 3-27B (Gemma Team et al., 2025), and GPT-OSS-20B (OpenAI et al., 2025), and three are proprietary API models from the GPT-5 family (GPT-5.4 mini, GPT-5.4, and GPT-5.5). In addition, we report a majority-vote ensemble that aggregates the predictions of these models. Comparing results across models allows us to assess how differences in scale, developer, and training data translate into performance across settings. Table 1 lists each model with its release date.

Expert Annotations: Finally, each author annotated the same set of texts using the identical codebooks provided to the LLMs. To ensure comparability, the authors completed the annotation independently, did not discuss difficult cases, and relied exclusively on the coding rule without consulting external references or additional information. By holding the available information constant across human and LLM annotators, our design isolates differences in the application of the coding rule rather than differences in background knowledge or information access. This enables a direct comparison of LLM–expert agreement with expert–expert agreement.

After three experts annotated each text, we calculated the mean pairwise agreement, defined as the average proportion of texts that received the same label across three possible pairs of experts. Table 2 reports this measure for each replicated study. In line with the existing literature (e.g., Gong et al. 2026), mean pairwise agreement varies substantially across papers.

3.2 Results

3.2.1 Observational Equivalence of LLM and Human Annotation

We begin by testing Hypothesis 1 by comparing average LLM–expert agreement with expert–expert agreement. Because expert agreement varies substantially across studies, we present the results separately for relatively easy tasks (mean pairwise agreement > 0.8) and relatively difficult tasks (mean pairwise agreement ≤ 0.8).

Figure 1 shows that LLM–expert agreement closely tracks expert–expert agreement.[2] Among the easy tasks, expert–expert agreement is 0.89, while all models except Mistral-7B achieve LLM–expert agreement between 0.89 and 0.91. Among the difficult tasks, LLM–expert agreement declines to between 0.69 and 0.76, but expert–expert agreement also falls to 0.67. Thus, lower LLM–expert agreement occurs precisely in settings where experts themselves agree less. Overall, the results support Hypothesis 1: with the exception

[2]As a robustness check, we present the results with Krippendorff's α in Appendix S3.1.1. The conclusions remain unchanged: LLM–expert intercoder reliability closely matches expert–expert intercoder reliability.

Table 2: The 14 replicated studies: concepts of interest and expert intercoder reliability (ICR), measured as the mean pairwise agreement among the three trained coders on each study's full sample of 100 texts. The studies are ordered in descending order of expert intercoder reliability.

Paper	Concept of interest	Intercoder reliability (ICR)
Feltovich and Giovannoni (2024)	Negative campaigning	0.987
Bush and Clayton (2023)	Taxes	0.933
Thrall (2025)	Commercial issues	0.900
Mattingly et al. (2025)	Western dysfunction	0.887
Clark and Dolan (2021)	Fiscal policy	0.880
Arias (2022)	Climate security	0.873
Blumenau and Lauderdale (2018)	Finance	0.860
Parthasarathy et al. (2019)	Fund allocation	0.853
Jung (2020)	Moral rhetoric	0.847
Blaydes et al. (2018)	Art of rulership	0.753
Lacombe (2019)	Gun control	0.720
Gilardi et al. (2021)	Smoking ban enforcement	0.700
Schub (2022)	Adversary domestic politics	0.640
Osnabrügge et al. (2021)	Emotive rhetoric	0.533

of the oldest model (Mistral-7B), recent LLMs achieve agreement with experts comparable to the agreement observed among experts themselves.

3.2.2 Disagreement Concentrates in Texts That Divide Experts

To test Hypothesis 2-a, we use a leave-one-expert-out procedure. For each text, we classify it as *clear* when two experts agree and as *ambiguous* when they disagree. We then evaluate each LLM against the remaining held-out expert, repeating the procedure for all three choices of held-out expert and averaging the results.

Figure 2 shows that LLM agreement declines sharply on texts that divide experts. Among the easy studies, average agreement with the held-out expert falls from approximately 0.91 on clear texts to 0.67 on ambiguous texts. Among the difficult studies, it falls from approximately 0.74 to 0.62. These results support Hypothesis 2-a. LLM–expert disagreement is concentrated in the same texts that generate disagreement among experts, implying that much of the disagreement reflects ambiguity in the text or codebook rather than model-specific error. Appendix S2 also provides illustrative examples of these contested texts.

We then test Hypothesis 2-b by examining whether disagreement across LLMs is greater for texts that divide experts. For each text i, let p_i denote the proportion of the 10 LLMs assigning the label 1. We measure cross-model disagreement using Gini impurity,

$$\text{Gini}_i = 2p_i(1 - p_i), \tag{1}$$

which ranges from zero when all models agree to 0.5 when they are evenly divided.

Table 3 shows that cross-model disagreement is much higher for texts on which experts disagree. In the easy studies, mean Gini impurity rises from 0.067 to 0.238; in the difficult studies, it rises from 0.141 to 0.254. Both differences are statistically significant. These results support Hypothesis 2-b and are robust to different measures of heterogeneity (see Appendix S3.1.4 for results based on Shannon's entropy). In Appendix S3.1.2, we obtain similar results by varying the prompts: for most models, classifications are significantly more sensitive to small changes in prompt for texts on which experts disagree.

3.2.3 Clarifying the Codebook Improves Agreement

Finally, we test Hypothesis 3 using the classification task from Schub (2022). Specifically, we developed a revised codebook with clearer definitions, inclusion and exclusion rules, and guidance for mixed cases. The

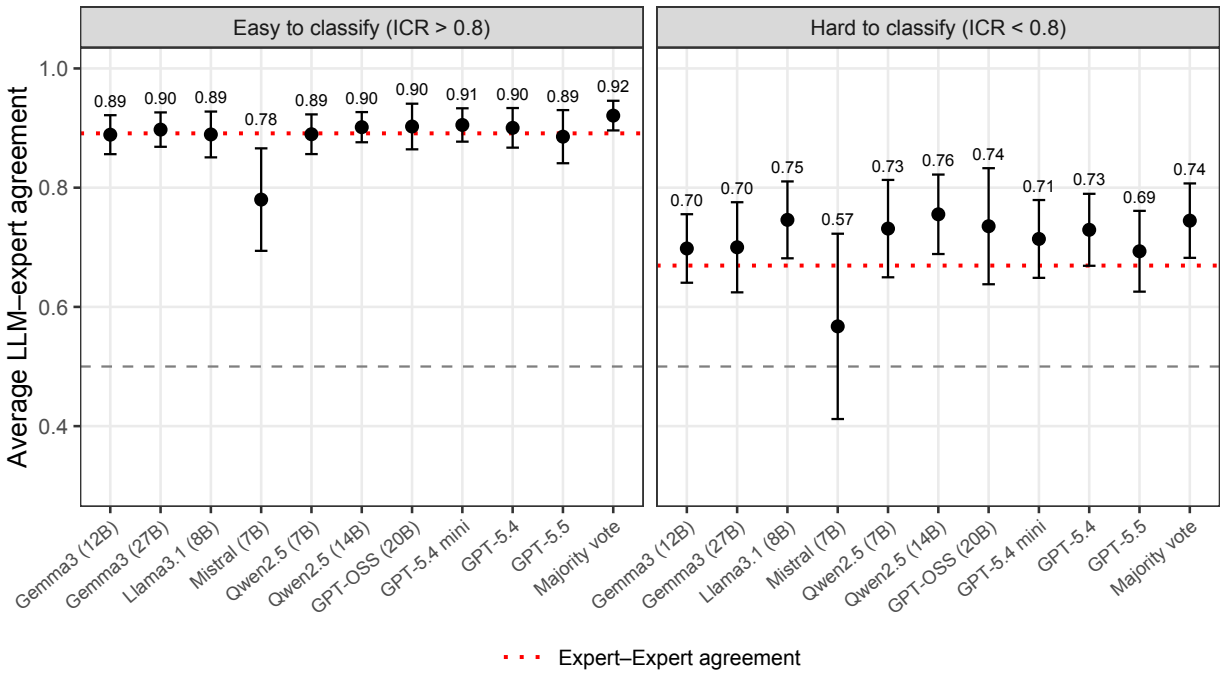

Figure 1: Average LLM–expert agreement by model, separately for relatively easy studies (expert ICR > 0.8; 9 studies) and relatively difficult studies (expert ICR ≤ 0.8; 5 studies). Points are model-specific means, and error bars are 95% between-study confidence intervals. The red dotted line indicates mean expert–expert agreement within each panel, and the gray dashed line indicates 0.50 agreement.

three experts and all 10 LLMs then reclassified the same texts using the revised instructions.

Figure 3 shows that the revised codebook raises expert–expert agreement from approximately 0.64 to 0.82. It also increases LLM–expert agreement for most recent models, several of which reach between 0.77 and 0.82 and approach the expert benchmark. The improvement is not uniform, however. Mistral-7B and Qwen2.5-7B perform worse under the more detailed instructions, suggesting that clearer but more complex codebook may be difficult for weaker models to apply consistently.

3.2.4 Same finding for Crowdsourced Workers, but they are less reliable than LLMs

Finally, we extend our analysis to crowdsourced annotation. We recruited 165 U.S. participants through Cloud Research Connect and randomly assigned each participant to one of the 14 papers we used in replication or to the revised codebook version of Schub (2022) analyzed in the previous section. For each paper, we limited the coding task to 20 texts to reduce respondent fatigue while retaining enough observations for analysis. Specifically, we selected the 10 texts with the highest agreement across LLMs and the 10 texts with the lowest agreement to sample both easy and difficult cases. Participants were asked to apply the same codebook used in the previous section to annotate the assigned texts. See Appendix S1.3 for the details ofr annotation procedure and Appendix S3.2 for the preregistered hypotheses and results.

We obtain two main findings. First, we find that crowd workers exhibit greater disagreement on ambiguous texts, regardless of whether ambiguity is measured by disagreement among LLMs or among experts. The difference in Gini impurity between ambiguous and non-ambiguous texts is both statistically and substantively significant: Gini impurity is 14 percentage points higher for the texts with the lowest versus highest LLM agreement and 10 percentage points higher for texts on which experts split versus those on which they reached unanimous agreement (Figure 4).

Second, we find that agreement between experts and LLMs is substantially higher than agreement be-

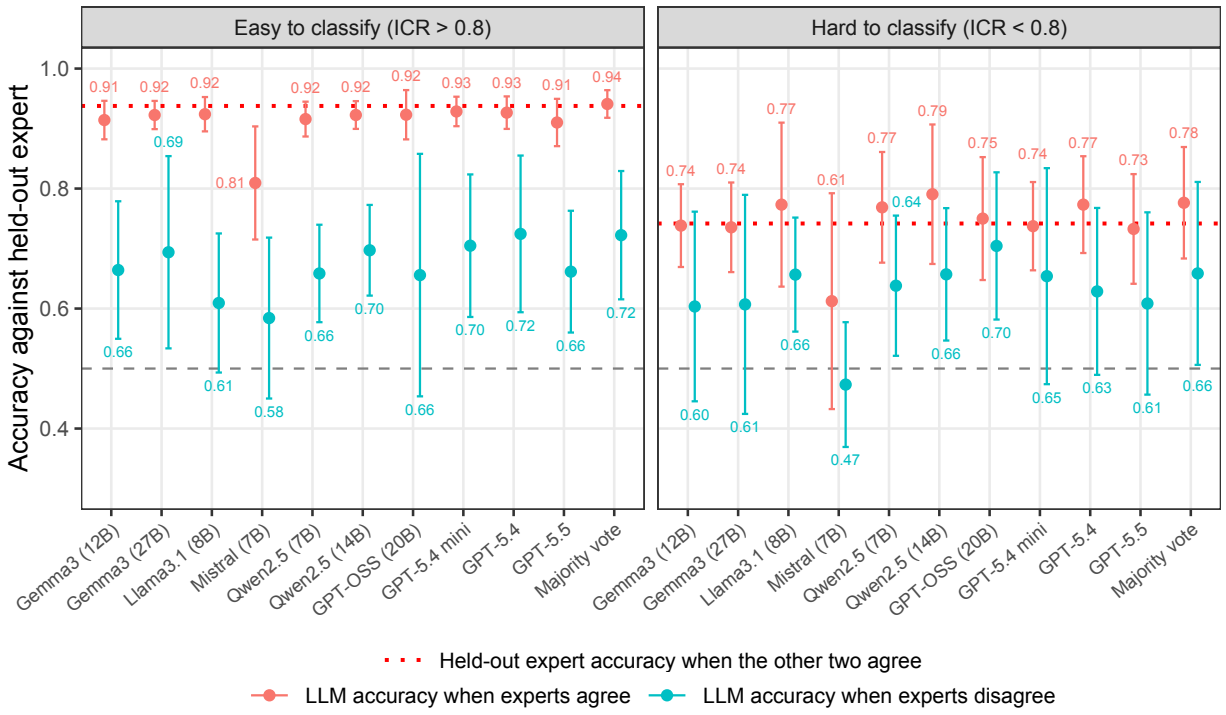

Figure 2: Agreement with a held-out expert, by model and by whether the two remaining experts agree. The red dotted line indicates the held-out expert's agreement with the common label of the other two experts on clear texts. Coral and teal points report LLM agreement with the held-out expert on clear and ambiguous texts, respectively. Error bars are 95% between-study confidence intervals.

tween experts and crowd workers. To facilitate this comparison, we replot our main analysis from Figure 1 using only the same 20 texts annotated by all three coders. This standardized comparison, based on identical texts and codebooks, replicates and extends the findings of Gilardi et al. (2023) by comparing inter-coder agreement rather than evaluating performance against an assumed ground truth. We find not only a substantial gap between expert–LLM and expert–crowd agreement, but also that the attainable level of agreement varies systematically with task difficulty. As shown in Figure 5, mean expert–expert agreement (red dotted line) declines considerably for difficult tasks (right panel) relative to easy tasks (left panel). Most strikingly, agreement among LLMs closely tracks agreement among experts across both types of tasks, whereas agreement among crowd workers remains consistently lower.

4 Practical Takeaways for Applied Researchers

Given these results, how should researchers analyze text data with LLMs? We propose a procedure that uses multiple LLMs to efficiently identify borderline cases, refine the codebook, and ultimately conduct honest statistical inference without imposing strong assumptions. Figure 6 summarizes the procedure. Below, we explain each step in detail.

4.1 STEP 1: Generate a clear Codebook

The first step is the construction of the codebook. Researchers should first define what they want to measure, and find the way to operationalize this. This step is consequential, as our results show that much of what appears to be annotation error, whether by LLMs or by humans, is in fact ambiguity introduced at the operationalization.

Table 3: Cross-model disagreement by expert agreement and study difficulty. Higher values indicate greater disagreement among the 10 LLMs.

	Mean LLM disagreement				
Study difficulty	Experts unanimous	Experts split	Difference	95% CI	N
Easy (ICR > 0.8)	0.067	0.238	0.171	[0.125, 0.217]	753 / 147
Hard (ICR ≤ 0.8)	0.141	0.254	0.113	[0.051, 0.175]	252 / 248

Notes: Cells are mean Gini impurity $2p(1-p)$ over the ten LLM votes per text. N counts texts (unanimous / split); confidence intervals are cluster-robust by paper.

While the best codebook depends on the situation and concept of interest, we generally see that the clear instruction of classification rules help improve the LLM's agreement with expert annotations (see Appendix S3.1.2). Researchers are expected to describe the clear classification rules and revise them through trials and errors in STEP2.

4.2 STEP 2: Test Codebook with LLMs / Experts and inspect Disagreements if necessary

Before annotating the full corpus, researchers should begin by manually coding a small simple random sample of texts. This initial random allows researchers to assess how frequently the proposed codebook yields clear classifications in the corpus and exposes common weaknesses in the codebook before annotation is scaled up. Because researchers generally have the clearest substantive understanding of the concept they intend to measure, they should inspect disagreements and difficult cases in this initial sample and revise the codebook accordingly. This step is necessary regardless of who ultimately performs the full annotation, because it concerns the quality of the codebook rather than the choice of annotator.

After this first round of human coding and codebook revision, researchers should apply the revised codebook to multiple LLMs across a larger set of subsets for additional pilots. They can then use disagreement across LLMs to identify additional cases that warrant manual review. [3] This step effectively expands the initial random human-coded sample: rather than selecting additional texts randomly, researchers deliberately add texts for which the LLMs disagree, because these cases are more likely to reveal residual ambiguity in the codebook. Researchers can inspect these targeted cases, further revise the codebook, and repeat the process until the remaining disagreements can no longer be resolved through clearer instructions.

Implementing this procedure requires selecting both LLMs and prompts. Because model capabilities evolve rapidly, we do not recommend a fixed set of models or prompts. Instead, researchers should first use a pilot to evaluate which models can reliably apply the codebook, ideally using multiple expert coders even for a small subset of texts. Models whose agreement with expert coders falls substantially below expert–expert agreement may generate disagreement because of limited model capability or poor instruction following rather than underlying ambiguity and should therefore be excluded from the analysis. Our empirical results suggest that this requirement is not especially restrictive: annotation quality varies relatively little across recent medium- and large-scale models, except the old and small model, and open-source models perform comparably to proprietary state-of-the-art models, consistent with existing evidence (e.g., Yang et al. 2025). Prompt choice appears less consequential once the codebook is clearly specified. In particular, LLM-specific prompting techniques, such as role assignment or chain-of-thought reasoning, contribute relatively little to agreement compared with clear classification rules (see Appendix S3.1.2).

We emphasize that STEP 1 and STEP 2 are the most important components and that statistical corrections

[3] Appendix S3.1.3 examines whether repeated runs of a single model can provide an alternative measure of annotation uncertainty. Under stochastic decoding, within-model instability is concentrated on texts that divide experts for three of the seven models we tried (GPT-OSS-20B, LLaMA 3.1-8B, and Qwen2.5-14B), but not for all models, possibly reflecting the difference in model training. Thus, repeated runs of one model can be informative, although their usefulness depends on the model.

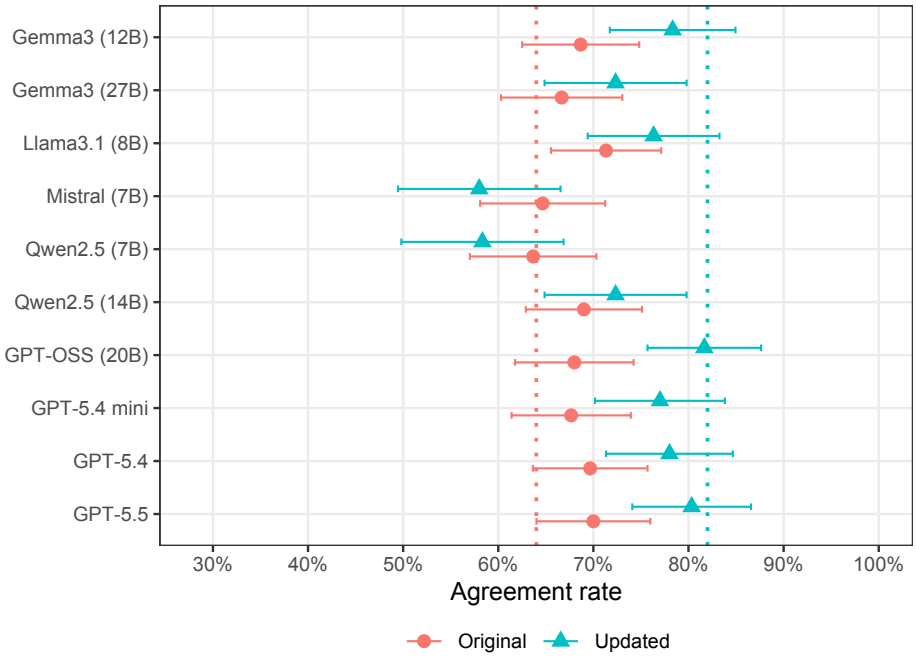

Figure 3: LLM–expert agreement under the original (red) and revised codebook (blue) for the Schub (2022) annotation task. Each point reports a model's average agreement with the three experts. The colored dotted lines indicate mean expert–expert agreement.

cannot substitute for them. Recently, researchers have proposed numerous methods for correcting measurement error in text annotations. Some of these methods rely on gold-standard labels (e.g., Angelopoulos et al. 2023; Carlson and Dell 2025; Egami et al. 2023, 2024; Fong and Tyler 2021; Nakamura 2025; Zrnic and Candès 2024), whereas others use latent-variable models (e.g., Balasubramanian et al. 2026; Dawid and Skene 1979; Egami and Shin 2026; Luo et al. 2020; Raykar et al. 2010; Zhao et al. 2026). Although these methods might be attractive, they cannot resolve ambiguity that remains under the codebook itself. When a codebook admits multiple reasonable interpretations, there may be no reader-independent truth to recover. Statistical methods may still produce an estimate, but if researchers cannot theoretically define what the true annotation should be under the codebook, that estimate is merely a statistical artifact rather than a substantively interpretable quantity.

4.3 STEP 3: Run Downstream Inference under Residual Ambiguity

Once the codebook has stabilized, researchers can annotate the full corpus using multiple LLMs that perform reasonably well in STEP 2. Although the codebook should be revised extensively in STEPS 1 and 2, some ambiguous texts may inevitably remain. Indeed, we find that some texts drawn from corpora used in published political science research are inherently ambiguous, such that the appropriate annotation is not clear even under a well-developed codebook. One solution for this might be to impose an explicit decision rule, for example, coding all unclear cases as zero, to increase intercoder agreement. Such a rule, however, may compromise the validity of the resulting measure by forcing ambiguous texts into categories that do not accurately reflect the underlying concept of interest. Because researchers typically use these annotations in downstream analyses, such as running regressions, the remaining ambiguity may distort the resulting estimates and statistical inference. How, then, should researchers proceed?

We suggest researchers to use *ambiguity-aware bounds*, which accounts for the fact that the truth can be totally undefined if texts are ambiguous given codebook. For the sake of explanation, consider the average

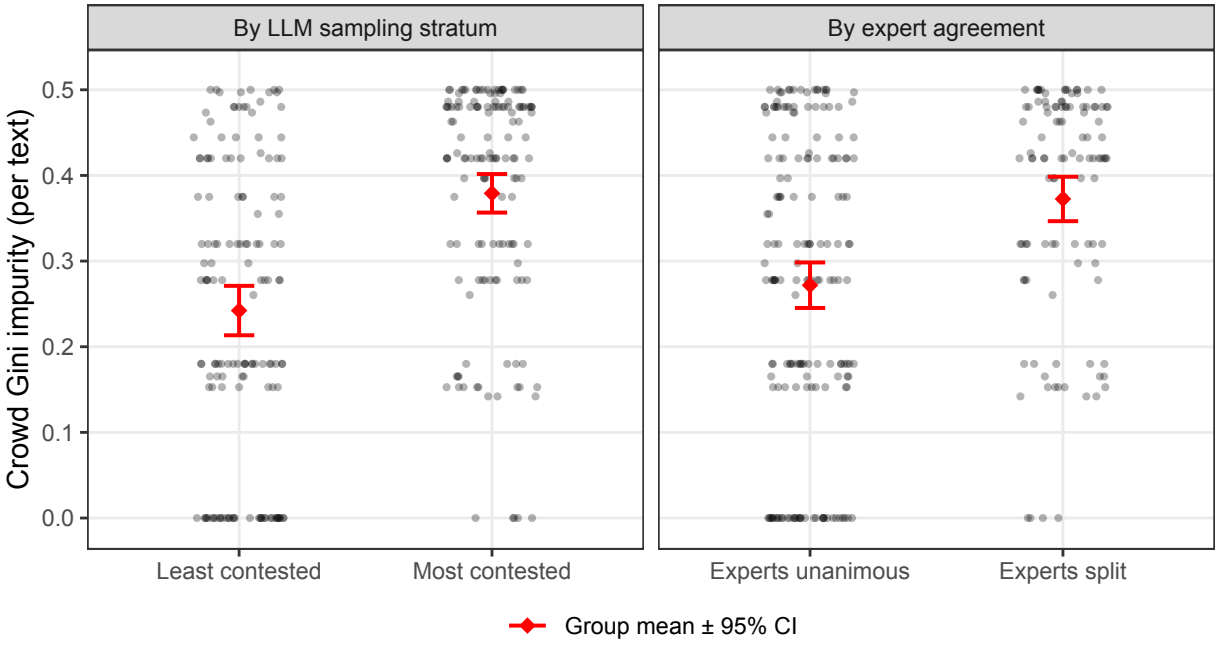

Figure 4: (Left) For each paper, texts are grouped as the 10 least contested texts across LLM coding versus 10 most contested ones, based on our stratified sampling. (Right) Texts are grouped into those which the three expert coders are unanimous versus split. In red are the group means with 95% confidence intervals; gray points are individual texts. We exclude the revised codebook version of Schub (2022) for consistency.

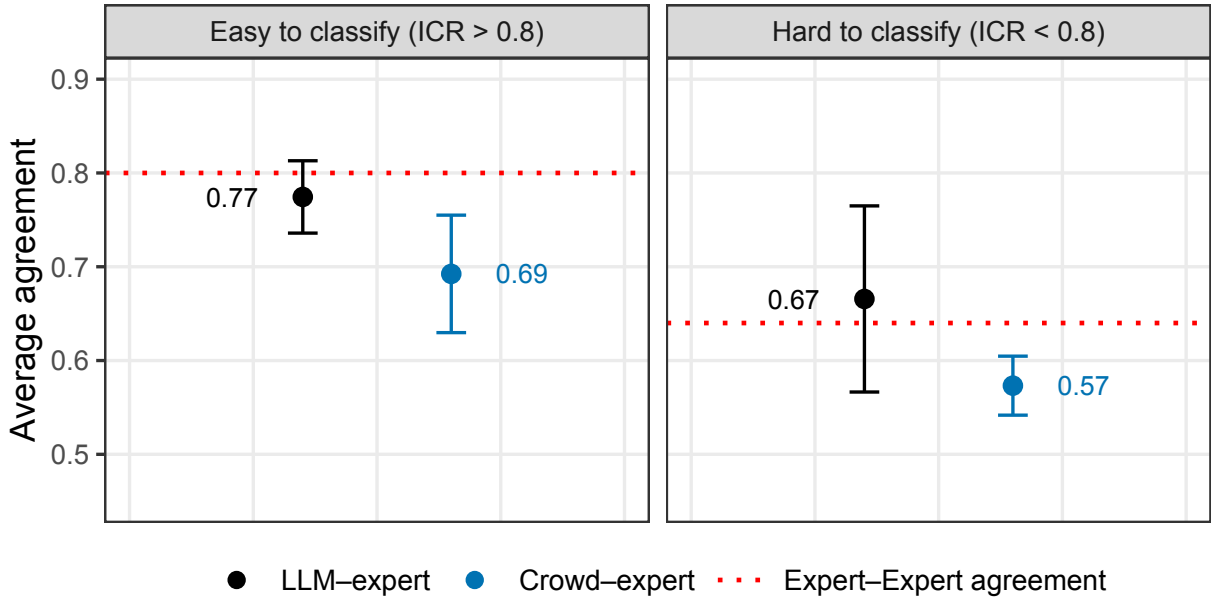

Figure 5: Survey extension of Figure 1, restricted to the 20-text-sample. The red dotted line is mean pairwise expert-expert agreement, black dot and bars are LLM-expert agreement, and blue the crowd-expert agreement. Each unit is the per-paper mean agreement over 20 texts by coder type, averaged across all coders.

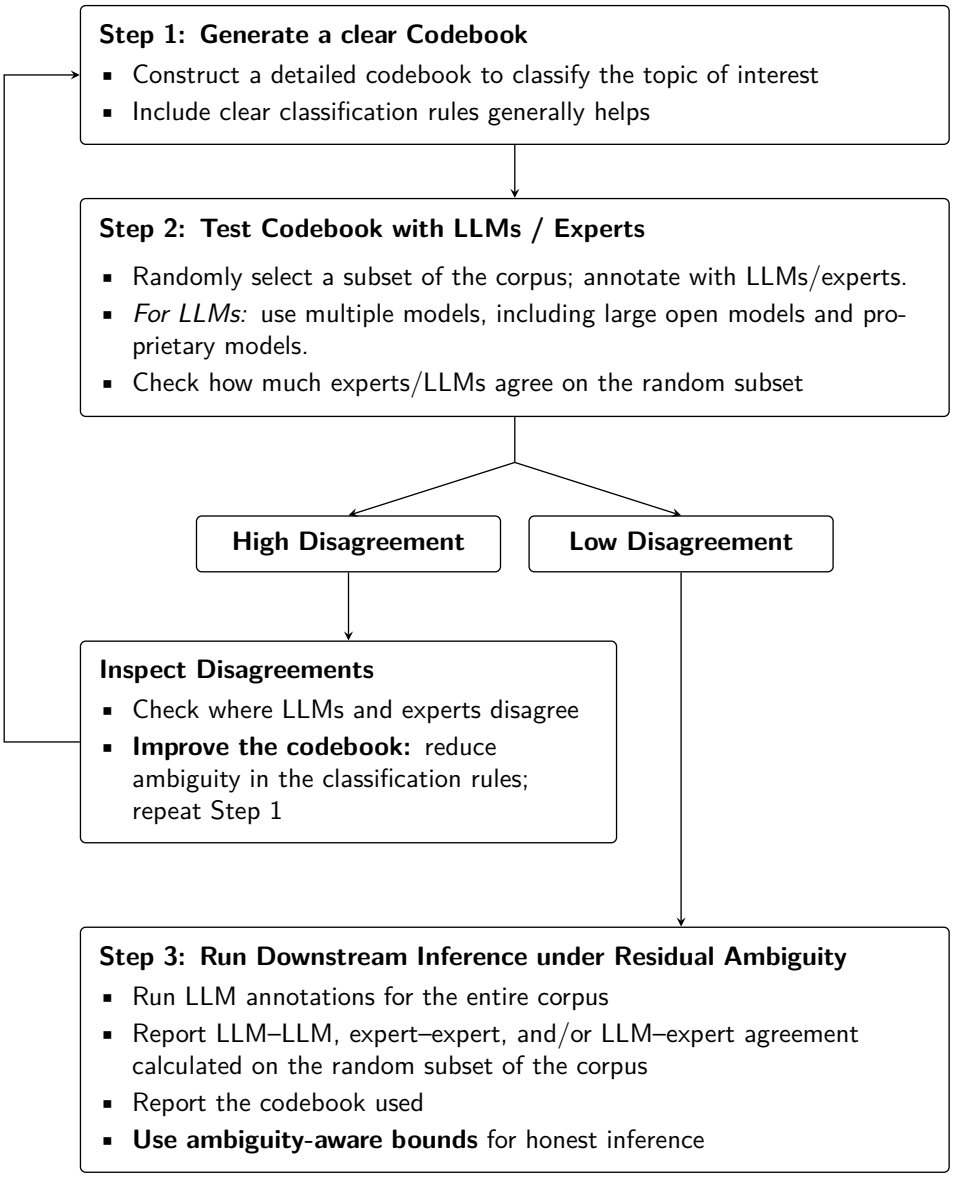

Figure 6: Proposed analysis procedure for text-as-data using LLMs.

of binary annotations across N texts, which is formally written as

$$\theta := \frac{1}{N}\sum_{i=1}^{N} Y_i, \tag{2}$$

where Y_i is the true annotation of text i ($i = 1, \ldots, N$). The problem here is that, if we cannot define the truth for some texts, the quantity above is also undefined. We can, however, still form the bounds without assuming the existence of the unique truth if we know which texts are ambiguous. To explain this, let A_i be an binary indicator that takes 1 if the annotation of text i is ambiguous and 0 otherwise. Because Y_i is binary,

$$\theta \in \left[\frac{1}{N}\sum_{i=1}^{N}(1 - A_i)\hat{Y}_i,\ \ \frac{1}{N}\sum_{i=1}^{N}\Big((1 - A_i)\hat{Y}_i + A_i\Big)\right], \tag{3}$$

where $\hat{Y}_i$ is the binary observed annotations that can be made by humans or LLMs. We assume that the codebook is sufficiently clear so that $\hat{Y}_i = Y_i$ for unambiguous texts. [4] The lower bound corresponds to the case in which all ambiguous texts ($A_i = 1$) have $Y_i = 0$, whereas the upper bound corresponds to the case in which they all have $Y_i = 1$. This approach enables researchers to conduct honest inference while acknowledging that a uniquely defined ground-truth label may not exist for some texts.

We recommend that researchers assess ambiguity because it is defined relative to the researchers' intended operationalization of the concept, and the human review therefore provides an independent substantive assessment of whether the codebook leaves more than one defensible classification. Importantly, errors in coding ambiguity have asymmetric consequences for our bounds. Classifying an unambiguous text as ambiguous is conservative because it widens the bounds, whereas failing to classify an ambiguous text as ambiguous may produce bounds that are too narrow. We therefore recommend coding A_i conservatively, treating a text as ambiguous whenever the researcher cannot rule out a substantively defensible alternative classification.

To scale the bounds in Equation (3), we propose using an ambiguity measure based on disagreement among multiple LLMs, such as entropy or Gini impurity. Although such a measure does not perfectly capture ambiguity, our results suggest that it provides a useful proxy. The procedure is as follows. First, after finalizing the codebook, researchers examine a random subset of texts and determine whether each text is ambiguous. Let S_i be an indicator equal to one if text i is included in the subset reviewed by the researchers and zero otherwise. Because researchers directly assess the texts in this subset, the true ambiguity status A_i is observed whenever $S_i = 1$. Second, researchers annotate the full corpus using multiple LLMs, calculate their level of disagreement, and construct a proxy for ambiguity, denoted by $\hat{A}_i$. Finally, researchers calculate the following bounds:

$$\theta \in \left[\frac{1}{N}\sum_{i=1}^{N}(1 - \tilde{A}_i)\hat{Y}_i,\ \ \frac{1}{N}\sum_{i=1}^{N}\Big((1 - \tilde{A}_i)\hat{Y}_i + \tilde{A}_i\Big)\right] \ \text{where}\ \ \tilde{A}_i = \hat{A}_i + \frac{S_i}{\Pr(S_i = 1)}\Big(A_i - \hat{A}_i\Big). \tag{4}$$

Here, $\tilde{A}_i$ is the design-adjusted ambiguity measures (Angelopoulos et al., 2023; Egami et al., 2024). These bounds can be calculated without observing A_i for the entire corpus and provide unbiased estimates of the bounds in Equation (3).[5] This approach can readily be extended beyond population averages. See Ap-

[4] In some applications, this assumption may be violated, for example, because LLMs or human annotators do not follow the codebook for certain types of texts. We relax this assumption in Appendix Section S4.

[5] To see why, first notice that

$$\mathbb{E}[\tilde{A}_i \mid A_i, \hat{A}_i, \hat{Y}_i] = \hat{A}_i + \frac{\Pr(S_i = 1)}{\Pr(S_i = 1)}\Big(A_i - \hat{A}_i\Big) = A_i,$$

pendix S4 for its generalizations and the empirical application. Our bounds are more informative than sensitivity analyses or bounds designed for non-random measurement error (e.g., Imai and Yamamoto 2010; Bisbee and Spirling 2026) by incorporating text-level variation in ambiguity and an empirically informative proxy for it.

The bounds in Equation (4) further underscore the importance of STEPS 1 and 2. Although these bounds allow researchers to report results without assuming the existence of a unique ground truth, they become uninformative when many texts are ambiguous and their annotations remain undetermined. They should therefore be viewed as a complement to codebook development rather than as a substitute for it.

5 Conclusion

In this paper, we evaluate the quality of LLM-based annotation in quantitative text analysis. Using replications of 14 political science text-classification tasks, we compare annotations from ten LLMs, three trained experts, and 165 crowd-sourced workers applying identical codebooks. We find that recent LLMs agree with experts at rates comparable to the agreement observed among experts themselves, establishing an observational equivalence between LLM and human coding. We further show that this equivalence arises because different annotators struggle with the same texts: LLM–expert disagreement is concentrated in texts that divide experts, and disagreement across LLMs is substantially greater on those same cases. Clarifying an underspecified codebook reduces disagreement among experts and sufficiently capable LLMs. Crowd-sourced workers exhibit the same ambiguity-related pattern, although their overall agreement with experts is substantially lower than that of recent LLMs.

These findings change both how annotation disagreement should be interpreted and how text-based measures should be used in downstream analysis. Divergence from a human label should not automatically be treated as model error, because humans and LLMs tend to disagree on the same borderline cases. Researchers should therefore invest first in conceptualization and codebook development, using disagreement across multiple LLMs to identify difficult cases and refine coding rules. Because some ambiguity may remain even after this process, we develop ambiguity-aware bounds that do not require every text to possess a unique ground-truth label. The broader lesson is that reliable text analysis requires not only capable annotators, but also careful operationalization and statistical inference that explicitly propagates residual ambiguity rather than treating it as ordinary annotation error.

---

where the first equality is because the subset researchers read is chosen randomly. As a result,

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(1-\tilde{A}_i)\hat{Y}_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[(1-\tilde{A}_i)\hat{Y}_i] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(1-\mathbb{E}[\tilde{A}_i \mid A_i, \hat{A}_i, \hat{Y}_i]\right)\hat{Y}_i\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(1-A_i)\hat{Y}_i\right]$$

where the second equality follows from the law of iterated expectations. The same argument applies to the upper bound. Importantly, the argument does not assume that the proxy $\hat{A}_i$ is binary.

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.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023). Prediction-Powered Inference. arXiv:2301.09633 [stat].

Arias, S. B. (2022). Who securitizes? climate change discourse in the united nations. *International Studies Quarterly*, 66(2):sqac020.

Atreja, S., Ashkinaze, J., Li, L., Mendelsohn, J., and Hemphill, L. (2024). Prompt Design Matters for Computational Social Science Tasks but in Unpredictable Ways. arXiv:2406.11980 [cs].

Balasubramanian, K., Podkopaev, A., and Kasiviswanathan, S. P. (2026). Dependence-aware label aggregation for llm-as-a-judge via ising models. *arXiv preprint arXiv:2601.22336*.

Benoit, K., Conway, D., Lauderdale, B. E., 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.

Bisbee, J. and Spirling, A. (2026). What to do when humans are no longer the gold standard: Large language models, state of the art and robustness for politics research. Working paper, Vanderbilt University and Princeton University. First draft: June 13, 2025. This version: February 10, 2026.

Blaydes, L., Grimmer, J., and McQueen, A. (2018). Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds. *The Journal of Politics*, 80(4):1150–1167.

Blumenau, J. and Lauderdale, B. E. (2018). Never let a good crisis go to waste: Agenda setting and legislative voting in response to the EU crisis. *The Journal of Politics*, 80(2):462–478.

Budak, C., Goel, S., and Rao, J. M. (2016). Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271.

Bush, S. S. and Clayton, A. (2023). Facing Change: Gender and Climate Change Attitudes Worldwide. *American Political Science Review*, 117(2):591–608.

Buzan, B., Wæver, O., and de Wilde, J. (1998). *Security: A New Framework for Analysis*. Lynne Rienner Publishers, Boulder, CO.

Carlson, J. and Dell, M. (2025). A unifying framework for robust and efficient inference with unstructured data. *arXiv preprint arXiv:2505.00282*.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. (2021). All that's 'human'is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296.

Clark, R. and Dolan, L. R. (2021). Pleasing the principal: U.S. influence in world bank policymaking. *American Journal of Political Science*, 65(1):36–51.

Dawid, A. P. and Skene, A. M. (1979). Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics*, 28(1):20.

Egami, N., Hinck, M., Stewart, B., and Wei, H. (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, 36:68589–68601.

Egami, N., Hinck, M., Stewart, B. M., and Wei, H. (2024). Using large language model annotations for the social sciences: A general framework of using predicted variables in downstream analyses. *Preprint from November*, 17:2024.

Egami, N. and Shin, S. (2026). Debiased inference for ai-generated data without gold-standard labels: Identification via multiple imperfect measurements. *arXiv preprint arXiv:2608.18294*.

Feltovich, N. and Giovannoni, F. (2024). Campaign Messages, Polling, and Elections: Theory and Experimental Evidence. *American Journal of Political Science*, 68(2):408–426. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12722.

Fong, C. and Tyler, M. (2021). Machine Learning Predictions as Regression Covariates. *Political Analysis*, 29(4):467–484.

Gemma Team, Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gielens, E., Sowula, J., and Leifeld, P. (2026). Goodbye human annotators? content analysis of social policy debates using chatgpt. *Journal of Social Policy*, 55(2):385–404.

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.

Gilardi, F., Shipan, C. R., and Wüest, B. (2021). Policy Diffusion: The Issue-Definition Stage. *American Journal of Political Science*, 65(1):21–35. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12521.

Gong, L., Hopkins, D. J., and Wolken, S. (2026). The limits of de-politicizing–and also of annotation: A case study in russian media outlets' social media posts, 2016–2024. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pages 889–909.

Goodall, L. S., Shilton, D., Mullins, D. A., and Whitehouse, H. (2026). Large language models struggle with ethnographic text annotation. *arXiv preprint arXiv:2601.12099*.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24:395–419.

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Grimmer, J. and Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297.

Halterman, A. and Keith, K. A. (2026). Codebook llms: Evaluating llms as measurement tools for political science concepts. *Political Analysis*, 34(2):188–204.

Hase, V. (2022). Automated content analysis. In *Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft–Standardized content analysis in communication research: Ein Handbuch-a handbook*, pages 23–36. Springer.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

Heseltine, M. and Clemm von Hohenberg, B. (2024). Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1):20531680241236239.

Hosking, T., Blunsom, P., and Bartolo, M. (2024). Human feedback is not gold standard. In *International Conference on Learning Representations*, volume 2024, pages 55864–55883.

Imai, K. and Yamamoto, T. (2010). Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis. *American Journal of Political Science*, 54(2):543–560. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5907.2010.00446.x.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., and El Sayed, W. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jung, J.-H. (2020). The Mobilizing Effect of Parties' Moral Rhetoric. *American Journal of Political Science*, 64(2):341–355. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12476.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., and Winter, N. J. G. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, 8(4):614–629.

Kracauer, S. (1952). The challenge of qualitative content analysis. *Public opinion quarterly*, pages 631–642.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Lacombe, M. J. (2019). The political weaponization of gun owners: The national rifle association's cultivation, dissemination, and use of a group social identity. *The Journal of Politics*, 81(4):1342–1356.

Levy, M. A. (1995). Is the environment a national security issue? *International Security*, 20(2):35–62.

Liu, A. and Sun, M. (2025). From voices to validity: Leveraging large language models (llms) for textual analysis of policy stakeholder interviews. *AERA Open*, 11:23328584251374595.

Luo, Y., Card, D., and Jurafsky, D. (2020). Detecting Stance in Media On Global Warming. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.

Mattingly, D., Incerti, T., Ju, C., Moreshead, C., Tanaka, S., and Yamagishi, H. (2025). Chinese state media persuades a global audience that the "China model" is superior: Evidence from a 19-country experiment. *American Journal of Political Science*, 69(3):1029–1046. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12887.

Mikhaylov, S., Laver, M., and Benoit, K. R. (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political analysis*, 20(1):78–91.

Mu, Y., Wu, B. P., Thorne, W., Robinson, A., Aletras, N., Scarton, C., Bontcheva, K., and Song, X. (2024). Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. *arXiv preprint arXiv:2305.14310*.

Nakamura, K. (2025). Surrogate representation inference for text and image annotations. *arXiv preprint arXiv:2509.12416*.

OpenAI, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., et al. (2025). gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Osnabrügge, M., Hobolt, S. B., and Rodon, T. (2021). Playing to the Gallery: Emotive Rhetoric in Parliaments. *American Political Science Review*, 115(3):885–899.

Park, J. Y. (2021). When do politicians grandstand? measuring message politics in committee hearings. *The Journal of Politics*, 83(1):214–228.

Parthasarathy, R., Rao, V., and Palaniswamy, N. (2019). Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India's Village Assemblies. *American Political Science Review*, 113(3):623–640.

Peer, E., Vosgerau, J., and Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, 46(4):1023–1031.

Phillis, Y. A., Chairetis, N., Grigoroudis, E., et al. (2018). Climate security assessment of countries. *Climatic Change*, 148:25–43.

Qwen Team (2024). Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., and Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34):e2308950121.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning From Crowds. *Journal of Machine Learning Research*, 11(43):1297–1322.

Rheault, L., Rayment, E., and Musulan, A. (2019). Politicians in the line of fire: Incivility and the treatment of women on social media. *Research & Politics*, 6(1):1–7.

Rønnfeldt, C. F. (1997). Review essay: Three generations of environment and security research. *Journal of Peace Research*, 34(4):473–482.

Schub, R. (2022). Informing the leader: Bureaucracies and international crises. *American Political Science Review*, 116(4):1460–1476.

Snow, R., O'connor, B., Jurafsky, D., and Ng, A. Y. (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*, pages 254–263.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., and Parnet, O. (2016). A bad workman blames his tweets: The consequences of citizens' uncivil twitter use when interacting with party candidates. *Journal of Communication*, 66(6):1007–1031.

Thrall, C. (2025). Informational lobbying and commercial diplomacy. *American Journal of Political Science*, 69(3):1147–1162. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12873.

Törnberg, P. (2025). Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 43(6):1181–1195.

Vogler, A. (2023). Tracking climate securitization: Framings of climate security by civil and defense ministries. *International Studies Review*, 25(2):viad010.

Weber, M. and Reichardt, M. (2024). Evaluation is all you need: Prompting generative large language models for annotation tasks in the social sciences: A primer using open models. *arXiv preprint arXiv:2401.00284*.

Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, 122(47):e2518075122.

Yang, E., Wang, Z., Zhou, C., and Xu, Y. (2025). Data annotation with large language models: Lessons from a large empirical evaluation.

Ying, L., Montgomery, J. M., and Stewart, B. M. (2022). Topics, concepts, and measurement: A crowdsourced procedure for validating topics as measures. *Political Analysis*, 30(4):570–589.

Zhao, J., Shin, C., Huang, T.-H., GNVV, S. S. S. N., and Sala, F. (2026). Care: Confounder-aware aggregation for reliable llm evaluation. *arXiv preprint arXiv:2603.00039*.

Zhu, Y., Zhang, P., Haq, E.-U., Hui, P., and Tyson, G. (2023). Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

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.

Zrnic, T. and Candès, E. J. (2024). Cross-Prediction-Powered Inference. arXiv:2309.16598 [stat].

Appendix

Appendix Contents

S1 Details of Replication Studies

S1.1 Summary of Each Study

For each study, we summarize the paper, the text corpus used, the topics into which those texts are classified, how the topic classification is used in the original analysis, and the topic of interest in our reanalysis.

Arias (2022), ***Who Securitizes? Climate Change Discourse in the United Nations*****.** Arias asks why some states frame climate change as a security issue in the United Nations, and which states are most likely to make such securitizing moves. The paper argues that securitization is shaped by agenda-control incentives: P5 states benefit from moving issues toward the Security Council, while highly vulnerable states such as SIDS may resist losing agenda control. Empirically, the paper finds that UN climate discourse overall is not securitized, P5 states are more likely to securitize climate change, and SIDS are less likely to do so. The text data are speeches by state representatives in the UN General Assembly General Debate, filtered to speech segments discussing climate change. The analysis uses 4,525 climate-related speech segments from 1,987 speeches, with the earliest climate-related speech segment appearing in 1984.

Blaydes et al. (2018), ***Mirrors for princes and sultans: advice on the art of governance in the medieval Christian and Islamic worlds*****.** The paper compares medieval Christian and Islamic political advice texts ("mirrors for princes") to examine how ideas of governance evolved across regions. Using text-as-data methods, it identifies shared thematic structures and tracks their evolution over time. It finds broad similarities across traditions but key divergences, including declining religious emphasis in Europe and shifting ruler-centered discourse in the Islamic world. The data consist of translated medieval political advice texts addressed to rulers in Christian Europe and the Islamic world. These texts include guidance on governance, morality, religion, and social order, often written by elites for kings, courts, or political audiences.

Blumenau & Lauderdale (2018), ***Never let a good crisis go to waste: Agenda setting and legislative voting in response to the EU crisis*****.** Blumenau and Lauderdale study how the 2008 global financial and sovereign debt crises affected agenda setting and legislative voting in the European Parliament. They argue that crises weaken legislators' attachment to the status quo, allowing pro-integration agenda setters to pass more integrationist policy than would otherwise have been possible. Empirically, they show that crisis-related votes increasingly divided MEPs along the pro- versus anti-European integration dimension rather than the left-right dimension. The text data are European Parliament legislative summary texts linked to roll-call votes from the sixth and seventh European Parliaments, covering 2004-2014. These summaries describe the purpose, background, and content of legislation under discussion, and are used to identify whether votes concern crisis-relevant financial and economic policy areas.

Bush & Clayton (2023), ***Facing Change: Gender and Climate Change Attitudes Worldwide*****.** The article argues that the gender gap in climate concern widens with national wealth because men in richer countries perceive greater costs from climate mitigation. The text analysis draws on open-ended survey responses, collected in an original ten-country survey, in which respondents describe the personal consequences they associate with acting on climate change. A structural topic model recovers recurring topics – among them denying any personal cost, anticipating higher taxes or prices, and pointing to environmental benefits. The topics are used at the response level, with the prevalence of each estimated directly as a function of respondent gender and national income. We reanalyze the material-costs topic, which most directly operationalizes the paper's proposed mechanism.

Clark & Dolan (2021), ***Pleasing the Principal: U.S. Influence in World Bank Policymaking*****.** The paper examines how World Bank policy conditionality reflects the preferences of powerful states, especially the United States. It shows that countries politically aligned with the U.S. receive fewer and less stringent policy conditions. The mechanism is primarily indirect, with World Bank staff designing programs consistent with U.S. preferences rather than explicit bargaining. The data consist of World Bank Development Policy

Financing (DPF) loan conditions from 2005–2018, including detailed text and categorical coding of policy requirements. Each observation captures policy conditions (prior actions) imposed on borrowing countries as part of loan agreements.

Feltovich & Giovannoni (2024), ***Campaign Messages, Polling, and Elections*****.** The article examines how candidates' track records and campaign messaging jointly affect voter welfare, pairing a formal model with a laboratory experiment. The texts are campaign messages written by experimental subjects playing incumbent and challenger politicians in repeated elections. Coders classify each message along five dimensions that may co-occur: positive self-promotion, negative campaigning against the opponent, claims about one's own competence, claims about past performance, and promises about future performance. These message-level labels are used directly as dependent variables in regressions on the experimental conditions. We reanalyze negative campaigning, the behavior behind the paper's sharpest prediction – that challengers attack incumbents more as incumbent performance worsens.

Gilardi et al. (2021), ***Policy Diffusion: The Issue-Definition Stage*****.** The paper studies policy diffusion at the issue-definition stage, asking whether prior smoking-ban adoptions in other U.S. states predict how later states frame the issue. Using structural topic models, it finds that prior adoptions predict practical, observable frames such as rules, enforcement, bars and restaurants, casinos, and local legislation, but not normative frames such as freedom. The paper concludes that diffusion shapes policymaking before adoption by affecting how issues are publicly defined. The data are newspaper paragraphs about anti-smoking laws from 49 newspapers covering 49 U.S. states between 1996 and 2013. The final corpus contains 52,675 relevant paragraphs after broad keyword retrieval and filtering with crowd annotation and machine-learning classification.

Jung (2020), ***The Mobilizing Effect of Parties' Moral Rhetoric*****.** The article asks whether parties' use of moral rhetoric mobilizes their own supporters. The texts are party manifestos from six English-speaking democracies (Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States), drawn from the Comparative Manifesto Project and divided into quasi-sentences. Each quasi-sentence is classified as moral or non-moral adapting from the Moral Foundations Dictionary, which flags language invoking concerns such as care, fairness, loyalty, authority, and sanctity. These document-level labels are not used directly: they are aggregated to the manifesto level as the share of moral quasi-sentences, which serves as an independent variable predicting copartisan turnout. We reanalyze the moral/non-moral distinction itself, the construct on which the paper's entire argument rests.

Lacombe (2019), ***The Political Weaponization of Gun Owners*****.** Lacombe studies why the NRA has been unusually effective at mobilizing gun owners and influencing U.S. gun politics. The paper argues that the NRA cultivated a politicized gun-owner social identity through long-term membership communications and programs, then used perceived threats to that identity to mobilize political action. The conclusion is that this identity-based mobilization is an important, often overlooked channel of interest-group power. The paper analyzes two original text corpora: NRA American Rifleman editorials from 1930-2008 and gun-control-related letters to the editor from four major U.S. newspapers: the New York Times, Arizona Republic, Atlanta Journal-Constitution, and Chicago Tribune. The newspaper letters are ordinary citizens' public arguments about gun control, excluding elites and NRA officials.

Mattingly et al. (2025), ***Chinese State Media Persuades a Global Audience*****.** The article shows, through a nineteen-country survey experiment, that exposure to Chinese state media moves global audiences toward seeing the "China model" as superior to the American one. The text analysis characterizes the messaging itself, using the descriptions of videos published by China's CGTN and the United States' ShareAmerica. A topic model sorts these descriptions into topics, grouped by the authors into portrayals of China's political model (responsive institutions, competent leadership, and Western dysfunction) and its economic model (poverty alleviation, infrastructure, and trade and innovation), alongside other news and cultural content.

The topics are used in aggregate, as corpus-level proportions that describe typical state-media messaging and motivate the choice of representative videos used as experimental stimuli; they do not enter a regression. We reanalyze the Western-dysfunction topic, the most clearly defined and substantively interpretable of the individual topics from our reading.

Osnabrugge et al. (2021), ***Playing to the Gallery: Emotive Rhetoric in Parliaments.*** The paper studies when and why politicians use emotive rhetoric in parliamentary debates. It argues that emotive rhetoric is strategically deployed to appeal to voters, especially when speeches reach a broader public audience. Using large-scale data from UK and Irish parliaments, it finds that high-profile debates (e.g., Prime Minister's Questions) feature significantly more emotive language. The data consist of nearly one million parliamentary speeches from the UK House of Commons (2001–2019) and a comparable dataset from the Irish Parliament. These are formal legislative speeches delivered in different types of debates that vary in public visibility and audience size.

Parthasarathyetal et al (2019), ***Deliberative Democracy in an Unequal World.*** The data are verbatim transcripts and translations of 50 village assembly meetings held on Republic Day 2014 in rural Tamil Nadu, India. Each document is an uninterrupted speech, with speaker gender and position identified; the full corpus contains 1,736 speeches from gram sabha proceedings. A text should be labeled Yes for fund allocation when Discussion of government funds, budgets, allocations, or rupee amounts.

Schub (2022), ***Informing the Leader: Bureaucracies and International Crises.*** The paper examines how bureaucratic position shapes the information advisers provide to leaders during international crises. Using over 5,400 advisory texts from U.S. Cold War crises, it argues that bureaucratic specialization affects informational content and uncertainty rather than policy preferences. Foreign policy bureaucracies emphasize political attributes of adversaries and express greater uncertainty, whereas military bureaucracies emphasize military characteristics. The data consist of internal U.S. foreign policy deliberation texts during Cold War international crises, including memoranda, NSC discussions, presidential advisory meetings, and FRUS documents. The corpus contains adviser-level speech acts from senior officials across bureaucracies such as the State Department, Defense Department, CIA, NSC, and Joint Chiefs of Staff.

Thrall (2025), ***Informational Lobbying and Commercial Diplomacy.*** The paper examines how informational lobbying shapes bilateral diplomacy, arguing that diplomats prioritize issues emphasized by the interest groups supplying them with information. Using the expansion of American Chambers of Commerce abroad and oral histories from U.S. diplomats, the paper shows that diplomats exposed to active AmChams devote greater attention to commercial diplomacy. The findings suggest that business interests influence foreign policy by subsidizing diplomats' information environment rather than through coercive lobbying. The paper uses approximately 1,500 oral history interviews conducted with retired U.S. diplomats by the Association for Diplomatic Studies and Training (ADST). The interviews describe diplomats' experiences across embassy postings worldwide from roughly the 1940s–2000s, including discussions of trade, investment, security, human rights, and bilateral relations.

Table S1: Summary of Text Classification Tasks for 14 Studies

Paper	Texts (unit)	Classification method	Reanalysis target	Use in original analysis
Arias (2022)	UN General Assembly General Debate speech	Structural topic model	Climate security	Analyzed as dependent variable
Blaydes et al. (2018)	Political advice texts	Structural topic model	Multiple political content	Describes topic trends over time
Blumenau and Lauderdale (2018)	European Parliament roll-call voting records	Structural topic model	Finances	Distinguishes crisis-related votes to estimate how crises altered voting cleavages
Bush & Clayton (2023)	Open-ended survey responses (response)	Structural topic model	Material-costs topic	Used directly; topic prevalence modeled on gender and income
Clark & Dolan (2021)	World Bank policy conditionality	Pre-existing categories	Fiscal policy	Reanalyzed as dependent variable
Feltovich & Giovannoni (2024)	Lab campaign messages (message)	Human coders, majority of three	Negative campaigning	Used directly; dependent variable on experimental conditions
Gilardi et al. (2021)	Newspaper paragraphs on smoking restrictions	Structural topic modeling	Smoking ban enforcement	Measures issue framing to examine how policy diffusion influences the definition of policy problems before adoption

S1.2 Details of Codebook Generation

For generating paper-specific prompts, we first upload the paper PDF to the ChatGPT interface and instruct the model to generate each component using the same prompt. We used the following prompt:

> We are developing a prompt for an LLM to classify texts used in a political science paper. I have attached the paper as a PDF. Our goal is to classify whether a text should be categorized as category name as defined and used in the paper's analysis. Please create a high-quality prompt for this task. Specifically, I want you to provide the following components:
>
> **Definition:** Provide a definition of the target category in no more than 10 words. Whenever possible, base the definition on how the concept is used in the attached paper.
>
> **Classification rules:** Provide a classification rule of no more than 10 words for each of the following:
>
> when the label should be Yes, when the label should be No, and what kinds of cases are tricky or borderline.
>
> **Label wording:** Provide five keywords and related expressions for the target category. Include alternative label names, synonyms, and close paraphrases when they are semantically relevant. Keep this concise while covering the major cases.
>
> **Summary of the paper:** Provide a summary of the paper in no more than three sentences, covering the research question, context, and main findings.

Table S1: Summary of Text Classification Tasks for 14 Studies (continued)

Paper	Texts (unit)	Classification method	Reanalysis target	Use in original analysis
Jung (2020)	Party manifestos (quasi-sentence)	Moral Foundations dictionary	Moral vs. non-moral rhetoric	Aggregated to party level; independent variable for turnout
Lacombe (2019)	NRA and gun-rights advocacy communications	Structural topic modeling	Gun regulation	Measures how advocacy organizations construct gun-owner identity; descriptive
Mattingly et al. (2025)	State-media video descriptions (description)	LDA topic model	Western dysfunction topic	Aggregated to corpus proportions; descriptive, not included in a regression
Osnabrugge et al. (2021)	Parliamentary debates	Supervised learning	Emotive vs. non-emotive	Reanalyzed as dependent variable
Parthasarathy et al. (2019)	Village assembly transcripts	Structural topic modeling	Fund allocation	Measures who speaks, sets the agenda, and is heard in deliberative forums
Schub (2022)	Foreign-relations speech (memo)	Supervised learning	Political content	Tests whether political departments convey more political-content advice
Thrall (2025)	Diplomat interview scripts	Supervised learning	Commercial issues	Reanalyzed as dependent variable

Summary of the data: Provide a summary of the text data in no more than two sentences. Describe the source and context of the texts (e.g., speeches in the United Nations General Assembly or parliamentary speeches in India). Focus on the data itself rather than how it is analyzed."

See Table S2 for the justification for each component.

For each study we submitted the prompt above, together with the full text of the article as a PDF, to LLM (GPT-5.5) and recorded the five returned components: definition, classification rules, label wording, paper summary, and data summary. Holding the prompt fixed across papers ensures that only the paper-specific content varies, while the length constraints (a ten-word definition, ten-word rules, five keywords, a three-sentence paper summary, and a two-sentence data summary) keep the resulting codebooks concise and comparable across studies. We then verified each generated component against the source article. See Table S3 for the components used for each paper.

Table S2: General Prompt Components and Their Justification (Codebook-Style)

Component	**Role in Prompt**	**Justification**
Role	Assigns the model a specific analytical perspective	Directs the model to adopt the relevant domain expertise and evaluative stance, improving task alignment and reducing inconsistent interpretations.
Definition	Provides a formal conceptualization of the construct	Aligns the model with the intended theoretical meaning, ensuring consistency with the research concept rather than colloquial usage.
Classification Rules (Yes / No / Tricky Cases)	Specifies inclusion (Yes), exclusion (No), and ambiguous (Tricky) conditions for classification	Operationalizes the abstract concept into an explicit decision boundary: inclusion rules define positive cases, exclusion rules prevent category leakage, and tricky cases guide decisions under ambiguity. This structured separation improves consistency, reduces false positives, and enhances robustness in edge cases.
Label Wording	Defines the vocabulary of allowable outputs	Standardizes the semantic space of labels, ensuring consistency across observations and enabling aggregation and downstream analysis.
Summary of Paper/-Data	Provides contextual information about the substantive setting and the dataset	Anchors interpretation in the research context, helping the model infer relevance when signals are indirect or mechanism-based rather than explicit. Clarifies the distribution of topics and potential co-occurrence of issue areas, helping calibrate expectations and avoid overgeneralization.
Reasoning	Specifies the inferential process for classification	Encourages the model to evaluate the text against the definition systematically and articulate the basis for its decision, improving transparency and conceptual consistency.

Table S3: Text classification across the studies

Study	Prompt
Arias (2022)	**Topic**: Climate security **Definition**: A text should be labeled Yes for climate security when Climate change is framed as an international security threat. **Classification Rules** Yes - Climate described as security, conflict, instability, or UNSC issue. No - Climate discussed without security, conflict, threat, or UNSC framing. Tricky/borderline - Existential harm or sea-level threats without security jurisdiction. **Label Wording** Climate security Climate securitization Climate as security threat Threat to international peace and security Climate-linked conflict, instability, terrorism, or state failure **Summary of Paper** Arias asks why some states frame climate change as a security issue in the United Nations, and which states are most likely to make such securitizing moves. The paper argues that securitization is shaped by agenda-control incentives: P5 states benefit from moving issues toward the Security Council, while highly vulnerable states such as SIDS may resist losing agenda control. Empirically, the paper finds that UN climate discourse overall is not securitized, P5 states are more likely to securitize climate change, and SIDS are less likely to do so. **Summary of Data** The text data are speeches by state representatives in the UN General Assembly General Debate, filtered to speech segments discussing climate change. The analysis uses 4,525 climate-related speech segments from 1,987 speeches, with the earliest climate-related speech segment appearing in 1984.
Blaydes et al. (2018)	**Topic**: Art of Rulership **Definition** Practical guidance on governing, exercising power, and ruling effectively. **Classification Rules** YES - Describes ruler's duties, governance practices, or statecraft strategies. NO - Focuses on religion, personal morality, or non-political life. Tricky / borderline - Moral advice tied directly to governing decisions or authority. **Label Wording** Statecraft, Governance advice, Ruling practices, Political management, Exercise of power / kingship guidance **Summary of Paper** The paper compares medieval Christian and Islamic political advice texts ("mirrors for princes") to examine how ideas of governance evolved across regions. Using text-as-data methods, it identifies shared thematic structures and tracks their evolution over time. It finds broad similarities across traditions but key divergences, including declining religious emphasis in Europe and shifting ruler-centered discourse in the Islamic world. **Summary of Data** The data consist of translated medieval political advice texts addressed to rulers in Christian Europe and the Islamic world. These texts include guidance on governance, morality, religion, and social order, often written by elites for kings, courts, or political audiences.

Continued on next page

Table S3 continued

Study	Prompt
Bush & Clayton (2023)	**Topic**: taxes **Definition** Tax-related material costs of climate action. **Classification rules** Yes -Mentions taxes, fees, prices, or cost increases. No - Mentions non-tax harms, benefits, or general climate concern. Tricky/borderline - Code broad "cost" claims only if tax-like or monetary. **Label wording** Taxes, Tax burden, Carbon tax / fossil-fuel tax, Fees, penalties, mandatory payments, Higher costs / cost-of-living increases **Summary of Paper** Bush and Clayton study why the gender gap in climate concern is larger in wealthier countries. They argue that as countries become wealthier, climate mitigation is perceived as more costly and less beneficial, especially among men, who are more likely to associate decarbonization with material and psychological costs. Using cross-national surveys, an original 10-country survey, and focus groups in Peru and the United States, they find that men's climate concern declines more sharply with economic development and that men in wealthier countries more often emphasize costs of climate action. **Summary of Data** The relevant text data are open-ended survey responses from citizens in 10 countries in the Americas and Western Europe, collected through Netquest in 2019–2020, asking how acting to stop climate change might personally harm respondents. The "Taxes" category appears in a structural topic model of these open-ended responses; the paper labels Topic 2 as "Taxes," associated with words such as "increase," "tax," and "cost," and finds it more common in wealthier countries.
Clark & Dolan (2021)	**Topic**: fiscal policy **Definition** Government taxation, spending, debt, and budgetary policy actions **Classification rules** Yes - Mentions taxes, spending, deficits, debt, or fiscal measures No - Discusses non-fiscal policies (e.g., monetary, foreign, social) Tricky / borderline - Economic discussion without explicit fiscal instruments or government action **Label wording** fiscal policy / budget policy, taxation and government spending, deficit, debt, public finance, revenues, expenditures, austerity, government budget management **Summary of Paper** The paper examines how World Bank policy conditionality reflects the preferences of powerful states, especially the United States. It shows that countries politically aligned with the U.S. receive fewer and less stringent policy conditions. The mechanism is primarily indirect, with World Bank staff designing programs consistent with U.S. preferences rather than explicit bargaining. **Summary of Data** The data consist of World Bank Development Policy Financing (DPF) loan conditions from 2005–2018, including detailed text and categorical coding of policy requirements. Each observation captures policy conditions (prior actions) imposed on borrowing countries as part of loan agreements.

Continued on next page

Table S3 continued

Study	Prompt
Feltovich & Giovannoni (2024)	**Topic**: Contains negative campaigning **Definition:** Criticizes opponent's behavior, quality, or performance. Yes - Attacks opponent's behavior, quality, competence, or record. No - Promotes sender without criticizing the opponent. Tricky/borderline - Neutral opponent references without criticism are borderline. **Label wording** attack; criticism; opponent failure; poor performance; low quality **Summary of Paper** Feltovich and Giovannoni study how politicians' track records and campaign messages interact to shape elections and voter welfare. They develop and test a laboratory election model where incumbents and challengers vary in quality, voters observe performance imperfectly, and candidates may send campaign messages before elections. The paper finds that both higher quality variability and campaigning improve voter welfare, while challengers are more likely to use negative campaigning when incumbents perform poorly. **Summary of Data** The text data are short, free-form campaign announcements written by candidates in a laboratory election experiment after a straw poll and before the final election. Messages were 0–140 characters, written in English, and sent by incumbents and challengers; research assistants coded whether each message was positive, negative, about quality, about past performance, or about future decisions.
Gilardi et al. (2021)	**Topic**: Smoking ban enforcement **Definition**: A text should be labeled Yes for smoking ban enforcement when Implementation and legal enforcement of smoking restrictions. **Classification Rules** Yes - Discusses compliance, penalties, lawsuits, or enforcement of smoking bans. No - Discusses smoking bans without enforcement or compliance issues. Tricky/borderline - Mentions rules, unless implementation or penalties are central. **Label Wording** enforcement, compliance, penalties or fines, lawsuits or legal challenges, implementation **Summary of Paper** The paper studies policy diffusion at the issue-definition stage, asking whether prior smoking-ban adoptions in other U.S. states predict how later states frame the issue. Using structural topic models, it finds that prior adoptions predict practical, observable frames such as rules, enforcement, bars and restaurants, casinos, and local legislation, but not normative frames such as freedom. The paper concludes that diffusion shapes policymaking before adoption by affecting how issues are publicly defined. **Summary of Data** The data are newspaper paragraphs about anti-smoking laws from 49 newspapers covering 49 U.S. states between 1996 and 2013. The final corpus contains 52,675 relevant paragraphs after broad keyword retrieval and filtering with crowd annotation and machine-learning classification.

Continued on next page

Table S3 continued

Study	Prompt
Jung (2020)	**Topic**: Moral rhetoric **Definition**: Frames politics as moral right, wrong, virtue, or transgression. **Classification Rules** Yes – Uses moral values to justify political claims. No – Only factual, instrumental, or policy-pragmatic language. Tricky / Borderline – Moral words used as policy labels or proper nouns. **Label Wording** Moral appeal — ethical appeal, moral framing Right and wrong — morally right, morally wrong Virtue/transgression — dignity, justice, fairness, harm Moral foundations — care, fairness, loyalty, authority, sanctity Values-based rhetoric — principles, duties, rights, common good **Summary of Paper** The paper asks whether parties' use of moral rhetoric affects voter behavior, focusing on party campaign communication in six English-speaking democracies. Jung argues that moral rhetoric activates positive emotions among copartisans who are exposed to party rhetoric, especially politically aware voters. The paper finds that higher moral rhetoric in party manifestos increases turnout among more educated copartisans, and that survey experiments and British Election Study data support the positive-emotion mechanism. **Summary of Data** The core text data are party manifestos from the Manifesto Corpus, covering 64 party manifestos from six English-speaking democracies across 18 elections. The unit of text classification is the quasi-sentence, a policy statement that may be a full sentence or part of a sentence. The appendix also validates the manifesto-based measure against a smaller set of campaign speech texts from Australia and New Zealand.
Lacombe (2019)	**Topic**: Gun regulation **Definition**: A text should be labeled Yes for gun control when Debates over firearm laws, restrictions, ownership, and gun rights. **Classification Rules** Yes - Text discusses firearm policy, laws, regulation, or gun rights. No - Text discusses guns without policy, regulation, or rights debate. Tricky/borderline - Identity, crime, or safety claims tied to regulation. **Label Wording** gun control, firearm regulation, gun laws, gun rights, Second Amendment / firearms legislation **Summary of Paper** Lacombe studies why the NRA has been unusually effective at mobilizing gun owners and influencing U.S. gun politics. The paper argues that the NRA cultivated a politicized gun-owner social identity through long-term membership communications and programs, then used perceived threats to that identity to mobilize political action. The conclusion is that this identity-based mobilization is an important, often overlooked channel of interest-group power. **Summary of Data** The paper analyzes two original text corpora: NRA American Rifleman editorials from 1930-2008 and gun-control-related letters to the editor from four major U.S. newspapers: the New York Times, Arizona Republic, Atlanta Journal-Constitution, and Chicago Tribune. The newspaper letters are ordinary citizens' public arguments about gun control, excluding elites and NRA officials.

Continued on next page

Table S3 continued

Study	Prompt
Mattingly et al. (2025)	**Topic**: Western dysfunction **Definition** Claims Western democracies suffer instability, racism, violence, or failure. **Classification rules** Yes - Criticizes Western political/social disorder or democratic failure. No - Discusses China's strengths without criticizing the West. Tricky/borderline - Comparisons count only with explicit Western dysfunction. **Label wording** Western dysfunction, Western political failure, U.S./Western instability, Racism, protests, or political violence in the West, Democratic disorder / liberal-democratic crisis **Summary of Paper** Mattingly et al. study whether Chinese and American state media can persuade global audiences to prefer China's authoritarian model or the U.S. democratic model. The paper combines content analysis of Chinese and American external media with a randomized survey experiment in 19 countries. It finds that Chinese state media, especially messages emphasizing performance, growth, stability, and competent governance, substantially increases support for the Chinese model, while U.S. messaging is less persuasive. **Summary of Data** The relevant text data are descriptions of external state-media video segments, especially CGTN videos posted on YouTube from 2020–2021 and ShareAmerica videos produced by the U.S. Department of State. For "Western dysfunction," the category comes from the paper's topic-model analysis of CGTN videos promoting China's political model, where this theme includes stories about protests, racism, and political violence in the United States that contrast Western disorder with Chinese stability and responsiveness.
Osnabrugge et al. (2021)	**Topic**: Emotive Rhetoric **Definition** Language eliciting emotional response beyond literal policy content **Classification rules** Yes- Uses emotionally charged words to evoke feelings or reactions No - Neutral, technical, or purely informational without emotional tone Tricky / borderline - Strong language without clear emotional appeal or intent **Label wording** emotional appeal / emotive language, affective rhetoric / loaded language, emotionally charged / expressive tone, fear, outrage, admiration framing, persuasive emotional signaling **Summary of Paper** The paper studies when and why politicians use emotive rhetoric in parliamentary debates. It argues that emotive rhetoric is strategically deployed to appeal to voters, especially when speeches reach a broader public audience. Using large-scale data from UK and Irish parliaments, it finds that high-profile debates (e.g., Prime Minister's Questions) feature significantly more emotive language. **Summary of Data** The data consist of nearly one million parliamentary speeches from the UK House of Commons (2001–2019) and a comparable dataset from the Irish Parliament. These are formal legislative speeches delivered in different types of debates that vary in public visibility and audience size.

Continued on next page

Table S3 continued

Study	Prompt
Parthasarathyetal et al. (2019)	**Topic**: Fund allocation **Definition**: A text should be labeled Yes for fund allocation when Discussion of government funds, budgets, allocations, or rupee amounts. **Classification Rules** Yes - Mentions allocating, approving, spending, or reporting public funds. No - Mentions services without money, budgets, or fund distribution. Tricky/borderline - Schemes count only when funding or amounts are discussed. **Label Wording** Fund allocation, Allocation of funds, Budget / public expenditure, Government funds / panchayat funds, Rupee amounts / allotted funds / scheme funding **Summary of Paper** The paper studies deliberation in constitutionally mandated village assemblies, or gram sabhas, in rural Tamil Nadu, India, using text-as-data methods on meeting transcripts. It asks whether these assemblies provide meaningful opportunities for citizens to speak, set the agenda, and receive responses from officials, especially under conditions of gender inequality. The authors find that the assemblies are not merely talking shops, but women are less likely than men to speak, shape discussion, or receive relevant official responses, while female village presidents improve women's likelihood of being heard. **Summary of Data** The data are verbatim transcripts and translations of 50 village assembly meetings held on Republic Day 2014 in rural Tamil Nadu, India. Each document is an uninterrupted speech, with speaker gender and position identified; the full corpus contains 1,736 speeches from gram sabha proceedings.
Schub (2022)	**Topic**: political content (vs. military) **Definition** Discussion of adversary domestic politics and political resolve. **Classification Rules** Yes - Focuses mainly on opponent's political situations, such as: -domestic political landscape: the configuration of domestic actors, institutions, and political constraints that shape a state's policy choices and strategic behavior. -difficulties of converting battlefield outcomes to desired political end states. -adversary resolve: the willingness of an opponent to continue fighting in pursuit of its objectives. No - (military, only two categories in total) Focuses mainly on military capabilities, quality of forces, force operations, force movements, or force locations. Tricky / Borderline -according to the paper, we should code "yes" if the text's focus or objective is about political content but mentioning its causes due to other factors such as military. That is, when it's not directly about military actions. -mixed operational-political assessments: sometimes it may be hard to distinguish whether political content is the focus when it co-appears with other factors such as military, it may be hard to code. **Label Wording** Political content, Political attributes, Domestic political conditions, Resolve and political will, Governance and regime dynamics **Summary of Paper** The paper examines how bureaucratic position shapes the information advisers provide to leaders during international crises. Using over 5,400 advisory texts from U.S. Cold War crises, it argues that bureaucratic specialization affects informational content and uncertainty rather than policy preferences. Foreign policy bureaucracies emphasize political attributes of adversaries and express greater uncertainty, whereas military bureaucracies emphasize military characteristics. **Summary of Data** The data consist of internal U.S. foreign policy deliberation texts during Cold War international crises, including memoranda, NSC discussions, presidential advisory meetings, and FRUS documents. The corpus contains adviser-level speech acts from senior officials across bureaucracies such as the State Department, Defense Department, CIA, NSC, and Joint Chiefs of Staff.

Continued on next page

Table S3 continued

Study	Prompt
Thrall (2025)	**Topic**: commercial issues **Definition** Commercial diplomacy and cross-border business-related diplomatic issues. **Classification rules** Yes - Discusses trade, investment, firms, taxation, or commercial regulation. No - Focuses solely on security, military, humanitarian, or cultural issues. Tricky / Borderline - Economic development discussion without explicit business or trade relevance. **Label wording** trade policy / trade relations, investment promotion / foreign investment, business regulation / commercial regulation, taxation / tariffs / market access, exports, firms, commerce, industry, intellectual property **Summary of Paper** The paper examines how informational lobbying shapes bilateral diplomacy, arguing that diplomats prioritize issues emphasized by the interest groups supplying them with information. Using the expansion of American Chambers of Commerce abroad and oral histories from U.S. diplomats, the paper shows that diplomats exposed to active AmChams devote greater attention to commercial diplomacy. The findings suggest that business interests influence foreign policy by subsidizing diplomats' information environment rather than through coercive lobbying. **Summary of Data** The paper uses approximately 1,500 oral history interviews conducted with retired U.S. diplomats by the Association for Diplomatic Studies and Training (ADST). The interviews describe diplomats' experiences across embassy postings worldwide from roughly the 1940s–2000s, including discussions of trade, investment, security, human rights, and bilateral relations.

S1.2.1 Updated Codebook

We use the updated codebooks for Section 3.2.3 and Appendix S4.2. For each paper, one of our authors carefully read the paper and then crafted a more detailed codebook. Because these updated prompts have study-specific structures, we present the complete prompts separately rather than in tabular form. For Arias, the full-corpus analysis in Appendix S4.2 uses the substantively refined prompt shown here rather than the standardized prompt 64 used for the main 100-text replication; the appendix therefore targets annotations under the refined codebook. This implements the refinement-before-scaling workflow described in Section 4.1.

Arias (2022)

Your role is an expert classifier of textual topics from a political science journal article.

Carefully read the text provided below and determine whether the text belongs to the topic of "climate security" or not. Before you start, carefully review the following instructions.

Codebook for Climate Security

Definition: A text should be labeled Yes when it securitizes climate change or the environment.

This definition is met if any of the following conditions are present:

(1) climate change or the environment is directly described as a security issue;

(2) climate change or the environment is directly compared to traditional security issues, such as war, conflict, or terrorism; or

(3) climate change or the environment is presented as an existential threat that requires emergency or urgent measures.

A text should not be considered securitizing when security terms or issues are discussed separately from climate change or the environment.

Classification Rules

Yes — Label Yes if the text explicitly frames climate change or the environment as a security threat; links it to conflict, violence, war, terrorism, military concerns, or instability; or presents it as a direct existential threat requiring urgent or emergency action.

Positive example 1:
"Of the major issues confronting us today there is probably none that has captured our imagination more than the deterioration of the natural environment. The greenhouse effect, global warming, the depletion of the ozone layer, acid rain, waste dumping, forest depletion, and drift-net fishing, threaten our existence. The greenhouse effect and the resulting global warming and the rise in sea level constitute a direct threat, no longer dismissible, to our survival. We in the Pacific have more cause than many to be deeply concerned about it since if the rise in sea level is significant some of our islands and coastal areas may become permanently inundated. The possibility that entire countries may also drown is almost beyond comprehension."

Positive example 2:
"Alone we can do little, but together we can achieve much. The challenges that we face globally, from climate change to refugees to war and violence, require urgent action now. The world does not have the luxury of time. We have talked. We have debated. We have postulated and hypothesized. We have studied and analysed. We must now act."

No — Label No when climate change or the environment is discussed without a security, conflict, threat, or urgent existential framing. Also label No when security-related terms are present but refer to separate issues rather than climate change or the environment.

Negative example:
"The United Nations Conference on Environment and Development to be held in 1992, in which so many high expectations have been placed, offers an opportunity for the international community to address the global ecological challenges that confront mankind. Any effective, collective action in the major areas that the Conference is expected to tackle—whether climate change or biodiversity and genetic heritage—entails envisaging new methods of funding, such as the introduction of some form of international taxation."

Tricky/borderline — Existential, survival, or sea-level language requires context. Label Yes only when it clearly presents climate or environmental harm as a direct security or existential threat and conveys urgency or the need for emergency action. Label No when it merely describes environmental harm, vulnerability, or physical impacts without this framing.

Label Wording
Climate security
Climate securitization
Climate as a security threat
Threat to international peace and security
Climate-linked conflict, instability, terrorism, violence, or state failure
Security, military, war, conflict, instability, unrest, violence, emergency, urgent, sovereignty, survival, existential

Note: These terms are diagnostic only when they are substantively connected to climate change or the environment; their presence alone is not sufficient for a Yes label.

Summary of Paper

Arias asks why some states frame climate change as a security issue in the United Nations, and which states are most likely to make such securitizing moves. The paper argues that securitization is shaped by agenda-control incentives: P5 states benefit from moving issues toward the Security Council, while highly vulnerable states such as SIDS may resist losing agenda control. Empirically, the paper finds that UN climate discourse overall is not securitized, P5 states are more likely to securitize climate change, and SIDS are less likely to do so.

Summary of Data

The text data are speeches by state representatives in the UN General Assembly General Debate, filtered to speech segments discussing climate change. The analysis uses 4,525 climate-related speech segments from 1,987 speeches, with the earliest climate-related speech segment appearing in 1984.

Let's think step by step through all the above instructions when you process each text.

In your output, I want you to respond with 'Yes' if the text closely aligns with the definition of climate security, otherwise respond with 'No'. Respond only with 'Yes' or 'No'. Do not provide any other outputs or any explanation for your output.

Schub (2022)

Your role is an expert classifier of textual topics from a political science journal article.

Carefully read the text provided below and determine whether the text belongs to the topic described in the classification codebook. Before you start, carefully review the following instructions.

Classification codebook

Topic: political content

Definition
Discussion of adversary domestic politics and political resolve.

Classification Rules

Yes
Focuses mainly on opponent's political situations, such as:

- **domestic political landscape**: the configuration of domestic actors, institutions, and political constraints that shape a state's policy choices and strategic behavior.
- difficulties of converting battlefield outcomes to desired political end states.
- **adversary resolve**: the willingness of an opponent to continue fighting in pursuit of its objectives.

No
Focuses mainly on military capabilities, quality of forces, force operations, force movements, or force locations.

Tricky / Borderline

- according to the paper, we should code "yes" if the text's focus or objective is about political content but mentioning its causes due to other factors such as military. That is, when it's not directly about military actions.
- mixed operational-political assessments: sometimes it may be hard to distinguish whether political content is the focus when it co-appears with other factors such as military, it may be hard to code.

Paper examples:

Military Texts

Example 1: Ground Attacks on Base Camps in Cambodia: Attached at Tab A is a brief summary of the two options for ground attacks on enemy base camps in Cambodia submitted by General Abrams on March 30. In developing plans for potential operations against enemy base areas, General Abrams was asked to consider two possibilities: An attack against targets of high military priority which could involve the use of US forces if necessary. Any other operation which would reduce the necessity of the involvement of US forces. With respect to military priority, MACV considered an attack on Base Area 352/353 (COSVN Hq) to be the most lucrative. He made the following significant points about this base area. [1]

Example 2: The Chiefs believe that ground action against the North Vietnamese effort is adequate to reverse the situation. Air strikes on the three targets are not necessary from a military point of view. However, a South Vietnamese attack on their target is acceptable. [2]

Political Texts

Example 1: Iran. The two leading US academic experts on Iran, James Bill and Marvin Zonis, recently were debriefed in the Department following their separate visits to Iran at the end of November. In a wide range of Iranian contacts, both men found intense rage against the Shah personally. This is a marked change from the past when Iranians were content to blame their troubles on the Government and the Shah's advisers. Both professors see a slim chance that the Shah might retain a minimal role as constitutional monarch, but only if he moves quickly to negotiate a political compromise. They assess the opposition as very strong and extremely

well-organized. Everywhere they found an eagerness for the US to play a decisive role in promoting a political solution to Iran's crisis. [3]

Example 2: In spite of economic difficulties there is no solid evidence that Trujillo's fall is imminent. Trujillo rules by force and will presumably remain in power as long as the armed forces continue to support him. While there is evidence of dissatisfaction on the part of a few officers there is as yet no cogent evidence of large-scale defection within the officer corps. The underground opposition to Trujillo composed of business, student and professional people is believed to be predominantly anti-Communist. They have substantially increased in numbers in recent years but have been unable to move effectively against Trujillo. In addition to opposition groups in the Dominican Republic, there are numerous exile groups located principally in Venezuela, Cuba, United States and Puerto Rico. In some cases these groups have been infiltrated by pro-Castro or pro-Communist elements. [4]

Label Wording

- Political content
- Political attributes
- Domestic political conditions
- Resolve and political will
- Governance and regime dynamics

Summary of Paper

The paper examines how bureaucratic position shapes the information advisers provide to leaders during international crises. Using over 5,400 advisory texts from U.S. Cold War crises, it argues that bureaucratic specialization affects informational content and uncertainty rather than policy preferences. Foreign policy bureaucracies emphasize political attributes of adversaries and express greater uncertainty, whereas military bureaucracies emphasize military characteristics.

Summary of Data

The data consist of internal U.S. foreign policy deliberation texts during Cold War international crises, including memoranda, NSC discussions, presidential advisory meetings, and FRUS documents. The corpus contains adviser-level speech acts from senior officials across bureaucracies such as the State Department, Defense Department, CIA, NSC, and Joint Chiefs of Staff.

Let's think step by step through all the above instructions when you process each text.

In your output, I want you to respond with 'Yes' if the text closely aligns with the definition of political content, otherwise respond with 'No'. Respond only with 'Yes' or 'No'. Do not provide any other outputs or any explanation for your output.

S1.3 Details of Crowd Sourced Annotation

S1.3.1 Procedure of Data Collection

We fielded the crowd-sourced annotation experiment on CloudResearch Connect in June 2026. Prior to fielding, we preregistered the design and analysis on the Open Science Framework (OSF), obtained Institutional Review Board approval, and fielded a small pilot survey to test our instrument. Our sample of $N = 165$ is a U.S. census-matched demographic, selected from Connect's demographic targeting template. Each respondent is tasked with coding 20 texts as belonging to a topic or not, based on a standardized codebook. Given the difficulty of the task and to ensure response quality, we further implemented the following inclusion criteria, which we preregistered: respondents must be native English speakers, pass all attention and bot checks, not exceed Connect's default maximum time spent of 48 minutes, and have a platform approval rate of at least 95%, following prior conventions (Peer et al., 2014; Kennedy et al., 2020).

Our survey starts with a short set of demographic and baseline questions to gauge respondent's political knowledge, which we used as moderators in exploratory analyses. Respondents are then randomly assigned to one of 15 survey tasks: one for each of the 14 replicated papers and a second Schub (2022) task with an improved codebook. The respondent reads each task's codebook – the same codebook used by LLMs and experts for each of the 14 papers – and then codes that task's 20 texts as Yes/No. Figure S1 illustrates an example while Table S4 details our survey questions.

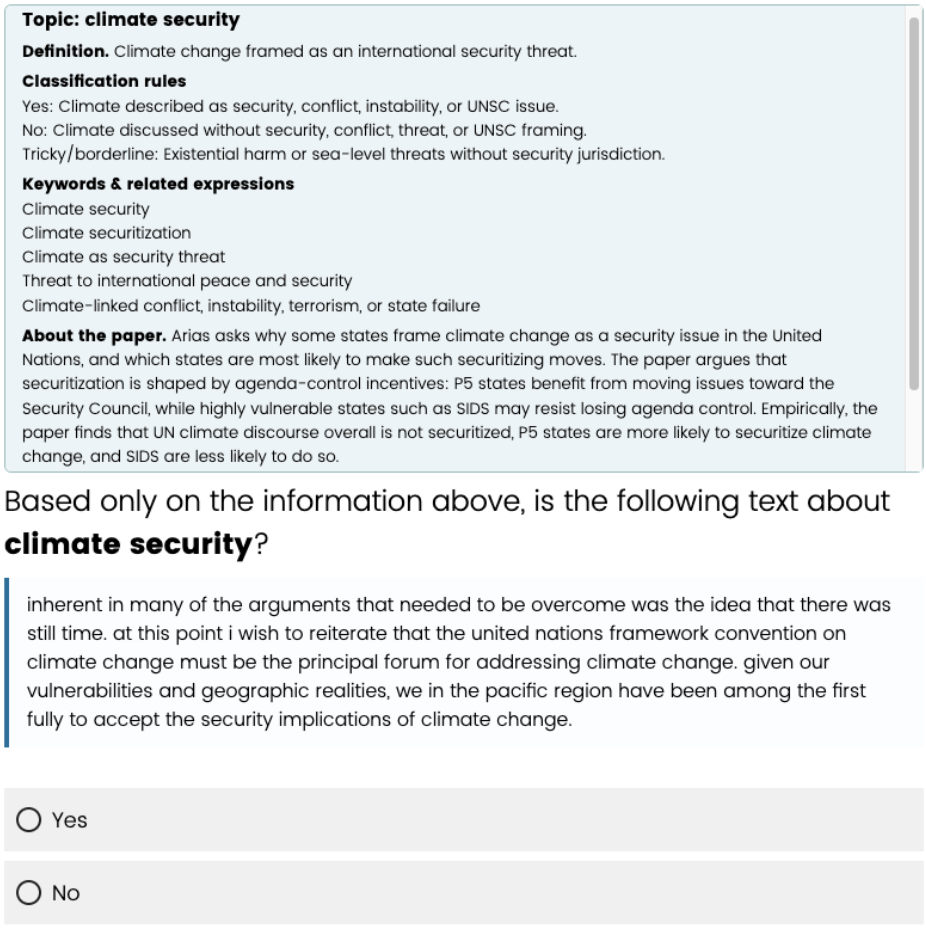

Figure S1: Example of coding task for the Arias (2022) paper.

Table S4: Survey Questionnaire

Item	Question
Baseline and moderator items	
Political knowledge	Which of the following best describes a political party manifesto or platform? *Options:* A document describing a party's policy positions (correct); A court ruling; A private tax document; A voter registration form; Not sure
News	How often do you follow news about politics or public policy? *Options:* Never; Once or a few times a month; A few times a week; Daily
Coursework	Have you ever taken a political science, government, public policy, or international relations course? *Options:* No; Yes, one; Yes, more than one; Not sure
Demographic items	
Age	What is your age?
Gender	What is your gender?
Education	What is the highest level of education you have completed?
Household income	What was your total household income last year (before taxes)?
Coding task	
Codebook	[Follows the same codebook as our expert and LLM annotation, comprising: topic, definition, classification rules, label wording, summary of paper, summary of data]
Classification ($\times$20)	Based only on the information above, is the following text about [the task's topic]? (Yes / No).
Post-task feedback	
Difficulty	How difficult was the classification task?
Confidence	How confident are you in how you classified the texts for the task?
Basis	Which best describes how you made your classifications? Select all that apply. *Options:* Used the definition; Used the rules; Used keywords / label wording; Used the summary of the paper and data; Used my own intuition; Used my background knowledge about the topic; Consulted other sources; Other

Our preregistered sample target is approximately 150 respondents, or 10 coders per task. Compared to having three expert coders, having 10 coders per paper reduces the likelihood of chance agreement across all coders by 128 times. Since we randomized the task assignment, we noticed a chance imbalance of less than 10 coders for some tasks and more than 10 for some others. To ensure we have at least 10 coders for each paper, we paused the survey after the initial 150 responses, excluded papers which already hit our target from our survey block randomizer, and continued the survey until we had 10 coders or more for the remaining papers. This gives us our final sample of 165 complete responses, with 10 to 13 coders per task. We report the sample demographics in Table S5 below.

Table S5: Demographic characteristics of survey respondents.

	n	%
Age		
18–24	12	7%
25–34	32	19%
35–44	37	22%
45–54	31	19%
55–64	27	16%
65 or older	25	15%
Prefer not to say	1	1%
Gender		
Woman	88	53%
Man	74	45%
Non-binary or another gender	1	1%
Prefer not to say	2	1%
Education		
High school diploma or GED	12	7%
Some college, no degree	28	17%
Associate degree	15	9%
Bachelor's degree	72	44%
Graduate or professional degree	37	22%
Prefer not to say	1	1%
Household income		
Under $25,000	22	13%
$25,000–$49,999	28	17%
$50,000–$74,999	30	18%
$75,000–$99,999	34	21%
$100,000–$149,999	26	16%
$150,000 or more	20	12%
Prefer not to say	5	3%
News consumption		
Daily	68	41%
A few times a week	64	39%
Once or a few times a month	28	17%
Never	5	3%
Politics/Policy coursework		
Yes, more than one	46	28%
Yes, one	53	32%
No	57	35%
Not sure	9	5%
Factual knowledge item		
Correct	149	90%
Incorrect	16	10%

S1.3.2 Text selection and Data Quality

For each coding task we selected 20 texts from the replicated papers, taken verbatim.[6] Out of the original 100 texts per paper used in our main analyses, we used stratified sampling to obtain the 10 "most agreed" and 10 "most disagreed" texts based on the LLM annotation. In other words, for each paper, we took the 10 texts with highest LLM disagreement (ranked by across-model Gini impurity) and 10 texts with lowest disagreement. This oversamples ambiguous texts by design, so that every task spans both straightforward and difficult cases and lets us compare coders across that range. For all analyses involving crowd annotations, we only use the annotations from the subsample of 20 texts per task, from crowd workers, experts, and LLMs.

Of note, Schub (2022) contributes one set of 20 texts, coded under both its original and its improved codebook. Holding the same set of 20 texts constant allows us to compare the effects of the codebook change since assignment to each task is randomized. Across the 14 papers, our text selection procedure yields 280 unique texts.

We capped each task at 20 texts to protect response quality and potential declines in data quality if the task was too long. We estimated the number of texts to assign for crowd coding based on the expected duration of the task, informed by our own expert coding and a pilot test. In our pilot test of $N = 42$ fielded shortly before the main study (also in June 2026 and on Connect with the same demographic selection and tasks), coding 20 texts took a median of 10.7 minutes and mean of 11.3. We found this to be short enough to hold a coder's attention while still giving enough texts to compare coders within a task. Our main survey took a median of 12.8 minutes and mean of 14.8. For all tasks, text order is randomized to reduce fatigue and learning effects.

Given that our analysis compares human crowd workers with LLMs, we took extensive care to prevent AI-assisted responses (Westwood, 2025). First, we programmed our survey on Qualtrics to prevent any copying and pasting of text content, raising the barrier to routing a text through an LLM. Second, we added a honeypot text field that is visible to bots but not to humans. Third, we enabled Qualtrics' bot-detection features, a CAPTCHA question, and duplicate-submission prevention.[7] Fourth, our survey provider, CloudResearch Connect, applies its own extensive anti-bot screening.[8]

We used our pilot to test the survey's length and data quality checks before fielding the main study. We generally found high data quality and good engagement across both the pilot and full survey: crowd agreement with the experts is fairly high, completion times are close to what we expected, and straight-lining is rare. The final sample of 165 respondents includes only those that pass every preregistered exclusion criteria. Numerous respondents failed our attention checks and are excluded while one respondent took more than eight hours and is excluded for exceeding Connect's maximum survey time. Respondents are pre-screened on Connect for native English language proficiency and a platform approval rate of at least 95%, so we do not collect any data from those that do not meet these criteria.

Additionally, among our final sample of 165, three respondents straight-lined all twenty answers. Given that we did not preregister straight-lining as an exclusion criterion, we retain these three responses in the main specification and drop them only as a robustness check. Appendix S3.2.6 reports that check and the other robustness analyses.

[6] Some texts are abruptly truncated at parts or contain mojibake. We chose to preserve all original texts without any modification to keep our replication exercise consistent. To reduce confusion, we informed respondents that "some texts may appear vague or incomplete — we ask that you answer to the best of your understanding."

[7] Accessed 23 July, 2026. `https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/fraud-detection/`

[8] Accessed 23 July 2026. `https://www.cloudresearch.com/resources/blog/ai-bot-threat-to-survey-research/`

S2 Examples of unambiguous and ambiguous texts

Table S6: Examples of unambiguous texts

Paper	Topic	Text
Arias (2022)	Climate security	Climate change is rightfully described as the most urgent threat facing mankind, and, at least for the next several months, it will remain at the top of the global diplomatic and negotiating agenda. But what is the challenge of climate change if not a risk to development, security and peace? What is the threat of climate change, if not a threat to the very notion of human survival and ecological balance?
Blumenau and Lauderdale (2018)	Finance	PURPOSE: to establish a new instrument for pre-accession, IPA II, in the framework of the reform of the EU external action financial instruments and following on from the unified Instrument for Pre-Accession Assistance, IPA, for the potential candidate countries to accession. PHILOSOPHY AND ACTION PLAN FOR EXTERNAL AID: what happens outside the borders of the EU can and does directly affect the prosperity and security of EU citizens. It is therefore in the interest of the EU to be actively engaged in influencing the world around us, including through the use of financial instruments.
Bush and Clayton (2023)	Taxes for climate policy	It would harm me through the increases in costs of energy.
Clark and Dolan (2021)	Fiscal policy	The Recipient has demonstrated enhanced predictability of budget execution as reflected by the fact that, during Fiscal Year 2009, at least 85% of Budget Heads expended between 95% and 105% of their total budgetary allocated funding for said year, as maintained to the date of this Agreement.
Feltovich and Giovannoni (2024)	Negative campaigning	The other challenger either failed in maths or lied to you.
Jung (2020)	Moral rhetoric	Changes have been made to the Sex Discrimination Act: no one can be discriminated against based on family responsibilities, while breastfeeding is now its own separate grounds of discrimination.
Mattingly et al. (2025)	Western dysfunction	The Wall Street showdown over GameStop share prices: is it a storm in a teacup or the tip of the iceberg? How troublesome is it for the U.S. bourse? A vaccine wall of defense against COVID slowly built.
Parthasarathy et al. (2019)	Fund allocation	Due to absence of ration shop facility due to rain water goes in. The food items are all spoiled. Please get it corrected.
Thrall (2025)	Commercial issues	But of even greater popular and political concern was the growing trade imbalance between Japan and the United States. I think the trade deficit with Japan was somewhere in the neighborhood of $3,000,000,000 annually. There was little point in making the argument that a negative bilateral trade imbalance had little or no economic significance. Despite all the efforts of the leading world economists beginning with Adam Smith, mercantilist theory still dominated popular thinking and the almost unanimous view was that a negative trade imbalance was a sign of economic weakness and that drastic action was necessary in order to see it eradicated or at least sharply reduced. Resentment on both sides grew as our manufacturers increasingly complained about the difficulties they were facing in exporting to Japan, which they blamed, primarily if not exclusively, on Japanese policies and practices designed to thwart imports.

Table S7: Examples of ambiguous texts

Paper	Topic	Text
Arias (2022)	Climate security	Climate change must be further mainstreamed into the work of the whole United Nations system, with a view to supporting efforts to help the transition to low-carbon economies consistent with sustainable development, to strengthen countries' adaptation and resilience in the face of climate change, and to minimize the possible security implications. In the light of diminishing natural resources, environmental degradation, extreme poverty, hunger and diseases, and social unrest, we agree with others that sustainable development has become the defining issue of our time. Our highly globalized and interdependent world means that we share not only the same challenges but a common fate.
Arias (2022)	Climate security	08-53141 40 being especially vulnerable. the recent hurricanes that left such a trail of destruction across the caribbean once again brought into sharp relief the acute vulnerabilities of small island states, such as the maldives, to global warming and climate change. for the maldives, climate change is not a distant possibility. it is happening now and is a reality that we are experiencing on a daily basis. the continuing degradation of the global environment is not only undermining our development process, but also seriously threatening the very survival of our people and the existence of our tiny country.
Blaydes et al. (2018)	Art of rulership	the Church of Christ. [5] We must make the same judgment about the end of the whole multitude as we do of one person.170 If, therefore, the end of a person were some good existing in that one alone, and the ultimate end of the multitude to be governed were similar in that the multitude should acquire such a good and preserve it, and if indeed such an ultimate end were a corporal one, either of one person or of the multitude, then the life and health of that body would be the duty of a physician. If the ultimate end were affluence and riches, a steward should be king of the multitude. If the good of knowing the truth were of the kind which the multitude could attain, the king would have the same duty as a professor. [6] It seems that the end of a multitude gathered together is to live according.
Blaydes et al. (2018)	Art of rulership	Any other person, have revealed to us the true sense of the oracle. Why do we then delay to offer him the crown, whom the Fates have ordained to be our king? The old men immediately quitted the sacred grove, and the chief of them, taking me by the hand, acquainted the people, who waited with impatience for their decision, that I had gained the prize. Scarcely had he done speaking, when a confused noise ran through the whole assembly. Everyone shouted for joy. The whole coast, and neighboring mountains, echoed with these words: "May the son of Ulysses, who resembles Minos, reign over the Cretans." After waiting a while, I made a sign with my hand, to ask to be heard. In the meantime, Mentor whispered thus in my ear: "Are you going to renounce your country? Will the ambition of reigning make you forget Penelope, who longs for you as her only …"
Clark and Dolan (2021)	Fiscal policy	The Recipient has reduced administrative costs arising from (A) the creation of a business and (B) the transfer of urban commercial real property (mutation d'immeubles urbains bâtis et non-bâtis) by amending the Recipient's Registration Code (Code de l'enregistrement) so as to reduce the registration fees in connection with: (i) the creation of a business; and (2) the transfer of urban commercial real property (mutation d'immeubles urbains bâtis et non-bâtis), each by at least fifty percent (50%).
Feltovich and Giovannoni (2024)	Negative campaigning	this is the best I can do. Q7. If anyone gives you more, they will have to sacrifice themselves, which they won't.
Lacombe (2019)	Gun control	The pioneer forefathers of the present generation of Californians would find it hard to believe their eyes and their ears if they were abroad in the Golden State today. Under the ostensible leadership of a member of the California Prison Board, a state-wide campaign has been launched to secure the necessary names on a petition to force a state-wide total-disarmament bill on the election ballot this fall. The bill on which the voters will be asked to register their vote would completely prohibit the retail or wholesale distribution of pistols or revolvers or any other weapon which may be concealed upon the person, and would also completely prohibit the possession of such weapons by any except the military and police. The proponents of the measure frankly admit that criminals will continue to obtain concealable weapons, and that as a matter of fact the principal source of supply for criminals at this time is the pistol bootlegger and not the legitimate dealer. The naive claim is made, however, that this continued bootlegging of pistols will not be as serious as the present legitimate sale and possession of such arms under California's permit system, because the police, when they find a man with a gun on him, will have all the evidence they need to send him to jail for a year. At the present time the police have the same opportunity under the existing law, but the crime isn't a felony. The Sullivan Law has failed to stop crime in New York State, according to the California reformers, because it does not go far enough.

Continued on next page

Table S7 continued from previous page

Paper	Topic	Text
Gilardi et al. (2021)	Smoking-ban enforcement	Snellville's ordinance would not be as stringent. It would allow smoking in private offices, for example, and would not require businesses to post no-smoking signs. Council Member Bruce Garraway, an outspoken tobacco opponent, urged council members to consider adopting a tougher ban, one that prohibits smoking in private offices. The city also should require outdoor smokers to stand away from doorways so smoke will not drift inside restaurants or offices, he said. "We'll have to look at that," said Oberholtzer. "I think we'll probably have some amendments brought up." The ban would authorize Snellville's code officers and police to fine anyone who violated the ordinance. The fines range from $50 for a first-time offender to $250 for people who consistently fire up where they should not. Snellville's proposals indicate that smoking is under fire as never before. The city would not be the first Georgia municipality to clamp down on indoor smoking; Grayson, Loganville and Bainbridge have smoking bans in effect, according to Georgians Against Smoking Pollution.
Jung (2020)	Moral rhetoric	But you don't need a safety net unless you've turned the health system into a highwire act and families are in danger of falling off.
Mattingly et al. (2025)	Western dysfunction	united nations human rights chief michelle bachelet urged the global community to fight discrimination against the people of asian origin if the world is to effectively combat the covid outbreak she made the remarks during a session of the un.
Osnabrügge et al. (2021)	Emotive rhetoric	May I thank the hon. Lady who has run such an effective campaign on this and the colleagues across the House who have written about this matter to my right hon. Friend the Home Secretary. As she knows, the previous Home Secretary, in her capacity as both Home Secretary and Minister for Women and Equalities, took this subject extremely seriously, as does the new Home Secretary. We are drawing together the evidence and looking at it very carefully, and we will of course let the House know the results of that review as soon as we can.
Osnabrügge et al. (2021)	Emotive rhetoric	My hon Friend has a distinguished record of more than four years of campaigning hard for local health care services in Redditch and her constituents should be proud of what she has done on their behalf fighting for Redditch hospital and local services I shall be delighted to meet her to talk further about the local challenges for maternity care.
Parthasarathy et al. (2019)	fund allocation	A meeting was held by the panchayat in which 13 persons were sanctioned OAP under the scheme. But, it has actually come only to just 2 persons and the remaining 11 are yet to get. It is 6 months and no action has been taken though they have been given orders. The money has not come so far. Help should be done to get them the money.
Parthasarathy et al. (2019)	fund allocation	That is what if we install a new pipe it will be okay. Next repairing of road, building public toilet. For that government is giving 11600. We have been saying nobody is coming forward.
Schub (2022)	political content (vs. military)	We have also considered diplomatic steps which we might take to stabilize the Angola-Zaire border situation. Our concern is that military assistance from French, Belgian, and Moroccan sources being funneled through Zaire to UNITA ma prompt Neto to encourage a resumption of Katangan gendarme attacks aimed at Zaire. My conclusion is that we cannot dissuade the French and Belgians from their view that their long-term interests are served by an ultimate Savimbi victory. Accordingly, we will limit ourselves to again warning Mobutu against diversion of US-supplied equipment. We will also share with him our concerns that his continued interference in Angola, even as an intermediary for others, could jeopardize Congressional support for our present economic and military programs and provoke the Angolans and Katangans into stepping up their activities in Shaba.
Schub (2022)	political content (vs. military	Has doubts about the Communists in charge'97CIA has no doubts. Rebels are not all of the same stripe. With [American] troops in the country it is difficult to talk with the rebels.

S3 Additional results for Replication Studies

S3.1 Additional results for LLM annotations against expert annotations

S3.1.1 Additional results under alternative measures of intercoder agreement

As a robustness check, we re-evaluate study-level expert intercoder reliability using Krippendorff's α, rather than relying exclusively on mean pairwise agreement. Mean pairwise agreement is the average proportion of texts on which each pair of expert coders assigns the same label. Although intuitive, this measure does not adjust for agreement that could arise by chance, especially when one category is substantially more common than the other. For nominal labels, Krippendorff's α is defined as

$$\alpha = 1 - \frac{D_o}{D_e}, \tag{S1}$$

where D_o is the observed disagreement among coders and D_e is the disagreement expected under the marginal distribution of the coded categories. Thus, $\alpha = 1$ indicates perfect reliability, $\alpha = 0$ indicates no more agreement than expected by chance, and $\alpha < 0$ indicates disagreement greater than would be expected by chance. We note that Krippendorff's α can take negative values when observed disagreement exceeds that expected by chance.

As shown in Table S8, Krippendorff's α is systematically lower than mean pairwise agreement because it adjusts for agreement expected by chance, yet it broadly preserves the same cross-study ordering. Studies with higher raw agreement generally also exhibit higher α, whereas those with lower agreement tend to show weaker chance-adjusted reliability. This consistency indicates that the agreement-based ranking captures meaningful differences in coding ambiguity, even though α provides a more conservative measure of reliability.

Table S8: The 14 replicated studies: concept of interest and expert intercoder reliability (ICR) measured by (i) the mean pairwise agreement among the three trained author coders and (ii) Krippendorff's α on each study's full 100-text sample.

Paper	Concept of interest	Mean pairwise agreement	Krippendorff's α
Feltovich and Giovannoni (2024)	Negative campaigning	0.987	0.869
Bush and Clayton (2023)	Taxes	0.933	0.745
Thrall (2025)	Commercial issues	0.900	0.372
Mattingly et al. (2025)	Western dysfunction	0.887	0.392
Clark and Dolan (2021)	Fiscal policy	0.880	0.734
Arias (2022)	Climate Security	0.873	0.431
Blumenau and Lauderdale (2018)	Finance	0.860	0.639
Parthasarathy et al. (2019)	Fund allocation	0.853	0.382
Jung (2020)	Moral rhetoric	0.847	0.460
Blaydes et al. (2018)	Art of rulership	0.753	0.363
Lacombe (2019)	Gun control	0.720	0.411
Gilardi et al. (2021)	Smoking ban enforcement	0.700	0.224
Schub (2022)	Adversary domestic politics	0.640	0.266
Osnabrügge et al. (2021)	Emotive rhetoric	0.533	-0.022

Figure S2 evaluates whether model–expert reliability varies with the ambiguity of the underlying coding task. On papers classified as easy, most models achieve average pairwise alphas between approximately 0.45 and 0.60. The majority-vote ensemble produces the highest point estimate, approximately 0.62, while several individual models approach or slightly exceed the mean expert–expert benchmark. On hard papers,

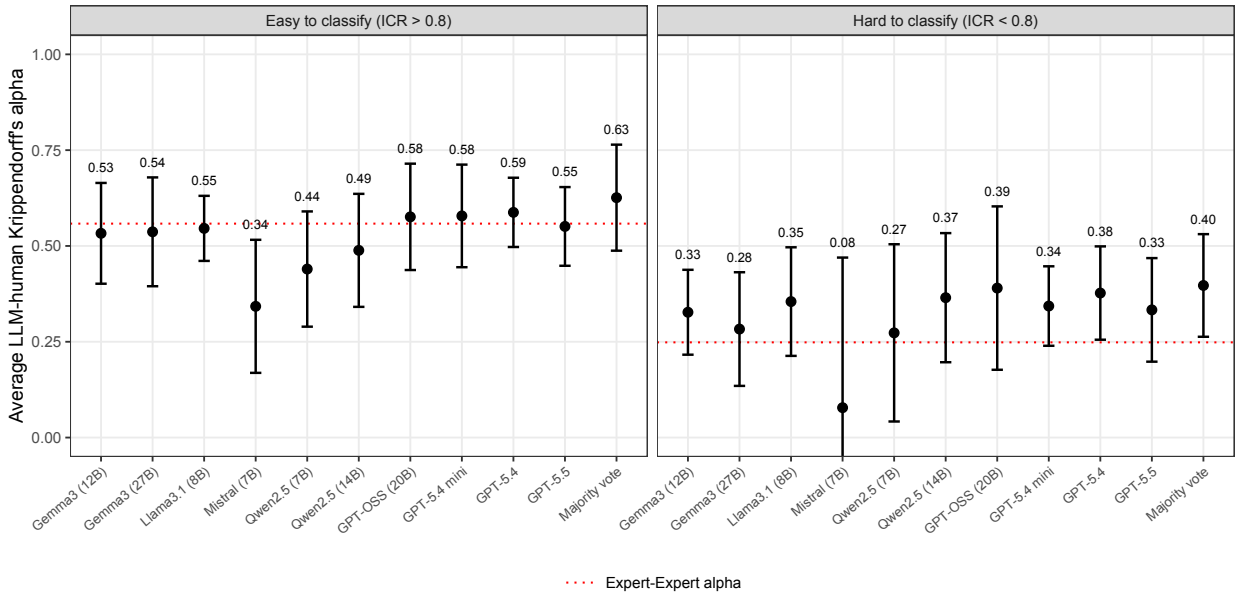

Figure S2: LLM–expert Krippendorff's α by coding difficulty. Points report the mean pairwise Krippendorff's α between each model and the individual expert coders, averaged first within papers and then across papers. Error bars show 95% confidence intervals calculated across papers. The red dotted line reports mean pairwise expert–expert Krippendorff's α within each difficulty group.

model–expert alpha generally declines, with most estimates falling between approximately 0.25 and 0.40. Mistral is a notable exception, producing an estimate close to zero in the hard-paper subset.

The principal pattern is therefore a substantial reduction in chance-adjusted model–expert reliability when expert coders themselves find the classification task more ambiguous. Confidence intervals are also wider among hard papers, indicating greater between-paper heterogeneity and making fine-grained model rankings uncertain. Estimates above the red expert benchmark should not be interpreted as evidence that a model is more accurate than experts against an objective ground truth. Rather, they indicate that the model agrees with individual expert coders at least as consistently as expert coders agree with one another under the same pairwise-alpha metric.

S3.1.2 Within-Model Annotation Stability under Prompt Variation

Our main replication analysis holds the prompt fixed, using a single structure built from six components (role, topic definition, classification rules, label wording, paper and data description, and a reasoning cue; see Table S2). This lets us compare models and studies on common ground, but raises a concern: we do not know how much our conclusions reflect this particular prompt rather than the task itself. To probe this, we toggle each component on or off to generate all $2^6 = 64$ prompts per paper and run them through seven open-source LLMs (Mistral 7B, LLaMa 3.1-8B, Qwen2.5-7B, Qwen2.5-14B, Gemma 3-12B, Gemma3-27B, GPT-OSS-20B). We test whether this sensitivity reflects *substantive ambiguity* rather than arbitrary instability, by asking whether prompts disagree most on the texts that human experts also find hard to code. The fully specified prompt, with all six components included, is our reference point and we refer to it as prompt 64.

The analysis uses all the 64 prompts applied to texts by each model. We measure within-model prompt disagreement using Gini impurity, which equals zero when all 64 prompts produce the same classification and reaches its maximum of 0.5 when the prompts are evenly divided between labels 0 and 1. We then

compare this prompt disagreement between texts on which all three expert coders agree and texts on which at least one expert assigns a different label. Because the overall level of coding difficulty varies across studies, we report the comparison separately for easy studies, defined by mean pairwise expert agreement above 0.8, and hard studies, defined by mean pairwise agreement at or below 0.8. For each model and difficulty group, the table reports the mean Gini impurity for expert-agreement texts and expert-disagreement texts, their difference, and the p-value from a two-sample comparison of the corresponding text-level Gini values.

Table S9 shows that within-model prompt disagreement is generally higher for texts on which experts disagree. This pattern is statistically significant for Gemma3 (12B), Llama3.1 (8B), Qwen2.5 (7B), Qwen2.5 (14B), and GPT-OSS (20B) in both easy and hard studies, and for Gemma3 (27B) in easy studies. The largest easy-study increases are 0.120 for Gemma3 (27B) and 0.119 for GPT-OSS (20B); for GPT-OSS (20B), the increase is 0.096 in hard studies. Mistral (7B) is the principal exception: prompt disagreement is slightly higher on expert-disagreement texts in easy studies (difference 0.016, $p = 0.342$) but lower in hard studies (difference -0.036, $p = 0.042$). Overall, the results indicate that prompt sensitivity often tracks expert-defined ambiguity, although the strength and direction of this relationship vary across models.

Table S9: Within-model prompt disagreement by model, expert agreement, and study difficulty. For each model and text, Gini impurity is calculated across the 64 prompt-specific predictions. Higher values indicate greater sensitivity to prompt specification.

Model	Study difficulty	Experts agree	Experts disagree	Difference	p-value
Gemma3 (12B)	Easy (> 0.8)	0.060	0.136	0.076	4.47×10^{-7}
Gemma3 (12B)	Hard (≤ 0.8)	0.150	0.200	0.049	2.72×10^{-3}
Gemma3 (27B)	Easy (> 0.8)	0.048	0.168	0.120	3.39×10^{-11}
Gemma3 (27B)	Hard (≤ 0.8)	0.112	0.123	0.011	4.44×10^{-1}
Llama3.1 (8B)	Easy (> 0.8)	0.066	0.163	0.097	3.81×10^{-9}
Llama3.1 (8B)	Hard (≤ 0.8)	0.182	0.248	0.066	1.94×10^{-4}
Mistral (7B)	Easy (> 0.8)	0.199	0.215	0.016	3.42×10^{-1}
Mistral (7B)	Hard (≤ 0.8)	0.202	0.166	-0.036	4.23×10^{-2}
Qwen2.5 (7B)	Easy (> 0.8)	0.039	0.126	0.087	7.14×10^{-9}
Qwen2.5 (7B)	Hard (≤ 0.8)	0.100	0.150	0.050	1.17×10^{-3}
Qwen2.5 (14B)	Easy (> 0.8)	0.028	0.115	0.087	1.58×10^{-9}
Qwen2.5 (14B)	Hard (≤ 0.8)	0.068	0.122	0.054	1.26×10^{-4}
GPT-OSS (20B)	Easy (> 0.8)	0.033	0.152	0.119	4.14×10^{-13}
GPT-OSS (20B)	Hard (≤ 0.8)	0.090	0.186	0.096	1.59×10^{-9}

To identify which prompt-design choices improve zero-shot classification, we exploit the full set of prompt variants applied across papers and models and estimate the marginal association between individual prompt components and LLM–expert agreement. The dependent variable is a model prediction's mean agreement with the three expert labels for a given text. The explanatory variables indicate whether the prompt includes each design component—such as an explicit expert role, a formal concept definition, classification rules, standardized label wording, step-by-step reasoning, or paper/data context. We also control for log text length and include paper and model fixed effects so that the coefficients are identified from differences across prompt specifications within the same studies and model families.

An OLS regression (Figure S3) analyzing the effect of prompt features on mean LLM–expert agreement reveals that structural instructions provide the largest gains, led by explicit *Classification rules* ($\hat{\beta} \approx +0.031$), standardized *Label wording* ($\hat{\beta} \approx +0.016$), and specifying an *Expert role* ($\hat{\beta} \approx +0.013$). Step-by-step *reasoning* and topic *definition* each have estimates of approximately $+0.006$, while incorporating *paper/data descriptions* has an estimate near zero ($\hat{\beta} \approx +0.003$). Additionally, document length has a negative associa-

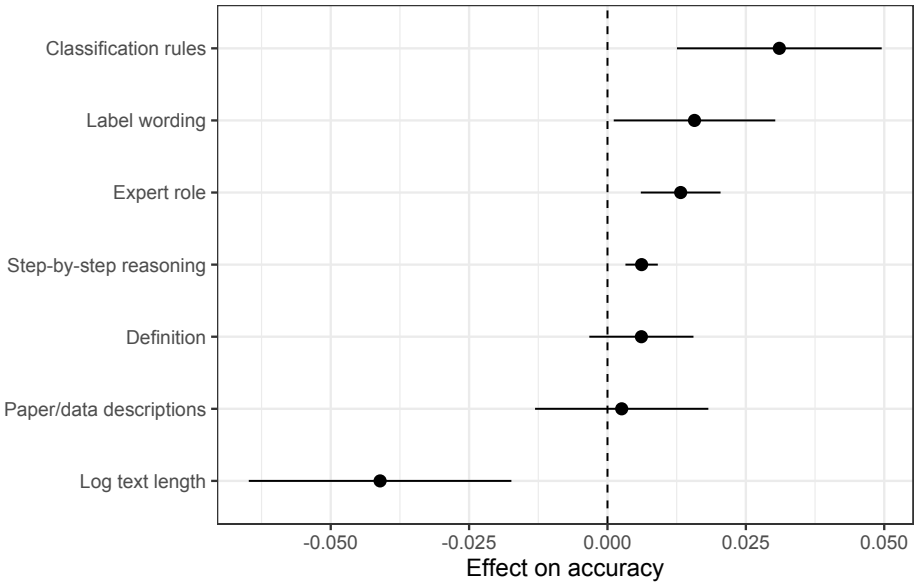

Figure S3: Marginal effects of prompt engineering components and text length on mean LLM–expert agreement (OLS coefficient estimates with paper-clustered 95% confidence intervals).

tion with agreement ($\hat{\beta} \approx -0.041$). The plotted 95% confidence intervals use CR1 standard errors clustered over the 14 papers.

S3.1.3 Within-Model Annotation Stability under Stochastic Decoding

Our main analysis identifies ambiguous texts using disagreement across different LLMs. However, most applied researchers often use a single model rather than a panel of models. We therefore examine whether repeated runs of the same model provide a useful within-model measure of annotation uncertainty. For each of the seven open-source LLMs, we classify the same 100 texts from each of the 14 studies ten times under stochastic decoding with temperature 1.0. For each model–text pair, we define instability as an indicator equal to one if the predicted label changes at least once across the ten runs. We then summarize the prevalence of instability by study and model and compare it between texts on which the experts agree and texts on which they disagree.

Figures S4 and S5 show that within-model instability is generally greater in the more difficult studies, but its magnitude varies substantially across models. LLaMA 3.1-8B and GPT-OSS-20B exhibit the greatest instability overall, followed by Mistral-7B, whereas the Gemma 3 and Qwen2.5 models produce substantially more stable annotations. Table S10 further shows that instability is concentrated on texts that divide experts for GPT-OSS-20B (0.406 versus 0.112), LLaMA 3.1-8B (0.308 versus 0.131), and Qwen2.5-14B (0.060 versus 0.010). All three differences are statistically significant. By contrast, the corresponding differences are small and statistically indistinguishable from zero for Mistral-7B, both Gemma 3 models, and Qwen2.5-7B. These results provide a partial support for stochastic-decoding instability as a within-model proxy for ambiguity: it identifies texts that divide experts for some models, but its informativeness depends substantially on the model.

S3.1.4 Additional results under an alternative measure of within-text disagreement

In Section 3.2.2 and S3.1.2, we used Gini impurity to measure the text-level annotation heterogeneity. As a robustness, we replicate the same results using the normalized Shannon entropy, which is defined as

$$H = -p \log_2(p) - (1-p) \log_2(1-p),$$

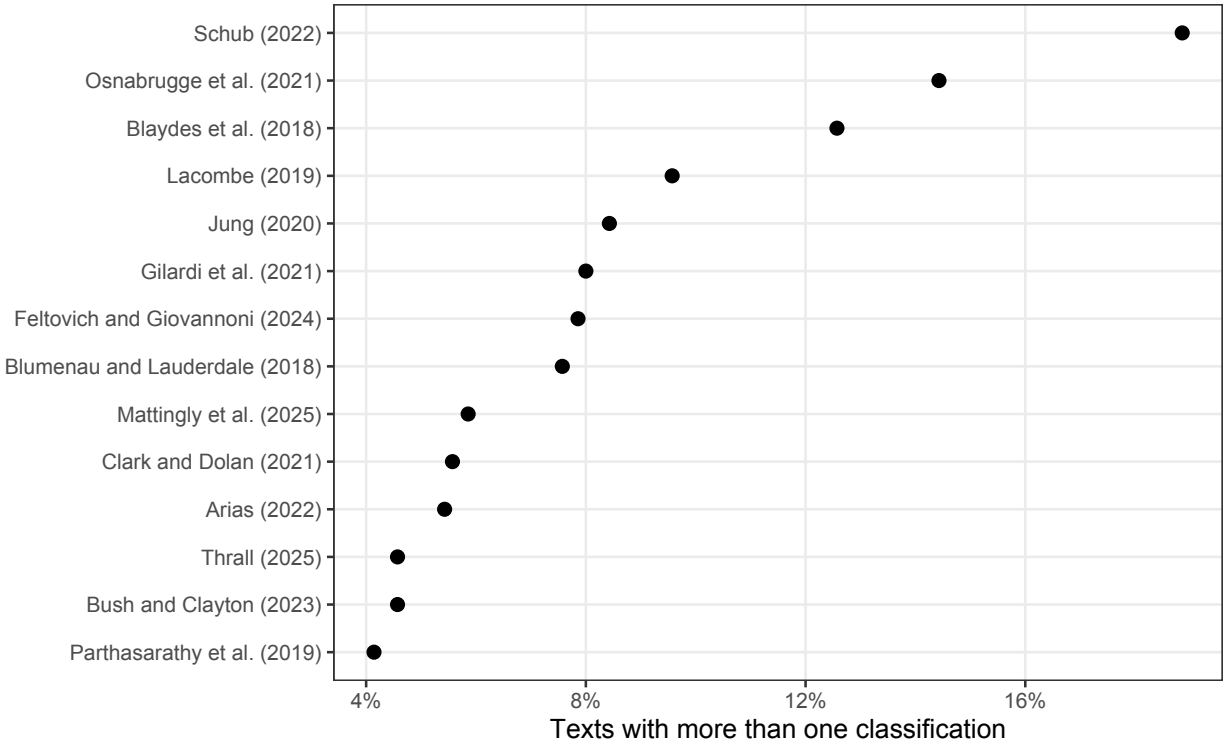

Figure S4: Instability of Text Annotations across seven open-source models for each paper. Instability is measured as the proportion of 100 randomly sampled texts that receive more than one classification label across 10 repeated runs under stochastic decoding (temperature = 1.0). The seven open-source models are GPT-OSS (20B), Gemma 3 (12B and 27B), Llama 3.1 (8B), Mistral (7B), and Qwen 2.5 (7B and 14B).

Table S10: Annotation Variation for each model under Stochastic Decoding (Temperature = 1) Across 10 Runs, Stratified by Expert Agreement.

Model	Experts agree	Experts disagree	Difference	p-value
GPT-OSS (20B)	0.112	0.406	0.293	<0.001
Llama3.1 (8B)	0.131	0.308	0.177	0.003
Qwen2.5 (14B)	0.010	0.060	0.050	0.002
Mistral (7B)	0.119	0.150	0.031	0.436
Gemma3 (12B)	0.021	0.033	0.012	0.257
Gemma3 (27B)	0.009	0.020	0.011	0.142
Qwen2.5 (7B)	0.020	0.019	-0.001	0.934

where p denote the proportion of LLMs assigning the positive label. In this binary classification setting, H ranges from 0 to 1: $H = 0$ indicates complete agreement across models, whereas $H = 1$ indicates a perfectly even split between positive and negative predictions. We calculate entropy separately for each text across the available LLM predictions and then compare the average entropy of texts on which human coders agree with that of texts on which human coders disagree. Higher average entropy therefore indicates that the models are more divided in their interpretation of the same texts.

Table S11 reports the same analysis as in Table 3 using Shannon entropy. This shows that cross-model disagreement is substantially greater on texts where experts disagree. Among studies classified as easy, mean

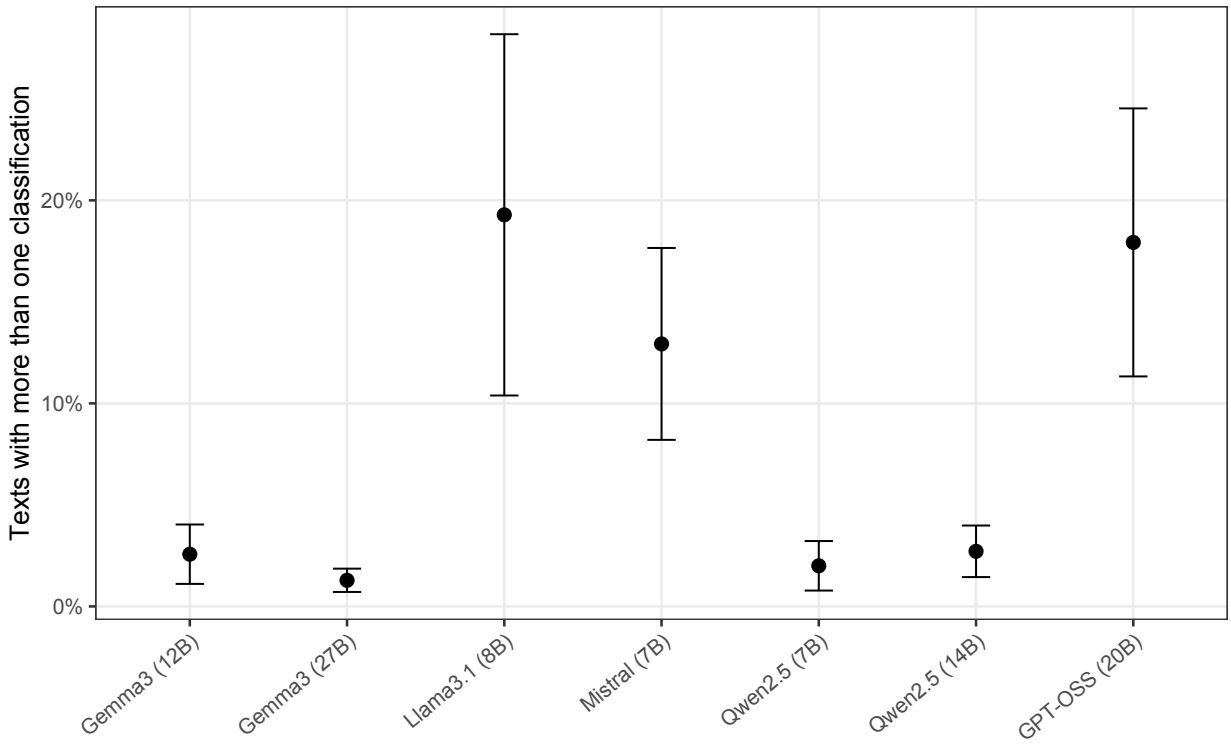

Figure S5: Instability of Text Annotations for each open-source model across 14 papers. Instability is measured as the proportion of 100 randomly sampled texts that receive more than one classification label across 10 repeated runs under stochastic decoding (temperature = 1.0).

entropy rises from 0.156 when experts agree to 0.525 when they disagree, a difference of 0.369 ($p < 0.001$). The same pattern appears among hard studies, where entropy increases from 0.323 to 0.559, a difference of 0.237 ($p < 0.001$). This suggests that LLMs are especially divided on texts that are also difficult for human experts to classify.

Table S11: Cross-model disagreement by expert agreement and study difficulty using entropy. Higher values indicate greater disagreement among the 10 LLMs.

Study difficulty	Experts agree	Experts disagree	Difference	p-value	N agree	N disagree
Easy (ICR > 0.8)	0.156	0.525	0.369	$< 2.2 \times 10^{-16}$	753	147
Hard (ICR ≤ 0.8)	0.323	0.559	0.237	4.69×10^{-13}	252	248

On the other hand, Table S12 replicates the results of within-model prompt disagreement in Table S9 using Shannon's entropy. This shows that, for most models, within-model prompt disagreement is substantially higher on texts where experts disagree than on texts where experts agree. This pattern is statistically significant for Gemma3 (12B), Llama3.1 (8B), both Qwen2.5 models, and GPT-OSS (20B) in both easy and hard studies, and for Gemma3 (27B) in easy studies. The largest easy-study differences are 0.267 for GPT-OSS (20B) and 0.253 for Gemma3 (27B). Mistral is the main exception: its prompt disagreement is slightly higher on expert-disagreement texts in easy studies but lower in hard studies, with the difference reaching statistical significance only among hard studies. Overall, the results indicate that prompt sensitivity generally concentrates on texts that are also difficult for experts to classify, although this relationship varies

across model families.

Table S12: Within-model prompt disagreement by model, expert agreement, and study difficulty. For each model and text, normalized Shannon entropy is calculated across the 64 prompt-specific predictions. Higher values indicate greater sensitivity to prompt specification.

Model	Study difficulty	Experts agree	Experts disagree	Difference	p-value
Gemma3 (12B)	Easy (> 0.8)	0.140	0.314	0.175	3.85×10^{-8}
Gemma3 (12B)	Hard (≤ 0.8)	0.332	0.448	0.115	7.29×10^{-4}
Gemma3 (27B)	Easy (> 0.8)	0.113	0.366	0.253	1.24×10^{-11}
Gemma3 (27B)	Hard (≤ 0.8)	0.256	0.280	0.025	4.34×10^{-1}
Llama3.1 (8B)	Easy (> 0.8)	0.153	0.370	0.217	2.36×10^{-10}
Llama3.1 (8B)	Hard (≤ 0.8)	0.396	0.537	0.140	1.06×10^{-4}
Mistral (7B)	Easy (> 0.8)	0.439	0.476	0.037	2.88×10^{-1}
Mistral (7B)	Hard (≤ 0.8)	0.441	0.359	-0.081	2.48×10^{-2}
Qwen2.5 (7B)	Easy (> 0.8)	0.090	0.288	0.198	7.09×10^{-10}
Qwen2.5 (7B)	Hard (≤ 0.8)	0.229	0.336	0.106	9.69×10^{-4}
Qwen2.5 (14B)	Easy (> 0.8)	0.068	0.266	0.198	1.58×10^{-10}
Qwen2.5 (14B)	Hard (≤ 0.8)	0.153	0.278	0.124	2.99×10^{-5}
GPT-OSS (20B)	Easy (> 0.8)	0.076	0.343	0.267	1.50×10^{-14}
GPT-OSS (20B)	Hard (≤ 0.8)	0.206	0.411	0.205	3.99×10^{-10}

S3.2 Additional results for Crowd Sourced Annotation

This section elaborates on the results from the crowd-sourced annotation, which affords a blinded, preregistered test of the paper's central claims. This section proceeds with reporting the results from the preregistered hypotheses for the experiment, followed by robustness checks, and alternative measures of intercoder agreement.

S3.2.1 Preregistered Hypotheses

We preregistered six hypotheses that extend the paper's theory from expert and LLM coders to a third set of coders: crowd workers. Among the six, three confirmatory hypotheses restate the paper's core predictions for the crowd.

- **H1 (text difficulty)**: When LLMs agree with one another, crowd-sourced workers are more likely to agree with each other. In other words, the Gini impurity of LLMs is correlated with that of crowd-sourced workers. This maps on to the main paper's Hypothesis 4a that disagreement reflects textual ambiguity rather than coder identity.
- **H2 (crowd vs. LLM agreement with experts)**: The average agreement of crowd-sourced workers with three experts (authors' coding) is lower than that of LLMs with three experts. In other words, the average LLM-expert agreement is significantly higher than the average crowd-expert agreement. This maps on to the main paper's Hypothesis 4b: LLMs align with experts better than crowd coders.
- **H3 (codebook improvement)**: Using an improved codebook decreases the Gini impurity among crowd-sourced workers (**H3a**) and increases the average agreement of crowd-sourced workers with experts (**H3b**). We test this on the original and improved Schub (2022) codebooks over the same 20 texts. This maps on to the main paper's Hypothesis 3: clearer coding rules reduce disagreement.

We further probe three exploratory hypotheses on coder-and-text level sources of agreement as moderators. We preregister these as exploratory hypotheses given these analyses involve slicing an already small sample size. A full copy of our preregistered hypotheses, research design, and analysis plan are available on OSF, at [LINK REDACTED].

- **H4 (coder knowledge):** We test whether crowd–expert agreement is higher among workers with higher baseline knowledge of politics/policy.(**H4a**), higher frequency of political news consumption, (**H4b**), or higher education levels (**H4c**).
- **H5 (coder attention):** We test whether crowd–expert agreement is higher on texts workers spend longer on (**H5a**) or on shorter texts (**H5b**).
- **H6 (text heterogeneity)**: Crowd Gini impurity is lower on "easy" texts than on "hard" ones, with difficulty defined by low LLM disagreement (**H6a**) or high expert inter-coder reliability (**H6b**). This maps on to the main paper's Hypothesis 4a, similar to H1 but contrasting texts across strata.

For all survey results, we use the 20-text-per-paper survey sample described in Appendix S1.3. Because this subsample oversamples contested texts by design, agreement and reliability levels on it are generally lower than in the full corpus. We conduct multiple hypothesis testing for the three confirmatory hypotheses (four tests: H1, H2, H3a, H3b) using the Benjamini–Hochberg correction. We report the summary of findings for each hypothesis in Table S13. Below, we walk through each result in turn.

Table S13: Preregistered tests: estimates, 95% confidence intervals, p-values, and support. The four confirmatory tests share a Benjamini–Hochberg correction (p_{BH}); exploratory tests are uncorrected.

Hypothesis	Test (N)	Est.	95% CI	p	p_{BH}	Support
Panel A: Confirmatory (Benjamini–Hochberg family)						
H1: Crowd tracks LLM disagreement	OLS slope (280)	0.319	[0.238, 0.400]	< 0.001	< 0.001	✓
H2: LLM closer to experts than crowd	Paired t (14)	0.086	[0.040, 0.132]	< 0.001	0.001	✓
H3a: Better codebook, less disagreement	Paired t (20)	-0.131	[-0.227, -0.034]	0.005	0.007	✓
H3b: Better codebook, closer to experts	Paired t (20)	0.055	[-0.061, 0.171]	0.167	0.167	(+) n.s.
Panel B: Exploratory tests (no correction)						
H4a: Political knowledge	OLS (165)	0.063	[-0.010, 0.137]	0.090	—	(+) n.s.
H4b: News consumption	OLS (165)	0.063	[0.009, 0.117]	0.022	—	(+)†
H4c: Education	OLS (165)	0.051	[0.005, 0.096]	0.030	—	(+)†
H5a: Total duration (per 10 min)	OLS (165)	-0.026	[-0.0538, 0.0011]	0.060	—	($-$)‡
H5b: Text length (per 100 words)	OLS (280)	-0.010	[-0.021, 0.001]	0.067	—	($-$) n.s.
H6a: LLM-contested − uncontested Gini	Welch t (280)	0.137	[0.100, 0.173]	< 0.001	—	✓
H6b: Expert-split − expert-agree Gini	Welch t (280)	0.101	[0.064, 0.138]	< 0.001	—	✓

Notes: ✓ supported; (+)/($-$), signed as predicted but not significant; n.s., not significant. † Significant only without task fixed effects. ‡ Sign opposite to the preregistered prediction but largely null.

S3.2.2 Survey Finding 1: Textual ambiguity drives crowd disagreement (H1 and H6).

First, to test H1, we estimate an OLS regression of each text's crowd Gini impurity on its LLM Gini impurity across all 280 texts. We exclude the improved Schub codebook for comparability. Here, a positive slope with 95% confidence interval excluding zero means that crowd disagreement correlates positively with LLM disagreement. We use three different specifications: naive OLS, including cluster-robust errors by paper, and including both paper fixed effects and cluster-robust errors.

Table S14 reports the results. We find that crowd disagreement concentrates on exactly the texts that experts and LLMs also find hard. We observe a positive and statistically significant slope: each text's crowd Gini impurity has a positive slope when regressed on LLM Gini impurity, with a coefficient of 0.32 (95% CI $[0.24, 0.40]$) for the naive OLS. This finding is consistent across specifications with similar estimates, which shows that it is not an artifact of which papers happen to contain more difficult texts.

Table S14: H1, Regression of crowd disagreement on LLM-panel disagreement

	(1) OLS	(2) CR2 by paper	(3) Paper FE + CR2
LLM Gini	0.319	0.319	0.315
	[0.238, 0.400]	[0.170, 0.468]	[0.165, 0.465]
p	< 0.001	< 0.001	< 0.001
Paper fixed effects	No	No	Yes
N (texts)	280	280	280
N (papers)	14	14	14

We probe this finding in a different way in H6 by splitting texts into "easy" versus "hard" strata, based on LLM and expert disagreement. This follows the same procedure as the main paper's analyses (see Section 3.2.2). To briefly reiterate: in H6a, disagreement among LLMs is an even split, incorporated by design given that we stratify-sample texts based on the 10 highest and lowest Gini impurity texts for each paper (averaged across the 10 LLM models). In H6b, disagreement among experts is considered "high" if experts are split in their coding and "low" for unanimous coding. We then compare mean crowd Gini impurity between

the easy and hard strata with a Welch t-test. We find that crowd disagreement rises from 0.24 on the least LLM-contested texts to 0.38 on the most contested, and from 0.27 where the experts agree to 0.37 where they split (Table S15). In all, from both H1 and H6, we demonstrate that crowd disagreement tracks the same underlying textual ambiguity that divides LLMs or experts.

Table S15: H6, Crowd disagreement by LLM sampling stratum and by expert unanimity.

	n texts	Mean crowd Gini (95% CI)
Panel A: By LLM sampling stratum		
Least contested (LLM agree)	140	0.242 [0.213, 0.271]
Most contested (LLM disagree)	140	0.379 [0.357, 0.402]
Difference (most − least)		+0.137 [0.100, 0.173]
Panel B: By expert unanimity		
Experts unanimous	172	0.272 [0.245, 0.298]
Experts split	108	0.373 [0.347, 0.398]
Difference (split − unanimous)		+0.101 [0.064, 0.138]

S3.2.3 Survey Finding 2: Expert-LLM agreement beats expert-crowd agreement (H2)

Second, to test our survey's H2, we use a one-sided paired t-test across the 14 papers, pairing each paper's mean LLM–expert agreement against its mean crowd–expert agreement. Here, we exclude the improved codebook for Schub for comparability. Here, we find a consistent and substantial gap between crowd–expert and LLM–expert agreement, regardless of LLM models used and for most papers.

Table S16 reports the results. We find that the LLM panel is closer to the experts than the crowd in 12 of the 14 papers, with one tie and a single, slight reversal. More surprisingly, this gap holds regardless of LLM models or task difficulty – even the oldest models and worst-performing ones have a closer average agreement with experts than crowd workers (Figure S6). This confirms the prediction in H2: although crowd workers track ambiguity in the same way that LLMs and experts do, crowd workers consistently show lower agreement with experts than LLMs. Our results corroborate existing works that find LLMs outperforming crowd workers on a range of text-annotation or related tasks (Gilardi et al., 2023; Ziems et al., 2024), and extend them by showing how "outperforming" can be measured by agreement without defining a ground truth. We further show how the LLM-crowd-worker gap applies across a wide set of papers, across task and text difficulty, and LLM model types.

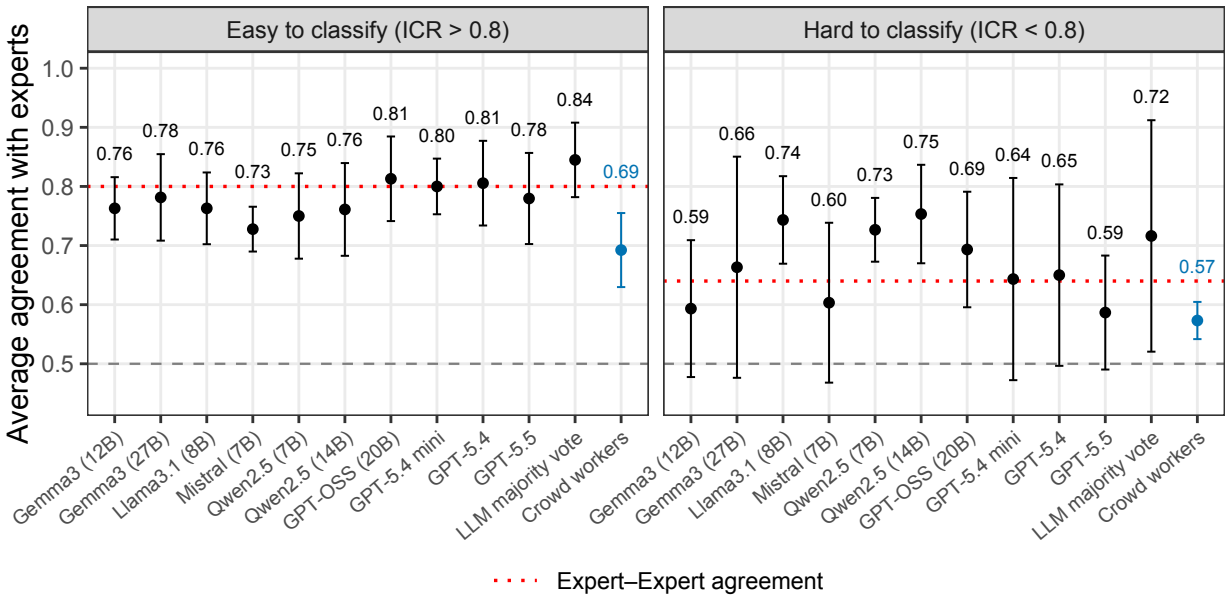

Figure S6: Agreement with expert coders on the 20-text survey sample, for each LLM, their majority vote, and the crowd workers (in blue). Points are mean pairwise agreement with the three experts; the red dotted line is the mean expert–expert agreement. Levels are lower than in the full corpus because the survey sample oversamples contested texts.

Table S16: H2, LLM–expert and crowd–expert agreement by paper

Paper	LLM–Expert	Crowd–Expert	Gap (LLM − Crowd)
Feltovich and Giovannoni (2024)	0.802	0.747	+0.055
Bush and Clayton (2023)	0.822	0.703	+0.118
Thrall (2025)	0.865	0.796	+0.069
Mattingly et al. (2025)	0.705	0.770	−0.065
Clark and Dolan (2021)	0.765	0.678	+0.087
Arias (2022)	0.760	0.538	+0.222
Blumenau and Lauderdale (2018)	0.720	0.720	0.000
Parthasarathy et al. (2019)	0.785	0.600	+0.185
Jung (2020)	0.747	0.679	+0.068
Blaydes et al. (2018)	0.670	0.560	+0.110
Lacombe (2019)	0.548	0.542	+0.006
Gilardi et al. (2021)	0.768	0.565	+0.203
Schub (2022), [original codebook]	0.695	0.603	+0.092
Osnabrügge et al. (2021)	0.647	0.595	+0.052
Mean (14 papers)	0.736	0.650	+0.086

S3.2.4 Survey Finding 3: A clearer codebook improves crowd coding (H3)

Third, to test H3, we conduct a one-sided paired t-test across the 20 texts from Schub, assessing changes from the original codebook to the improved codebook, which includes clearer classification rules and examples. The two Schub codebooks are randomly assigned as separate tasks in our survey, while respondents code the same 20 texts regardless of which codebook they receive.

Figure S7 visualizes the results (see Table S17 for the exact values). We find that sharpening the code-

book improves agreement and reduces disagreement among crowd workers, consistent with our findings for LLMs and experts. For H3a, Gini impurity falls from 0.42 to 0.29, indicating that the improved codebook reduces disagreement. For H3b, mean crowd–expert agreement rises from 0.60 to 0.66 under the improved codebook. Both findings are consistent with our preregistered expectations, although only the result for H3a is statistically significant at the 5% level. These findings are robust to alternative metrics, such as using Krippendorff's α instead of pairwise agreement. Thus, the improved codebook leads crowd workers to agree more with one another and with experts, supporting the paper's central conclusion that clearer coding rules can reduce ambiguity and disagreement.

Table S17: H3, Effects of codebook improvement for Schub (2022)

	Original codebook (v1)	Improved codebook (v2)	Difference	p
Crowd Gini impurity (H3a)	0.418	0.288	−0.131	0.005
Crowd–expert agreement (H3b)	0.603	0.658	+0.055	0.167
Krippendorff's α (crowd)	0.047	0.355	+0.308	—
Crowd workers (n)	10	12		

Notes: H3a and H3b are the preregistered tests; Krippendorff's α is descriptive and serves as an alternative measure to agreement.

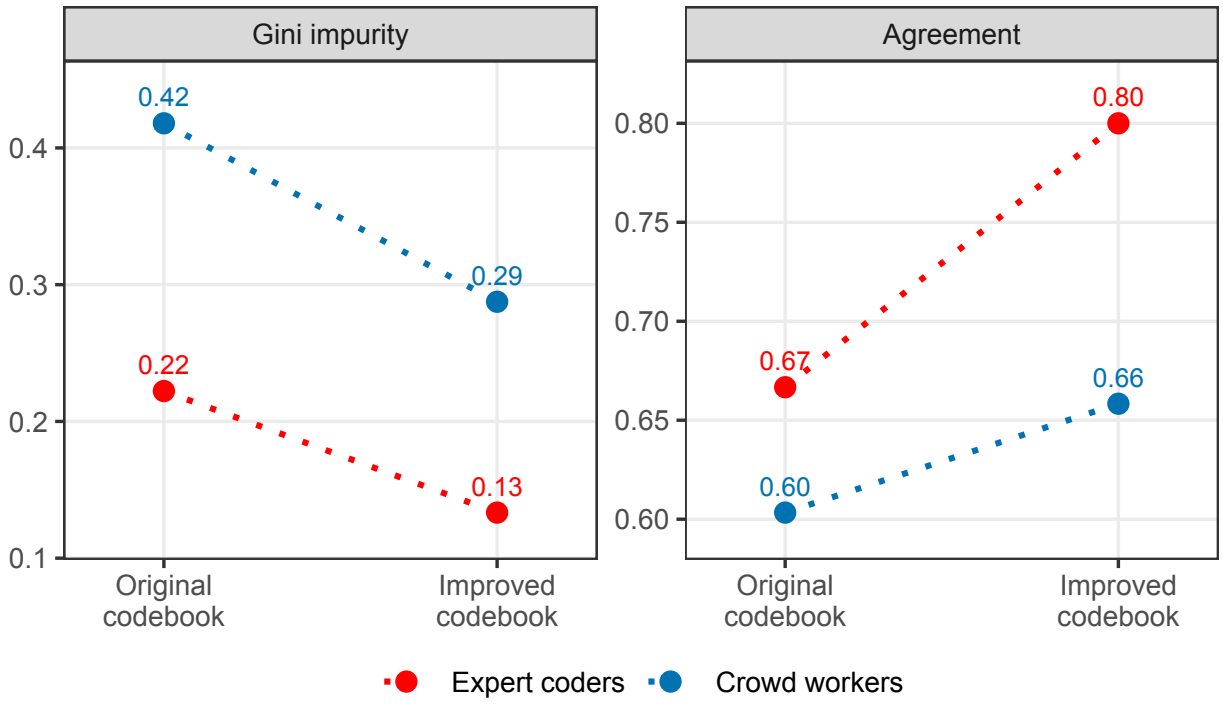

Figure S7: (Left:) Gini impurity, measure of within-source disagreement; fall in Gini corresponds to H3a (Better codebook, less disagreement). (Right:) agreement of crowd workers with experts and experts with one another; directionally consistent with H3b (Better codebook, closer to experts).

S3.2.5 Survey Finding 4: Coder characteristics do not predict annotation quality (H4/H5).

Finally, we test whether coder characteristics predict annotation quality (H4 and H5). We estimate H4 (coder knowledge as moderator) by respondent-level regressions of each worker's mean crowd–expert agreement on their political knowledge, news frequency, and education, each binarized given the modest sample size. For hypothesis 5 (coder attention as moderator), we regress each worker's mean crowd–expert agreement on their total survey duration (H5a) and per-text crowd–expert agreement on text length (H5b). Here, H5a uses total survey time rather than the preregistered per-text time. This is because we could not recover the per-text time spent since we programmed our survey to randomize text order. We run all regressions with two specifications: a univariate OLS with only the predictor of interest in each respective hypothesis and a joint specification with all moderators and task fixed effects.

Figure S8 plots the univariate OLS results while Table S18 reports both specifications in full detail. At first glance, political knowledge, news consumption, and higher education appear positively correlated with crowd–expert agreement while survey duration appears negatively correlated. However, these results are not robust. While we find directional support for H4a,b and c, these correlations are not only marginal but are not robust when we include task fixed effects. Every moderator becomes indistinguishable from zero under the fixed effect models (see Column 2 in Table S18). Neither political knowledge, news consumption, education, nor survey duration reliably separates better crowd coders from worse ones. This null is consistent with the paper's central claim that annotation quality is primarily a property of the text and the coding rule, more than of the coder. Overall, we find little evidence that *who* the coder is or their attention predicts how well they agree with the experts, contrary to our preregistered expectations.

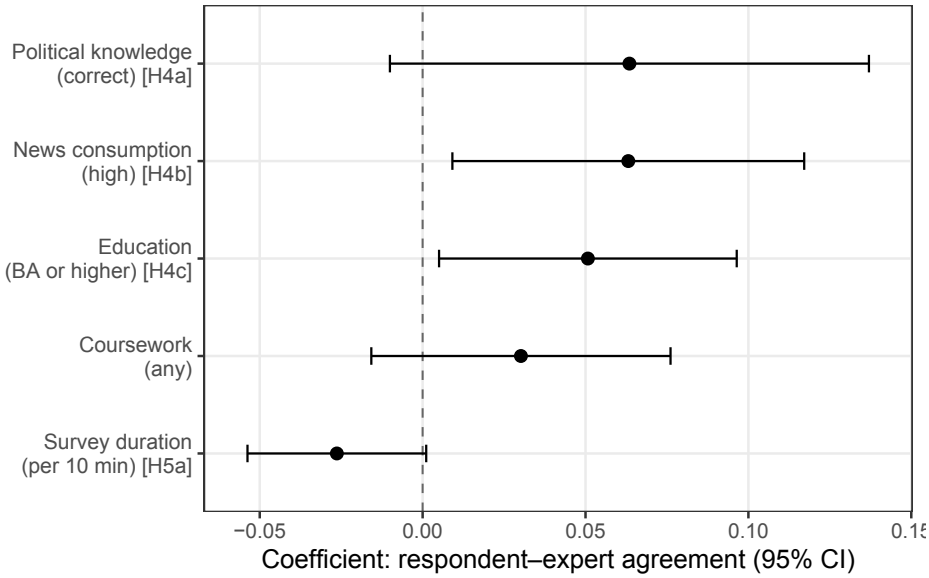

Figure S8: H4/H5, Preregistered univariate associations between coder characteristics and crowd–expert agreement. Every association is indistinguishable from zero once task fixed effects are added.

Table S18: Exploratory moderators of crowd coding quality (H4/H5)

	(1) Univariate (preregistered)			(2) Joint, task fixed effects		
Moderator	Est.	SE	p	Est.	SE	p
Panel A: Respondent's agreement with experts ($N = 165$ respondents)						
H4a: Political knowledge (correct)	0.063	0.037	0.090	0.020	0.044	0.659
H4b: News consumption (high)	0.063	0.027	0.022	0.018	0.020	0.387
H4c: Education (BA or higher)	0.051	0.023	0.030	0.033	0.037	0.382
Coursework (any)	0.030	0.023	0.196	−0.018	0.026	0.503
H5a: Survey duration (per 10 min)	−0.026	0.014	0.060	0.003	0.009	0.755
Panel B: Per-text crowd–expert agreement ($N = 280$ texts)						
H5b: Text length (per 100 words)	−0.010	0.006	0.067	−0.003	0.003	0.376
Task fixed effects		No			Yes	
Standard errors		Classical			CR2, clustered by task	

Notes: Exploratory tests (H4/H5 were preregistered as exploratory; no multiplicity correction); all p-values two-sided. Panel A's dependent variable is each respondent's mean agreement with the three expert coders over their 20 texts; Panel B's is the per-text mean crowd–expert agreement for the 280 texts of the 14 papers (Schub v2 arm excluded). Column (1) follows the preregistered specification: plain OLS, one moderator at a time, classical standard errors. In column (2) the five Panel A moderators enter jointly with task fixed effects and CR2 standard errors clustered by task (15 tasks; Satterthwaite degrees of freedom); the Panel B regression adds task fixed effects and CR2 clustering by task (14 tasks) to the univariate text-length specification. Moderators: political knowledge = correct answer on the manifesto knowledge check; high news consumption = follows news daily or a few times a week; education = bachelor's degree or higher; coursework = any response other than "No" to the related-coursework item; duration = total survey duration. A joint test of the five Panel A moderators gives $F(5, 14) = 0.30$, $p = 0.907$.

S3.2.6 Robustness to Quality Checks and Exclusions

We conducted a series of robustness checks to address the threat of AI use in survey research (Westwood, 2025). In our preregistration and Section S1.3, we applied a set of inclusion/exclusion criteria to screen for our final sample of 165: (1) we sample only from respondents who indicated English as a native language on Connect, given the difficulty of the task; (2) we automatically screen out respondents that fail any of our attention or bot checks; (3) we sampled only from respondents with an approval rate of 95% or more (Peer et al., 2014; Kennedy et al., 2020); (4) We exclude respondents that exceed a maximum time spent of 48 minutes (Connect's default maximum time) and drop any duplicate responses that are identical to others. Ultimately, we removed one respondent that exceeded the maximum time and found no respondents with duplicate responses across all questions; the other exclusions are automatic screens on Connect and Qualtrics.

Besides the above preregistered exclusions, we conducted robustness checks with two additional filters: we dropped respondents that straight-line all text classification responses (answered all 20 texts as "Yes" or all as "No") and those that score below 0.5 on Qualtrics' CAPTCHA scores. Both checks are soft exclusions given that they do not lead to a respondent being screened out based on our preregistration but they nonetheless raise data quality concerns and we thus remove them sequentially and rerun our main analyses without those responses.

Table S19 reports every headline result across the various exclusion specifications, from the preregistered sample ($N = 165$) to a strict sample that drops all flagged respondents ($N = 156$). We find that every conclusion holds across exclusion specifications with largely the same coefficients even if we exclude all potential low-quality responses: the H1 slope (crowd disagreement rising with LLM disagreement) stays between 0.319 and 0.326, the H2 gap (LLMs agreeing with experts more than the crowd) between 0.083

and 0.086, the H3a codebook effect (lower crowd Gini under the improved codebook) between -0.117 and -0.131, while H3b (higher crowd–expert agreement under that codebook) remains non-significant throughout.

Table S19: Headline estimates under three respondent-exclusion specifications (p-values in parentheses).

Specification	N	H1 slope	H2 gap	H3a ΔGini	H3b Δagr.	H6a gap	Crowd–expert agreement
Main: preregistered exclusions	165	0.319 (< 0.001)	0.086 (< 0.001)	-0.131 (0.005)	0.055 (0.167)	0.137 (< 0.001)	0.650
Drop flagged straightliners	162	0.326 (< 0.001)	0.083 (< 0.001)	-0.131 (0.005)	0.055 (0.167)	0.141 (< 0.001)	0.653
+ low reCAPTCHA score	156	0.323 (< 0.001)	0.084 (< 0.001)	-0.117 (0.011)	0.060 (0.142)	0.138 (< 0.001)	0.652

Notes: The final row additionally drops respondents with a Qualtrics reCAPTCHA bot-score below 0.5. It is retained from the authors' prior run because the de-identified package does not contain the reCAPTCHA flag needed to regenerate it.

S3.2.7 Additional results under alternative measures of intercoder agreement

We reanalyzed our main survey findings using an alternative inter-coder reliability as a robustness check, Krippendorff's α. This follows the procedure described in Section S3.1.1 which we now apply to the main crowd annotation findings: for Findings 1 and 3, we pool all workers coding a text into a single multi-coder α, as in that section; for Finding 2 we instead take the mean pairwise α between each source and individual experts. This alternative measure corrects for chance agreement since our main analyses' mean pairwise agreement can be inflated if one label dominates (e.g. most texts are "No" in a paper). We note that these alternative measures were not preregistered and we report them as supplementary checks. Below, we reproduce the analyses for Findings 1, 2, and 3 using Krippendorff's α as the outcome measure.

First, for finding 1 (textual ambiguity), we reanalyze H6 where we split texts into strata by LLM and expert disagreement but used crowd's Krippendorff's α as the outcome measure instead of Gini impurity. We expect that more contested texts based on LLM or expert coding correlates with lower crowd α values, the inverse pattern of Gini impurity as the outcome. As we show in Figure S9, within-crowd α decreases from about 0.36 to 0.16 when moving from the least contested to most contested LLM stratum, and from 0.34 to 0.17 from expert unanimity to expert disagreement. This reanalysis is consistent with H6 that text ambiguity tracks low reliability (α) just as it tracks disagreement (Gini). We do not reanalyze the regression for H1 since unlike Gini impurity, Krippendorff's α is not defined per text but estimated over a set of texts, so it cannot serve as a text-level outcome.

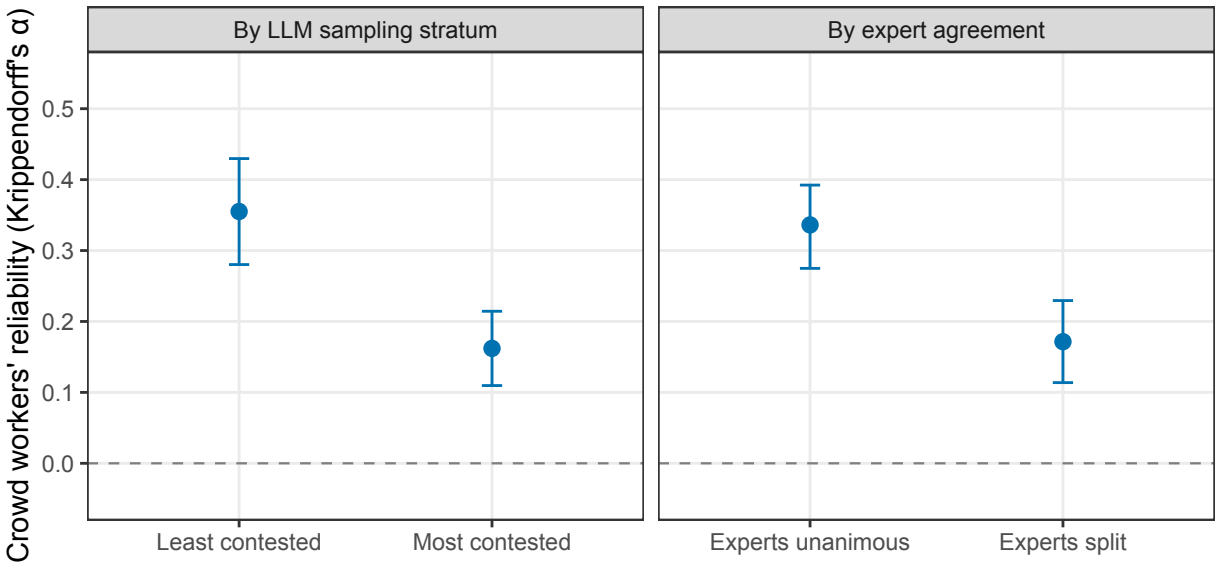

Figure S9: H6, Crowd-worker reliability by LLM sampling stratum and by expert agreement. Points are Krippendorff's α over the crowd workers within each group of texts, with bootstrap 95% confidence intervals; the dashed line marks the chance level of zero.

Second, for finding 2 (crowd workers trail LLMs), we replicate Figure 5 using Krippendorff's α between each source and the experts. Consistent with H2, crowd–expert α trails LLM–expert α on both the easy and hard papers (Figure S10). Of note, given that this analysis is using the 20-text-per-paper sample, levels of agreement across all measures are generally lower than the full corpus' with much wider confidence intervals. Nonetheless, the gap in point estimates between crowd and LLM or expert agreement persists regardless of the outcome measure used.

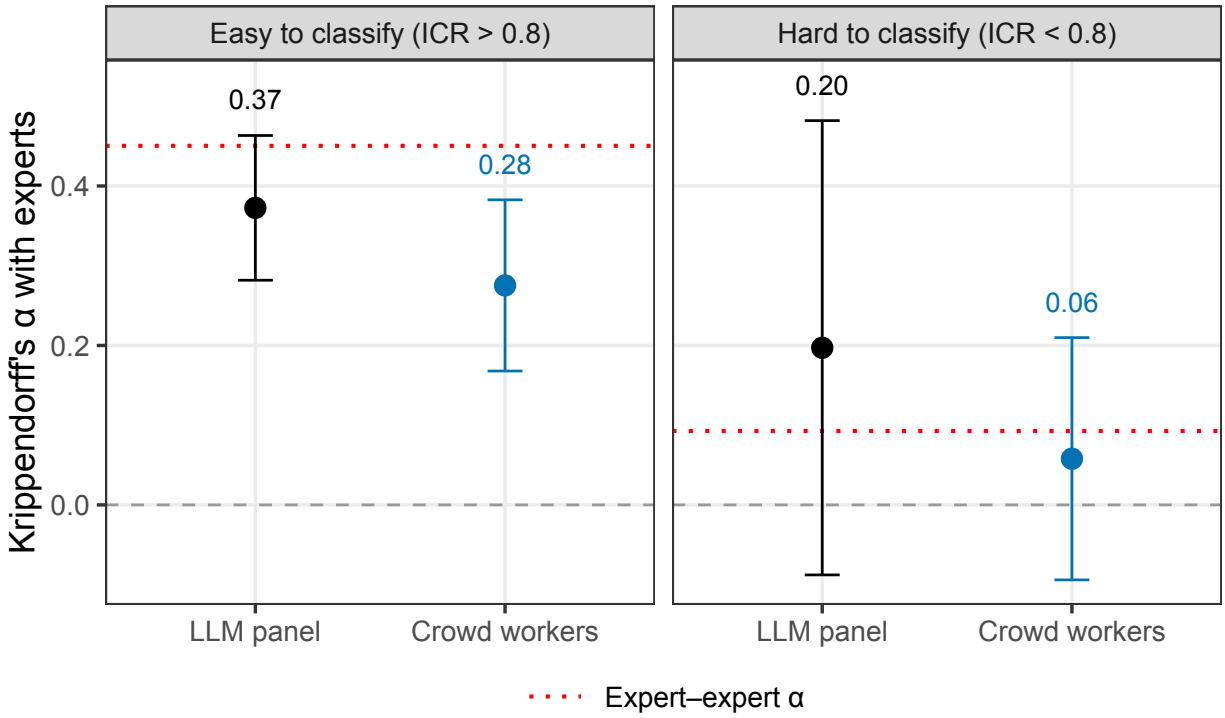

Figure S10: H2, Krippendorff's α between each source and the individual experts (mean pairwise, averaged within papers then across papers), with 95% confidence intervals across papers. The ten LLMs are collapsed into one panel average; the dotted line is the experts' own mean pairwise α; the dashed line marks chance ($\alpha = 0$); crowd workers in blue.

Third, for finding 3 (improved codebook helps agreement), we compare the crowd's Krippendorff's α under the original and improved codebooks. Because α is estimated over the set of 20 texts rather than per text, this is a descriptive comparison rather than the paired t-test we used for Gini impurity and agreement. Here, we not only see the same directional pattern that an improved codebook leads to higher agreement but a larger magnitude change of $\Delta\alpha = +0.308$ compared to using the crowd–expert pairwise agreement with $\Delta = +0.055$. Table S17 reports these estimates while Figure S11 plots the agreement and alpha analyses side-by-side for visual comparison. Overall, we find that all three of our crowd annotation's key findings hold when we substitute our outcome measure with Krippendorff's α.

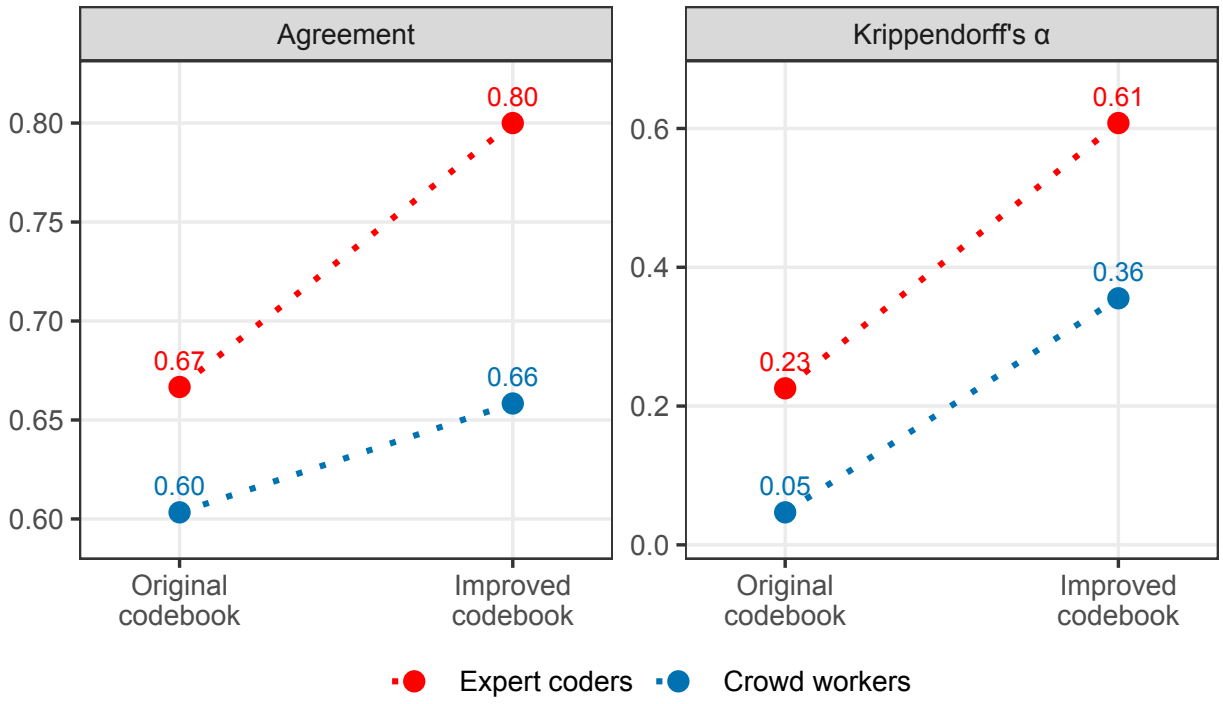

Figure S11: H3, Crowd–expert agreement (left) and within-source Krippendorff's α (right) under the original and improved Schub (2022) codebook.

S3.2.8 Additional results under an alternative measure of within-text disagreement

Alongside Krippendorff's α as an alternative to mean pairwise agreement, we reanalyze the findings that use Gini impurity under the normalized Shannon entropy introduced in Section S3.1.4. Here, we compute entropy per text over the crowd codings rather than over the LLM predictions.

Table S20 reports crowd disagreement on texts where the experts agree against texts where they disagree, alongside the same split by the LLM sampling stratum used to draw the survey texts. We observe that similar to Gini impurity, crowd entropy rises from 0.537 on the least contested texts to 0.800 on the most contested, as defined by LLM disagreement that informed our sampling of texts for the crowd survey (H6a). Likewise, crowd entropy rises from 0.593 where the experts agree to 0.789 where they disagree (H6b).

Splitting by study difficulty, the text heterogeneity finding holds among the easy studies but only partially among the hard studies. Among the hard studies, the LLM-stratum contrast holds, with crowd entropy rising from 0.662 to 0.883 when moving from the least to most contested stratum. The expert-agreement contrast among the hard studies is the one comparison that is directionally consistent but null.

We also reanalyze Finding 3 on codebook effects using entropy. We find that crowd entropy falls from 0.862 under the original codebook to 0.631 under the improved one (one-sided $p = 0.011$), reproducing the direction and the significance of the Gini impurity estimate for H3a. Therefore, we find that using Gini impurity or entropy as a measure of within-text heterogeneity does not change our substantive findings for the crowd annotation results.

Table S20: Crowd disagreement (normalized Shannon entropy) by expert agreement and by LLM sampling stratum

	Mean crowd entropy				
Panel A: By expert agreement					
	Experts unanimous	Experts split	Diff.	p	N
All studies	0.593	0.789	0.196	< 0.001	172 / 108
Easy (ICR > 0.8)	0.543	0.769	0.226	< 0.001	126 / 54
Hard (ICR ≤ 0.8)	0.729	0.809	0.080	0.151	46 / 54
Panel B: By LLM sampling stratum					
	Least contested	Most contested	Diff.	p	N
All studies	0.537	0.800	0.263	< 0.001	140 / 140
Easy (ICR > 0.8)	0.468	0.754	0.286	< 0.001	90 / 90
Hard (ICR ≤ 0.8)	0.662	0.883	0.221	< 0.001	50 / 50

Notes: Cells are the mean normalized Shannon entropy $-p \log_2 p - (1-p) \log_2 (1-p)$ of each text's share p of crowd Yes codes, an alternative to the Gini impurity of Table S15. N counts texts (unanimous / split in Panel A, least / most contested in Panel B); p-values are two-sided Welch t-tests over texts. Schub (2022) v2 arm excluded (280 texts, 14 studies).

S4 Details of Ambiguity Aware Bounds

This section extends the ambiguity-aware bounds introduced in Section 4.3 beyond the mean and then applies the framework to the empirical analysis in Arias (2022).

S4.1 Generalization beyond Average

As in the main text, suppose that we have N texts, indexed by $i = 1, \ldots, N$. Let Y_i denote the true annotation for text i. Although Y_i is latent, we observe an annotation produced by either a human or an LLM, denoted by $\hat{Y}_i$. We allow Y_i to be undefined. When it is well-defined, however, we assume that its support, denoted by $\mathcal{Y}$, is bounded. Thus, the annotation may be discrete, such as a yes-or-no classification, or continuous but bounded, such as a feeling thermometer score ranging from 0 to 100.

We are interested in a parameter θ characterized by the estimating equation

$$\mathbb{E}[\boldsymbol{m}(\boldsymbol{D}_i, Y_i; \theta)] = 0,$$

where $\boldsymbol{D}_i$ denotes the collection of all other variables entering the estimating equation. If there are no such variables, then $\boldsymbol{D}_i = \emptyset$. For example, when the parameter of interest is $\theta = \mathbb{E}[Y_i]$,

$$\boldsymbol{m}(\boldsymbol{D}_i, Y_i; \theta) = Y_i - \theta.$$

This formulation encompasses a wide variety of models, including regression models.

As described in the main text, we assume that researchers read a random subset of texts, indicated by $S_i = 1$, and annotate an ambiguity indicator $A_i \in \{0, 1\}$, which equals one if text i is ambiguous and the value of Y_i is therefore undefined. We further assume that, whenever researchers determine that an LLM has failed to follow the coding instructions (i.e., if there exists some i with $A_i = 0$ in which $Y_i \neq \hat{Y}_i$), they annotate the corresponding ground-truth label Y_i. We formalize this assumption as follows.

ASSUMPTION 1 (RANDOM SAMPLING) Researchers randomly select the texts to read such that

$$S_i \perp\!\!\!\perp \{A_i, Y_i\} \mid \hat{Y}_i, \hat{A}_i, \boldsymbol{D}_i. \tag{S2}$$

Moreover, the sampling probability $\pi := \Pr(S_i = 1 \mid \hat{Y}_i, \hat{A}_i, \boldsymbol{D}_i)$ is known to the researchers.

Under this assumption, we can derive the identified set by the same procedure as in the main paper. Specifically, for each candidate value θ, define the oracle bounding moments

$$\underline{\boldsymbol{M}}_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \Big[(1 - A_i)\, \boldsymbol{m}(\boldsymbol{D}_i, Y_i; \theta) + A_i \min_{y \in \mathcal{Y}} \boldsymbol{m}(\boldsymbol{D}_i, y; \theta) \Big],$$

$$\overline{\boldsymbol{M}}_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \Big[(1 - A_i)\, \boldsymbol{m}(\boldsymbol{D}_i, Y_i; \theta) + A_i \max_{y \in \mathcal{Y}} \boldsymbol{m}(\boldsymbol{D}_i, y; \theta) \Big],$$

where the minimum and maximum are taken element-wise. The identified set is

$$\Theta_I = \{\theta : \underline{\boldsymbol{M}}_N(\theta) \leq \mathbf{0} \leq \overline{\boldsymbol{M}}_N(\theta)\},$$

which includes any value θ that is consistent with the data if and only if some assignment of defensible annotations to the ambiguous texts solves the estimating equation.

Because A_i and, for unambiguous texts, Y_i are observed only when $S_i = 1$, we apply the design adjustment directly to each complete bounding-moment contribution. Specifically, define the feasible bounding moments $\underline{\boldsymbol{M}}^{\dagger}_N(\theta)$ and $\overline{\boldsymbol{M}}^{\dagger}_N(\theta)$ by

$$\underline{\boldsymbol{M}}^{\dagger}_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \Bigg[(1 - \hat{A}_i) \boldsymbol{m}(\boldsymbol{D}_i, \hat{Y}_i; \theta) + \hat{A}_i \min_{y \in \mathcal{Y}} \boldsymbol{m}(\boldsymbol{D}_i, y; \theta)$$

$$+\frac{S_i}{\pi}\Bigg\{(1-A_i)\boldsymbol{m}(\boldsymbol{D}_i,Y_i;\theta)+A_i\min_{y\in\mathcal{Y}}\boldsymbol{m}(\boldsymbol{D}_i,y;\theta)$$
$$-(1-\hat{A}_i)\boldsymbol{m}(\boldsymbol{D}_i,\hat{Y}_i;\theta)-\hat{A}_i\min_{y\in\mathcal{Y}}\boldsymbol{m}(\boldsymbol{D}_i,y;\theta)\Bigg\}\Bigg],$$

$$\overline{\boldsymbol{M}}_N^{\dagger}(\theta)=\frac{1}{N}\sum_{i=1}^{N}\Bigg[(1-\hat{A}_i)\boldsymbol{m}(\boldsymbol{D}_i,\hat{Y}_i;\theta)+\hat{A}_i\max_{y\in\mathcal{Y}}\boldsymbol{m}(\boldsymbol{D}_i,y;\theta)$$
$$+\frac{S_i}{\pi}\Bigg\{(1-A_i)\boldsymbol{m}(\boldsymbol{D}_i,Y_i;\theta)+A_i\max_{y\in\mathcal{Y}}\boldsymbol{m}(\boldsymbol{D}_i,y;\theta)$$
$$-(1-\hat{A}_i)\boldsymbol{m}(\boldsymbol{D}_i,\hat{Y}_i;\theta)-\hat{A}_i\max_{y\in\mathcal{Y}}\boldsymbol{m}(\boldsymbol{D}_i,y;\theta)\Bigg\}\Bigg].$$

Here, when $A_i = 1$, the term involving Y_i is not evaluated because the corresponding bounding value is determined by the minimum or maximum over $\mathcal{Y}$. We then obtain the feasible analogue of the identified set as

$$\Theta_I^{\dagger}=\left\{\theta:\underline{\boldsymbol{M}}_N^{\dagger}(\theta)\leq\mathbf{0}\leq\overline{\boldsymbol{M}}_N^{\dagger}(\theta)\right\}.$$

S4.2 Empirical Application

We finally illustrate the ambiguity-aware bounds by revisiting Arias (2022), in which a text-based measure serves as the outcome of a linear regression. This paper asks which states frame climate change as a security issue in the United Nations, a process known as *climate securitization*. The study argues that securitization follows agenda-control incentives: P5 states (Permanent members of the United Nations: China, France, Russia, United Kingdom, United States) benefit from moving climate change toward the Security Council, whereas highly vulnerable states such as Small Island Developing States (SIDS) may resist such a shift because it would reduce their agenda control. The outcome underlying this claim is constructed from text, because each climate-related speech segment must be classified as securitizing or not. We revisit this application because, as our replication shows, the concept of securitization is demanding even for trained coders. Some segments, such as those invoking existential harm without referring to security institutions, remain genuinely ambiguous (Table S7).

Following the original analysis, we estimate the linear probability model

$$Y_i=\beta_1\,\mathrm{P5}_{c(i)}+\beta_2\,\mathrm{SIDS}_{c(i)}+\boldsymbol{X}_{c(i),t(i)}^{\mathsf{T}}\boldsymbol{\gamma}+\delta_{t(i)}+\varepsilon_i, \tag{S3}$$

where Y_i is the true securitization label of segment i, delivered by country $c(i)$ in year $t(i)$. For genuinely ambiguous segments, this label may be undefined. The vector $\boldsymbol{X}_{c,t}$ contains country-level controls, including public concern about climate change, democracy, experienced climate disasters, changes in average temperature, voting affinity with the United States, military expenditure, logged GDP per capita, and logged population, and δ_t denotes year fixed effects.

To implement the bounds, we first develop a codebook and conduct an independent human review of a simple random sample of 200 segments.[9] To develop the codebook, one of the three authors reviewed the relevant literature on climate securitization and formulated the coding rules..[10] Using this codebook, all three

[9]We drew 200 segments at random for human review. One is a European Union observer statement that lies outside the country-level analysis corpus because it carries no country-year covariates, so 199 reviewed segments enter the later analysis frame.

[10]This includes Buzan et al. (1998), a canonical book on securitization, as well as other works that review or discuss the concept of securitization, including Levy (1995); Rønnfeldt (1997); Vogler (2023).

authors independently recorded both a binary classification and whether the codebook left a substantively defensible alternative classification for each segment. Finally, the author who conducted the literature review independently reclassified all segments for which at least one of the three authors identified ambiguity or for which the initial binary classifications disagreed. In the end, 24 texts are classified as ambiguous.

Next, we annotate all 4,525 segments using seven open-weight LLMs (Mistral 7B, LLaMA 3.1-8B, Qwen2.5-7B, Qwen2.5-14B, Gemma 3-12B, Gemma 3-27B, and GPT-OSS-20B) using the updated prompt introduced in Appendix S1.2.1. We take their majority vote as $\hat{Y}_i$ and construct the ambiguity proxy $\hat{A}_i$ from cross-model disagreement, measured by Gini impurity.

Using these annotations, we calculate the ambiguity-aware bounds using the following procedure. For a target coefficient β_k, let w_{ki} denote the OLS weight placed on segment i's outcome when estimating that coefficient. Because OLS is linear in the outcome, if all labels were well-defined, the coefficient could be written as

$$\hat{\beta}_k = \sum_{i=1}^{N} w_{ki} Y_i. \tag{S4}$$

The weights depend only on the observed regressors, including the controls and year fixed effects, and are therefore known once the regression design is fixed.[11] In particular, assigning segment i the label $Y_i = 1$ rather than $Y_i = 0$ changes $\hat{\beta}_k$ by w_{ki}. Thus, when $w_{ki} < 0$, assigning the label 1 lowers the coefficient, whereas when $w_{ki} > 0$, assigning the label 1 raises it. The oracle lower bound therefore assigns label 1 to ambiguous segments with negative weights and label 0 to those with positive weights; the oracle upper bound does the reverse.

Applying the design adjustment directly to the complete lower- and upper-bound contributions gives the feasible bounds

$$\begin{aligned}
\hat{\beta}_k \in \Bigg[& \sum_{i=1}^{N} w_{ki} \Bigg[(1-\hat{A}_i)\hat{Y}_i + \hat{A}_i \, \mathbb{1}\{w_{ki} < 0\} \\
& \qquad + \frac{S_i}{\pi} \Big\{ (1-A_i)Y_i + A_i \, \mathbb{1}\{w_{ki} < 0\} - (1-\hat{A}_i)\hat{Y}_i - \hat{A}_i \, \mathbb{1}\{w_{ki} < 0\} \Big\} \Bigg], \\
& \sum_{i=1}^{N} w_{ki} \Bigg[(1-\hat{A}_i)\hat{Y}_i + \hat{A}_i \, \mathbb{1}\{w_{ki} > 0\} \\
& \qquad + \frac{S_i}{\pi} \Big\{ (1-A_i)Y_i + A_i \, \mathbb{1}\{w_{ki} > 0\} - (1-\hat{A}_i)\hat{Y}_i - \hat{A}_i \, \mathbb{1}\{w_{ki} > 0\} \Big\} \Bigg] \Bigg].
\end{aligned} \tag{S5}$$

The terms inside the design adjustment are evaluated only for reviewed segments with $S_i = 1$, and Y_i is required only when the reviewed segment is unambiguous, so that $A_i = 0$. Under Assumption 1, the two endpoints in Equation (S5) are design-unbiased for the corresponding oracle lower and upper endpoints.

For comparison, we consider two benchmark approaches. First, we estimate the original regression using the majority vote of the LLM annotations as the outcome, without accounting for ambiguity. Second, we apply design-based supervised learning (DSL; Egami et al. 2024), treating the majority vote of the human annotations as the ground-truth label. DSL corrects for measurement error in LLM annotations, but it does not accommodate fundamental ambiguity in the classification task because it assumes that each text has a uniquely defined gold-standard label.

[11] Let $\boldsymbol{Z}$ denote the full design matrix and let $\mathbf{e}_k$ be the unit vector selecting coefficient k. Then w_{ki} is the ith element of $\mathbf{e}_k^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top$.

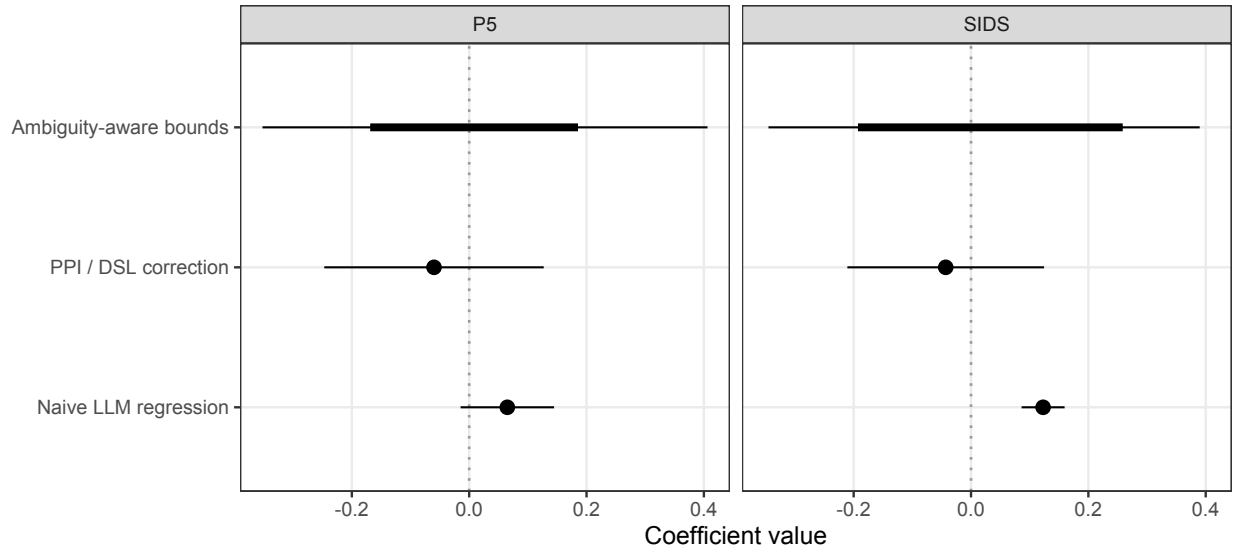

Figure S12: Coefficients of β_1 (P5) and β_2 (SIDS), along with their 95% confidence intervals, based on (i) the proposed ambiguity-aware bounds, (ii) PPI / DSL correction (which corrects the measurement error in LLM annotations by regarding human annotation as ground truth), and (iii) the naive LLM regression (which uses LLM predictions as the outcome without accounting for the underlying ambiguity of the concepts).

Figure S12 compares the ambiguity-aware bounds with the naive LLM regression and DSL. Both benchmark approaches yield substantially more precise estimates because they impose a unique label on every text and therefore do not account for conceptual ambiguity in the annotation task. This precision can be misleading: the underlying concept of climate securitization remains ambiguous under the paper's definition,[12] and alternative defensible classifications of the ambiguous texts can even reverse the sign of the estimated coefficient.

These results underscore the importance of developing a precise operationalization and translating it into a clear codebook. Researchers should resolve avoidable ambiguity through substantive refinement of the coding rules before turning to statistical adjustment. The ambiguity-aware bounds are designed for the ambiguity that remains after this process, making explicit the extent to which the substantive conclusion depends on unresolved borderline cases.

[12]In the paper, climate securitization is defined as "emphasizing the security dimensions of an issue, specifically the language of existential threat" and indicating the United Nations Security Council's (UNSC) jurisdiction over an issue (Arias, 2022, p3). However, some texts are borderline cases given this definition. Take the following excerpted text as an example:

> *"[W]e must all assume our fair share of responsibility in effective international cooperation to find solutions to global warming and climate change, which pose, as never before, a grave threat to humankind's survival".*

While the text does not explicitly call climate change a security issue, it refers to it as a "grave threat to humankind's survival" a phrase that could be interpreted as "existential threat"; on the other hand, it contains no element of indicating UNSC jurisdiction, which is also absent from most texts that were classified as "securitization" in the paper's original structural topic model.

Another ambiguous example arises from the text discussing multiple issues simultaneously, including both climate change and security, leaving room for interpretation about whether the two topics are referred to jointly or separately. For example:

> *"Alone we can do little, but together we can achieve much. The challenges that we face globally — from climate change to refugees to war and violence — require urgent action now.".*