MindVote: When AI Meets the Wild West of Social Media Opinion

Xutao Mao[1], **Ezra Xuanru Tao**[1], **Leyao Wang**[2]
[1]Vanderbilt University
[2]Yale University
xutao.mao@vanderbilt.edu, ezra.tao@vanderbilt.edu, leyao.wang.lw855@yale.edu

Abstract

Large Language Models (LLMs) are increasingly used as scalable tools for pilot testing, predicting public opinion distributions before deploying costly surveys. However, the prevailing paradigm for evaluating these models relies on traditional structured surveys—a methodology misaligned with the more realistic scenarios like social media where opinions are rich in digital contexts. By design, surveys strip away the social and cultural context that shapes public opinion, and LLM benchmarks built on this paradigm inherit these critical limitations. To bridge this gap, we introduce MindVote, the first benchmark for public opinion prediction grounded in authentic social media discourse. MindVote is constructed from 3,918 naturalistic polls sourced from Reddit and Weibo, spanning 23 topics and enriched with detailed annotations for platform and topical context. Using this benchmark, we conduct a comprehensive evaluation of 15 LLMs, revealing a critical "*survey-based specialization pitfall*" where models fine-tuned on traditional surveys underperform their general-purpose counterparts and demonstrating the necessity of context in social media. MindVote provides a robust, ecologically valid framework to move beyond survey-based evaluations and advance the development of social intelligent AI systems.

1 Introduction

A core application for Large Language Models (LLMs) in understanding public opinion is serving as a rapid, scalable tool for pilot testing—predicting how a population will respond to a question before deploying costly, large-scale surveys (Rothschild et al. 2024; Bisbee et al. 2024; Suh et al. 2025; Sinacola, Pachot, and Petit 2025; Cao et al. 2025a; Anthis et al. 2025; Binz et al. 2025). This capability is crucial across a wide spectrum of domains. For example, LLMs can be used to predict reactions to new entertainment content and technological innovations, understand career aspirations, navigate complex social issues, and track evolving lifestyle trends (Argyle et al. 2023; Qian et al. 2025; Chen, Wang, and Ma 2025). However, the central goal is not simply to identify the majority opinion, but to predict how opinions are distributed across all possible choices. Understanding the full distribution is important because it reveals whether society is unified, divided, or contains strong dissent—information that a single majority number cannot capture. (Zhou, Nie, and Bansal 2022; Moon et al. 2024; Meister, Guestrin, and Hashimoto 2025; Cao et al. 2025b).

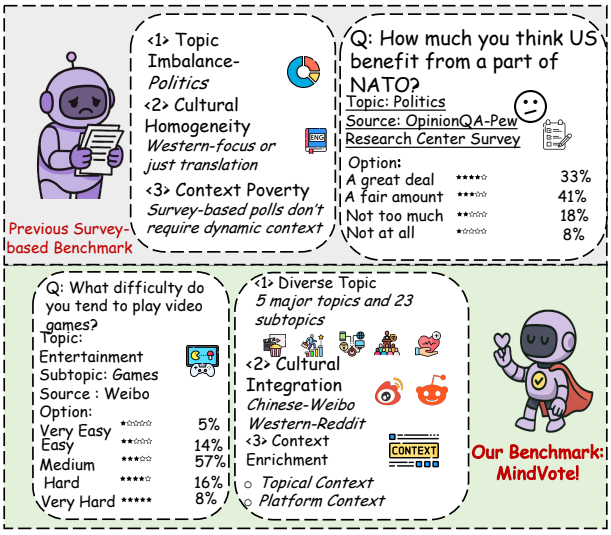

Figure 1: MindVote benchmark addresses three key limitations of previous survey-based approaches. Our benchmark provides diverse topics, cultural integration, and rich contextual metadata, overcoming the topic imbalance, cultural homogeneity, and context poverty of traditional survey datasets.

However, the prevailing paradigm for public opinion distribution prediction relies on traditional structured surveys (Krogstad, Passel, and Cohn 2016; Santurkar et al. 2023; Elkjær and Wlezien 2024; Suh et al. 2025). This approach does not align with the fast-paced, context-rich digital environments—such as social media platforms—where opinions are now actively formed and expressed (Boutyline and Soter 2021). While surveys include basic demographic questions—such as age, gender, or education level—these static variables are poor proxies for the dynamic, situated nature of social context (Zaller and Feldman 1992). Demographics may indicate who a person is in broad categories, but fail to reflect how opinions are specifically shaped by lived experiences, group affiliations, ongoing conversations, or cultur-

ally reasoning patterns (Kahan, Jenkins-Smith, and Braman 2011; Markus and Kitayama 2014; McIntosh and Wright 2019).

Consequently, LLM benchmarks built on this flawed paradigm, such as OpinionQA and SubPop (Santurkar et al. 2023; Suh et al. 2025), inherit these critical limitations. This reveals a crucial gap in the responsible deployment of LLMs. As these models are increasingly used to act as a proxy for human populations, ensuring their accuracy becomes essential. A proper evaluation framework must identify when models fail at context-rich, real-world tasks. Without this, we risk promoting systems that misrepresent the groups they are meant to simulate. This misrepresentation has severe downstream consequences, potentially compromising strategic decisions across critical topics like public policy, marketing, technology development, and community management (Qi 2024; Radivojevic, Clark, and Brenner 2024; Qu and Wang 2024; Panickssery, Bowman, and Feng 2024; Wang, Morgenstern, and Dickerson 2025). This reliance on traditional survey-based evaluation creates three critical gaps in our ability to assess an LLM's true social understanding:

The Gap about Topic Imbalance. Real-world online discourse is thematically diverse (Song et al. 2023); for instance, entertainment drives 70% of traffic on Weibo (China Marketing Corp 2024), while Reddit's largest communities focus on gaming and technology (Proferes et al. 2021). Yet, current survey-based benchmarks are heavily skewed towards formal, institutional topics like politics, which represent only a fraction of naturalistic opinion expression (Santurkar et al. 2023; Zhao et al. 2024). Because of this mismatch, we cannot evaluate if a model can adjust its reasoning from formal political surveys to the everyday language used on social media (Karjus and Cuskley 2024; Reveilhac and Morselli 2024).

The Gap about Cultural Homogeneity. Existing cross-cultural survey benchmarks often exhibit cultural homogeneity, creating test conditions on Western-centric questions even with other languages translation (Haerpfer et al. 2022; DURMUS et al. 2024; Zhao et al. 2024). A model might perform well on a translated question, but we are left unable to determine if this success stems from genuine cultural understanding or a superficial linguistic competence (Singh et al. 2025).

The Gap about Context Poverty. Opinions on social media are not noise but are helpful predictive signals (López-Rabadán 2021). This is correlated with *context priming*, a phenomenon where exposure to a specific context influences how individuals perceive and respond (Doyle and Lee 2016). On social platforms, users are primed by factors such as platform norms (e.g., Reddit's anonymity vs. Weibo's public-facing profiles (Zhu 2024)), temporal events (Banducci and Stevens 2015), and community-specific discourse (Qiao 2023). In contrast, survey-based benchmarks systematically remove this contextual priming to achieve standardization (Tourangeau, Singer, and Presser 2003), which prevents a true evaluation of a model's ability to leverage these critical predictive cues (Li et al. 2023).

To overcome these limitations, we move beyond surveys and instead use naturalistic social media polls—a more direct source for analyzing public opinion formation (Scarano et al. 2024). These polls address three critical gaps and provide a robust foundation for evaluation. The comparison between existing survey-based benchmarks and our benchmark is in Figure 1.

Our Contributions. We introduce **MindVote**, the first benchmark for public opinion distribution prediction grounded in realistic social media discourse. We advance the field with the following innovations:

- We construct and release MindVote, a benchmark of 3,918 authentic polls from two platforms (Reddit and Weibo) in their native English and Chinese. It spans 5 major topics and 23 sub-topics (shown in Table 1) and is enriched with detailed social context annotations.
- We benchmark leading LLMs, identify top performers, and reveal that models fine-tuned on traditional surveys face a critical *survey-based specialization pitfall*.
- We demonstrate that enhancing a model's capacity with social-context reasoning is more effective than fine-tuning on context-stripped data, offering a new direction for developing more socially intelligent systems.

2 Related Work

Opinion Distribution Prediction Benchmarks. U.S.-focused benchmarks like OpinionQA (Santurkar et al. 2023) and SubPop (Suh et al. 2025) concentrate on political topics, while international efforts like GlobalOpinionQA (DURMUS et al. 2024) is Western-centric frameworks and WorldValuesBench (Zhao et al. 2024) is framed as multicultural value prediction. Cao et al. (2025a) introduces three datasets (two English, one Chinese) specifically for group-level distribution prediction. These approaches contain only demographic-value pairings, structured survey format, or Western-centric lens that abstract away the naturalistic cultural contexts that are essential for authentic opinion distribution prediction.

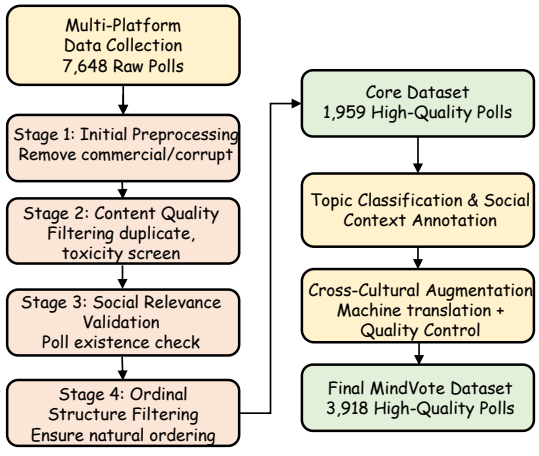

Figure 2: This flow pipeline demonstrates the dataset construction from data source to final composition of datasets.

Main Category	Sub-category	Percentage (%)
Entertainment (23%)	Game	6
	Movies/TV	5
	Music	4
	Sports	3
	Social Media	3
	Celebrities	2
Technology (22%)	AI/ML	8
	Tech Products	7
	Software/Apps	4
	Cybersecurity	3
Lifestyle (27%)	Finance	8
	Travel	6
	Cooking	5
	Mental Health	3
	Fitness	3
	Nutrition	2
Social Issues (19%)	Politics	7
	Social Justice	6
	Climate	3
	Immigration	2
	Tech Policy	1
Career (9%)	Job	4
	Prof. Dev.	3
	Work-Life	2

Table 1: Distribution of Topics by Category

3 Benchmarking Setup

3.1 Dataset Construction

MindVote's construction involved transforming 7,648 raw polls into 3,918 high-quality polls through strategic data sourcing, social context annotation, rigorous quality control, and the cross-cultural augmentation. Figure 2 shows the flow of dataset construction.

Platform Selection Strategy. Our platform selection strategy is designed to evaluate distinct aspects of LLM opinion prediction capabilities. Reddit provides an anonymous environment where users express unfiltered, authentic opinions without identity-based social constraints, requiring models to interpret and predict genuine thoughts purely through content analysis without relying on user profiles or reputation cues. Weibo enables the evaluation of model performance in culture-specific contexts, testing models' understanding of Chinese cultural contexts, social norms, and culture-laden discourse patterns.

Multi-Platform Data Collection. We collected 7,648 polls across two platforms spanning 2019-2025. Weibo dataset contributed 3,757 polls (2,026 from existing datasets (Lu et al. 2021) from 2019 to 2021, 1,734 newly crawled) and are anonymized for users' personal identifiable information, capturing Chinese social media dynamics across pandemic and post-pandemic periods. Reddit provided 3,891 polls from diverse subreddits including *r/poll* during 2021-2025.

Quality Control Pipeline. Our four-stage pipeline efficiently produced high-quality 1,959 core polls:

1. **Initial Preprocessing:** This process included the removal of commercial votes, targeting promotional polling content that lacked authentic user engagement. Furthermore, format-corrupted poll data was identified and discarded through a systematic validation of structural integrity.
2. **Content Quality Filtering:** This filtering begins with duplicate content removal using a MinHash algorithm to eliminate items with greater than 95% overlap (Shrivastava and Li 2014), effectively targeting redundant trending topics, reposted polls, and cross-platform duplicates. Subsequently, automated toxicity screening was conducted using the Google Perspective API, with polls retaining a toxicity score below 0.4 being retained to filter out hate speech and overly controversial subjects (Lees et al. 2022).
3. **Social Relevance Validation:** Human verification of voting patterns (removing polls $\leq$ 100 votes for social relevance) and verification of poll and vote existence in all two platforms to ensure authentic social engagement and meaningful community participation.
4. **Ordinal Structure Filtering:** We performed systematic filtering to ensure all selected polls exhibit natural ordering relationships between options to ensure structural alignments with existing benchmarks (Santurkar et al. 2023). We systematically excluded polls with purely nominal options (e.g., preference choices among unordered alternatives). This filtering employed LLM-as-a-judge (DeepSeek-R1 (DeepSeek-AI Team 2025)), validated through human annotation achieving Fleiss'κ = 0.59 agreement.

Topic Classification and Social Context Annotation. Unlike survey-based benchmarks that focus on pan-political and social topics, MindVote includes 5 major topics and 23 specific topics. We assigned topic labels through a classification process combining automated judgment and human validation. Initial classification used content-based detection employing LLM-as-a-judge (Deepseek-R1) to identify primary topic. Human validation was applied selectively to ambiguous cases where automated classification was uncertain or polls exhibited mixed topic characteristics (Chen et al. 2024). Trained annotators using standardized topic definitions achieved inter-annotator agreement of Fleiss'κ = 0.55 for these edge cases, with expert consensus resolving conflicts and assigning polls spanning multiple topics to their primary thematic focus.

We manually annotate each poll includes rich metadata for social context enrichment: general platform context (e.g., user statistics and user behaviors) and topical context (topic-specific discourse patterns).

Cross-Cultural Augmentation. We created a parallel bilingual corpus for all two platforms through machine translation with rigorous quality control to demonstrate both linguistic and cultural effects: 10% back-translation validation (BLEU $>$ 35 (Papineni et al. 2002)), 5% native speaker

Poll: How threatened do you feel by AI replacing your job? (Platform: Reddit, Date: April 14, 2025, Votes: 6,567)	
Options	1. Not at all threatened 2. Slightly threatened 3. Moderately threatened 4. Very threatened
Platform Context	Reddit user base: 58% US, 46% college-educated, generally tech-oriented.
Topical Context	Broad AI adoption. 78% of organizations use AI; 55% of Americans use AI regularly.

Table 2: An example poll from our MindVote dataset, demonstrating the structure of the question, options, and associated social context provided for model evaluation.

rewriting, and expert review (Fleiss' κ = 0.51). The total number of polls becomes 3,918 after augmentation.

Final Dataset Composition. The final MindVote dataset is composed of 3,918 polls, with 2,158 sourced from Reddit and 1,760 from Weibo. Each poll is enriched with a comprehensive set of metadata, including its creation time, total vote count, and layers of social context: general platform context and topical context within that platform. To ensure broad accessibility and ease of use, the entire dataset is provided in both CSV and JSON formats. Table 2 shows an example with its simplified metadata keywords.

3.2 Experiment Design

Task. We evaluate LLMs on opinion distribution prediction within naturalistic social discourse in MindVote dataset including 3,918 polls. Given a poll question with metadata including social contexts, models predict the probability distribution over answer choices.

Models. We evaluate 15 leading LLMs across closed-source, open-source and specialization categories: Claude-3.7-Sonnet-02-24 (Anthropic 2025), GPT-4o-2024-11-20 (OpenAI 2025b), GPT-4.1-04-14 (OpenAI 2025a), Gemini-2.5-Pro-05-06 (Google Cloud 2025), o3-medium-04-16 (OpenAI 2025c), Deepseek-R1-05-28 (DeepSeek-AI Team 2025), Qwen2.5-32B-Instruct (Qwen Team 2024), Gemma-2-9B-it (Team, Gemma et al. 2024), Mistral-7B-Instruct-v0.1 (Chaplot 2023), Llama-2-13B-Chat (Touvron et al. 2023), Llama-3-70B-Instruct (Grattafiori et al. 2024), and Llama-4-Maverick-17B-128E (Meta AI 2025). We also evaluate three opinion distribution prediction specialization models from Suh et al. (2025). These models are fine-tuned using LoRA (Hu et al. 2022) on base models: Llama-3-70B-Base (Grattafiori et al. 2024), Llama-2-13B-Base (Touvron et al. 2023), and Mistral-7B-v0.1 (Chaplot 2023) which we named them respectively as: SubPop-Llama-3-70B, SubPop-Llama-2-13B, SubPop-Mistral-7B.

Pipeline. All primary evaluations use a greedy decoding strategy (temperature=0) with default hyperparameter settings under zero-shot with context annotation (Kojima et al. 2022; Han et al. 2025). For the specialization fine-tuned models, we adapt those models into our pipeline by first loading the respective pretrained base model and then applying the publicly available LoRA weight checkpoints provided by Suh et al. (2025). All experiments are conducted on two H100 GPU with CUDA 12.1.

Prompt. To ensure consistent and machine-readable outputs, we employ a structured prompting strategy where the model is given a JSON object containing the poll and its context. The template instructs the model to assume the role of a *"opinion distribution prediction expert analyzing voting patterns and social dynamics."* The prompt includes the poll question and is enriched with social context metadata, with instruction for step-by-step reasoning. The model's task is to return a JSON object with a schema identical to the input, replacing placeholder fields with its numeric predictions for the voting distribution.

Evaluation Metrics. We adopt four distinct metrics to provide a comprehensive evaluation. Our primary metric is **1 - Wasserstein Distance (1-Wass.)** (Santurkar et al. 2023; Meister, Guestrin, and Hashimoto 2025; Suh et al. 2025). The Wasserstein Distance measures the minimum cost for transforming one distribution into another, crucially accounting for semantic similarity between answer choices by treating them as points in a metric space. To complement this, we also report **Spearman's Rank Correlation Coefficient (ρ)** (Zhou, Nie, and Bansal 2022), a non-parametric measure of how well the predicted ranking of options matches the true ranking of vote shares; **1 - KL Divergence** (Meister, Guestrin, and Hashimoto 2025), which quantifies the information loss when using the model's predicted distribution to approximate the ground truth; and **One-hot Accuracy** (Zhou, Nie, and Bansal 2022; Santurkar et al. 2023; Suh et al. 2025), which provides a strict measure of whether the single most likely predicted answer is correct.

Evaluation Boundary. We include upper bounds and lower bounds for comparisons following (Suh et al. 2025). The upper bound is established by sampling subsets of the original results, calculating the four metrics between subsampled and original distributions, and performing bootstrapping to obtain a robust estimate that captures the intrinsic variance arising from the respondent sampling process in opinion. The uniform distribution lower bound establishes a performance floor equivalent to random chance.

4 Results

4.1 Performance of General Purpose Models

We analyze the overall performance of ten leading general purpose LLMs on the MindVote benchmark in Table 3. Across all four metrics, the closed-source model o3-medium consistently outperforms all other models, establishing the highest performance ceiling among current general-purpose LLMs. Among open-source models, Deepseek-R1 demonstrates the strongest results, narrowly outperforming other models in its class. However, a critical gap remains between the top-performing models and the upper bound for all metrics. This highlights substantial opportunities for improvement in modeling authentic public opinion distribution.

4.2 Survey-based Specialization Pitfalls

Our results in Table 3 show that fine-tuning on a structured survey dataset leads to a significant decrease in performance when predicting opinion distributions on social media. To demonstrate this, we compare our specialization models (base models that fine-tuned on traditional but context-stripped survey data) against their original base instruction-tuned models (the general-purpose models).

As shown in Table 3, specialization models consistently underperform their base counterparts, even when evaluated with full context annotations. For example, the SubPop-Llama-2-13B specialization model experiences a 3.3 percentage point decrease in its 1-Wass. score relative to the base Llama-2-13B. These results indicate that fine-tuning on sanitized data does not equip models to handle the complexities of real-world contexts. Rather than enhancing performance, such narrow specialization actually impairs the models' ability to generalize to authentic social discourse.

Model	1-Wass.	1-KL Div.	Spearman.	Acc.
Closed-source Models				
o3-medium	**0.892**	**0.859**	**0.756**	**0.581**
Gemini-2.5-Pro	0.891	0.845	0.751	0.564
Claude-3.7-Sonnet	0.891	0.851	0.722	0.551
GPT-4o	0.880	0.836	0.691	0.515
GPT-4.1	0.874	0.845	0.688	0.524
Open-source Models				
Deepseek-R1	0.876	0.831	0.739	0.558
Qwen2.5-32B	0.866	0.787	0.605	0.483
Llama-4-17B	0.820	0.731	0.659	0.429
Llama-3-70B	0.844	0.752	0.641	0.461
Llama-2-13B	0.807	0.718	0.592	0.369
Gemma-2-9B	0.802	0.705	0.575	0.362
Mistral-7B	0.808	0.719	0.597	0.365
Specialization Models				
SubPop-Llama-3-70B	0.805	0.713	0.593	0.417
SubPop-Llama-2-13B	0.774	0.693	0.558	0.378
SubPop-Mistral-7B	0.782	0.695	0.546	0.370
Upper Bound	0.972	0.976	0.961	0.964
Lower Bound	0.701	0.663	0.000	0.307

Table 3: Opinion distribution prediction performance of LLMs on the MindVote Benchmark. Scores are presented as Mean values, evaluated on four different metrics: 1-Wasserstein distance (1-Wass.), 1-KL Divergence (1-KL Div.), Spearman's Rank Correlation (Spearman.), and One-hot Accuracy (Acc.). **All metrics are the higher the better.**

4.3 Performance Across Topics and Cultures

Performance varies by Topic. As shown in Table 4, model performance differs notably across topical domains. On average, models perform better on *Social Issues* and *Lifestyle* topics, which typically exhibit more structured discourse aligned with standard pre-training corpora. In contrast, performance declines in domains such as *Technology* and *Entertainment*, where specialized jargon, community-specific norms, and rapidly evolving vernacular are prevalent. This disparity indicates that models have difficulty adapting their reasoning from general knowledge to the nuanced linguistic and social norms of diverse online communities.

Model	Tech.	Soc.	Enter.	Career	Life.
Closed-source Models					
o3-medium	0.868	**0.925**	0.869	**0.904**	0.903
Gemini-2.5-Pro	0.860	0.917	0.881	0.896	**0.906**
Claude-3.7-S.	0.865	0.908	**0.896**	0.891	0.894
GPT-4o	0.859	0.893	0.872	0.895	0.884
GPT-4.1	0.866	0.913	0.854	0.878	0.868
Open-source Models					
Deepseek-R1	**0.889**	0.892	0.868	0.879	0.887
Qwen2.5-32B	0.846	0.876	0.867	0.868	0.875
Llama-4-17B	0.824	0.844	0.832	0.866	0.857
Llama-3-70B	0.835	0.853	0.836	0.845	0.854
Llama-2-13B	0.798	0.824	0.801	0.789	0.817
Gemma-2-9B	0.802	0.839	0.832	0.848	0.885
Mistral-7B	0.809	0.836	0.812	0.804	0.828
Specialization Models					
SubPop-Llama-3-70B	0.794	0.821	0.798	0.786	0.813
SubPop-Llama-2-13B	0.762	0.798	0.769	0.751	0.785
SubPop-Mistral-7B	0.771	0.803	0.778	0.764	0.792
Average	0.830	0.863	0.838	0.844	0.857

Table 4: Opinion distribution prediction performance across different topics. The final row shows the average performance across all models for each major topic.

Performance Reflects Cultural Origin. The analysis in Figure 3 reveals a *Cultural Gap*, showing models have a strong cultural alignment with their origin. Western models like Gemini-2.5-Pro excel on Reddit, while Chinese models such as DeepSeek-R1 dominate on Weibo. We confirmed this gap is primarily cultural, not linguistic, by evaluating translated content; the performance penalty from translation was minimal compared to the large drop across cultural sources. This systematic variation demonstrates a clear alignment effect, where models are more adept at interpreting the cultural norms of their training data.

4.4 Analysis

Section 4.2 highlights a pitfall of survey-based specialization: base models fine-tuned on sanitized survey datasets show degraded performance, even when *context priming*—adding relevant situational or background information to guide predictions—is applied (Doyle and Lee 2016). This unexpected finding raises three questions that we expect to investigate:

- Does social context in social media serve as noise or a helpful signal?
- How does increasing contextual complexity affect model performance, especially for survey-specialized models?
- Does adding social contexts better enhance models' opinion prediction compared with few-shot learning?

These questions emerge naturally from our specialization findings. The failure of fine-tuned models, even when

provided with full context, suggests two possibilities: either (1) platform and topic are less important than previously assumed, or (2) these models are unable to effectively utilize such contextual information. This motivates our first investigation into whether social context is truly beneficial signal or noise. Moreover, the suboptimal performance of specialized models reflects their difficulty in processing social context, motivating our second analysis on factors contributing to context complexity. Third, because our context-rich approach relies on contextual understanding rather than example-based training as in few-shot learning, a direct comparison is needed to assess the effectiveness of both strategies.

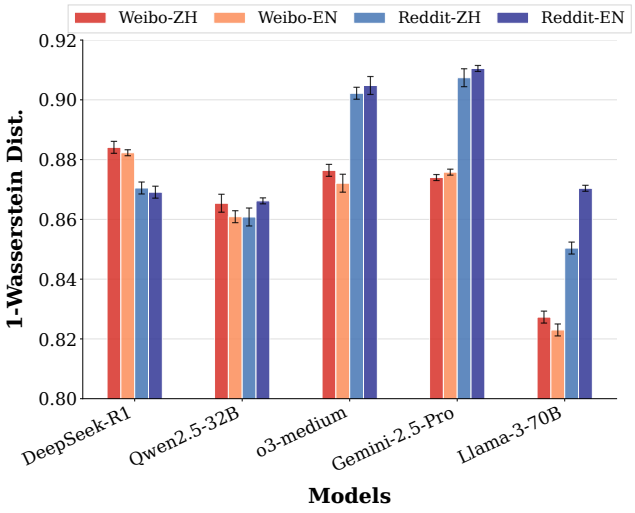

Figure 3: LLMs exhibit a cultural alignment with their origin. Models trained from Western data better on Reddit (EN-English), while those from China (e.g., DeepSeek-R1) excel on Weibo (ZH-Chinese). This cultural gap persists after controlling for linguistic effects. Error bars represent 95% CI.

Social Context is a Critical Signal. Our context ablation study (Table 5) confirms that social context is not noise but a vital signal for accurate opinion prediction. Removing all contextual metadata (No Context) results in the most substantial performance drop on average (*-5.91%*), followed closely by platform-specific information removal (*-5.12%*) and topical context removal (*-4.52%*).

Notably, specialized fine-tuned models exhibit dramatically larger degradation across all ablation conditions. This severe degradation approaches the theoretical lower bound of uniform distribution performance, suggesting these models have become overly dependent on contextual signals used for priming. Since these models were fine-tuned on structured survey data in standardized formats, they struggle to adapt when deployed on natural social media polls when contextual information may be fragmented or entirely absent. While these models may excel in controlled survey environments, their heightened sensitivity to context removal in real-world settings reveals a limitation, constraining their applicability across diverse online environments.

Model	w/o Plat.	w/o Topic.	No Ctx.
Closed-source Models			
o3-medium	-3.03	-2.31	-4.13
Gemini-2.5-Pro	-2.37	-3.30	-4.34
Claude-3.7-Sonnet	-4.46	-3.91	-5.53
GPT-4o	-5.75	-3.39	-4.92
GPT-4.1	-4.92	-3.36	-5.30
Open-source Models			
Deepseek-R1	-4.19	-3.55	-5.98
Qwen2.5-32B	-4.42	-3.81	-6.24
Llama-3-70B	-4.82	-4.78	-5.19
Llama-4-17B	-5.80	-5.33	-6.08
Llama-2-13B	-4.95	-5.26	-5.92
Gemma-2-9B	-5.37	-4.74	-6.47
Mistral-7B	-5.92	-4.84	-6.55
Specialization Models			
SubPop-Llama-3-70B	-6.82	-6.68	-6.95
SubPop-Llama-2-13B	-6.94	-6.11	-7.23
SubPop-Mistral-7B	-6.89	-6.41	-7.54
Average	-5.12	-4.52	-5.91

Table 5: Opinion distribution prediction performance degradation from the full-context baseline. All scores represent the drop in 1-Wasserstein Distance (%). Abbreviations: Plat. (Platform), Topic. (topical), Ctx. (Context).

Contextual Complexity. While our ablation study confirms that social context is critical for performance, its inherent complexity presents a significant challenge. To dissect this, we investigate how model performance degrades as complexity increases about our annotated context. We define contextual complexity using three metrics: (1) context length by tokens of combined metadata; (2) language informality in the context; and (3) niche topic which constitutes a small fraction of the overall dataset. Our analysis in Figure 4 reveals that increasing complexity in any of these areas has negative correlations with the 1-Wass. performance.

This effect is most pronounced for our specialization models. Specifically, these models, fine-tuned on structured and formal survey data, exhibit strong performance degradation when faced with longer contexts, higher degrees of informality, and more niche topics. This demonstrates a critical brittleness: while trained to be domain experts, their reliance on formal data structures makes them particularly vulnerable to the unstructured, multifaceted, and messy nature of authentic social media discourse.

Contextual Priming vs. Few-Shot Learning. Our "survey-based specialization pitfall" finding (Section 4.2) raised a key question: do models benefit more from understanding a poll's environment via contextual priming or from imitating examples via few-shot learning? To evaluate this, we compared the model's performance across three distinct settings. The first was a zero-shot setting with context priming, which followed our default context annotation pipeline. We contrasted this against two settings without context priming: a standard zero-shot setting without context priming, same as we have done in Table 5 and a few-shot (1-4 examples) set-

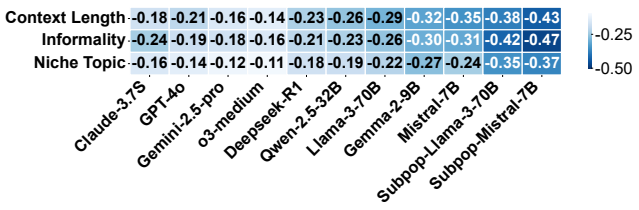

Figure 4: Correlation patterns between model's 1-Wass. performance and complexity dimensions. Survey-specialized models exhibit strong brittleness to social media contextual complexity.

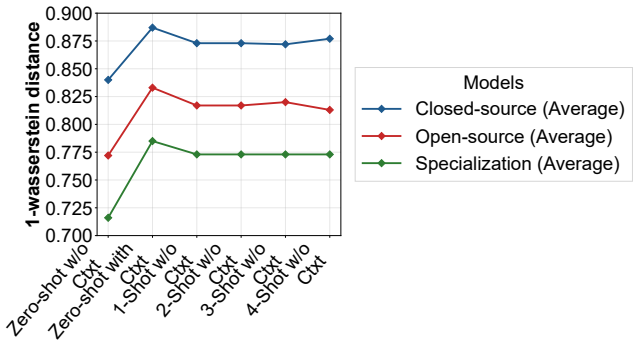

Figure 5: Contextual priming outperforms few-shot learning across all model categories, highlighting the importance of social context over example-based pattern matching.

ting. For the few-shot configuration, we supplied the model with examples predicted from Claude-3.7, where each example contained a poll, its distribution prediction, and the corresponding reasoning steps.

The results in Figure 5 are stark: contextual priming delivers a substantial performance boost across all models, often exceeding the gains achieved through few-shot learning. In contrast, the benefit from few-shot examples is unreliable. As the figure shows, performance does not reliably increase with the number of examples; for instance, the average 4-shot performance of open-sourced models is worse than that of 3-shot performance. This discrepancy suggests that the ability to situate a problem within its broader social context is a more critical and robust capability than simple pattern-matching from isolated examples.

5 Error Analysis

We conduct an error analysis to categorize prediction reasoning errors [1]. Using the LLM-as-a-judge, we identify three main categories of error type shown in Table 6.

Platform Mis-Adaptation. In the TikTok subscriptions case, the model applies a generic economic lens, predicting high tendency toward worth option based on the influencer economy. This reasoning ignores the context of Reddit, where the user tend to be negative of influencer monetization. The prediction is thus inverted from the ground truth which shows a much higher proportion of not worth it.

Cultural Misunderstanding. In the phone update case, models systematically misinterpret Chinese users' pragmatic spending attitudes, failing to capture the genuine cost-conscious preferences common among Weibo users, which often contradict Western-centric consumer marketing narratives. While model recognizes about the technical enthusiasm in Weibo, their predictions fail to reflect the cost-conscious approach that characterizes technology adoption for many Chinese social media users.

Temporal Dislocation. In the Twitter-to-X rebranding case, the model assumes that users would quickly adopt the new term. However, this prediction overlooks that Reddit users, in particular, demonstrated strong attachment to the familiar terminology, viewing continued use of the old name as both habitual behavior and subtle resistance. The model's temporal reasoning thus misaligned with the gradual and reluctant pace of real-world language adoption.

[1] We consider the individual poll's 1-Wass. < 0.8 as error.

Error Type	Case Study & Failed Reasoning
Platform Mis-adaptation	*Influencer Subscription Worth or Not (Reddit):* Failing to recognize Reddit attitude toward ad-centric monetization and user engagement patterns.
Cultural Misalignment	*New phone Buy or Wait (Weibo):* Assumed universal tech-enthusiasm drives upgrades, overlooking Chinese market's specific socio-economic factors and consumer behavior patterns.
Temporal Dislocation	*Calling it "X" vs. "Twitter" (Reddit):* Reasoning anchored to official rebrand timeline, underestimating persistent colloquial usage and real-world adoption resistance.

Table 6: Analysis of Claude 3.7 Sonnet prediction errors categorized by failure modes: **Platform Mis-Adaptation** (42.5%) – misapplying knowledge about specific platforms; **Cultural Misalignment** (36.6%) – ignoring local and cultural contexts; **Temporal Dislocation** (20.0%) – misunderstanding about the temporal information.

6 Conclusion

We introduce MindVote, the first benchmark for public opinion distribution prediction in social media. Our comprehensive evaluation demonstrates the critical importance of assessing models in naturalistic, context-rich environments. We argue that the path to socially intelligent AI requires enhancing a model's capacity for in-context reasoning. Our results demonstrate that models perform best when they can explicitly identify, weigh, and interpret the social cues present in the immediate context—a skill that requires flexible reasoning rather than memorized associations. MindVote provides the essential tool to guide and measure this necessary shift.

References

Anthis, J. R.; Liu, R.; Richardson, S. M.; Kozlowski, A. C.; Koch, B.; Evans, J.; Brynjolfsson, E.; and Bernstein, M. 2025. Llm so-

cial simulations are a promising research method. *arXiv preprint arXiv:2504.02234*.

Anthropic. 2025. Claude 3.7 Sonnet System Card.

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3): 337–351.

Banducci, S.; and Stevens, D. 2015. Surveys in context: how timing in the electoral cycle influences response propensity and satisficing. *Public Opinion Quarterly*, 79(S1): 214–243.

Binz, M.; Alaniz, S.; Roskies, A.; Aczel, B.; Bergstrom, C. T.; Allen, C.; Schad, D.; Wulff, D.; West, J. D.; Zhang, Q.; et al. 2025. How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, 122(5): e2401227121.

Bisbee, J.; Clinton, J. D.; Dorff, C.; Kenkel, B.; and Larson, J. M. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4): 401–416.

Boutyline, A.; and Soter, L. K. 2021. Cultural schemas: What they are, how to find them, and what to do once you've caught one. *American Sociological Review*, 86(4): 728–758.

Cao, Y.; Liu, H.; Arora, A.; Augenstein, I.; Röttger, P.; and Hershcovich, D. 2025a. Specializing Large Language Models to Simulate Survey Response Distributions for Global Populations. *arXiv preprint arXiv:2502.07068*.

Cao, Y.; Liu, H.; Arora, A.; Augenstein, I.; Röttger, P.; and Hershcovich, D. 2025b. Specializing Large Language Models to Simulate Survey Response Distributions for Global Populations. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3141–3154. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Chaplot, D. S. 2023. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*, 3.

Chen, G. H.; Chen, S.; Liu, Z.; Jiang, F.; and Wang, B. 2024. Humans or LLMs as the Judge? A Study on Judgement Bias. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8301–8327. Miami, Florida, USA: Association for Computational Linguistics.

Chen, Z.-S.; Wang, S.-L.; and Ma, Z. 2025. Large Language Models in Job Position Prediction: An Exploratory Study. *Available at SSRN 5098914*.

China Marketing Corp. 2024. Weibo Marketing in 2024: Essential Insights and Algorithms of Weibo Marketing. Entertainment-related topics make up at least 70% of the traffic on the platform.

DeepSeek-AI Team. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Doyle, E.; and Lee, Y. 2016. Context, context, context: Priming theory and attitudes towards corporations in social media. *Public Relations Review*, 42(5): 913–919.

DURMUS, E.; Nguyen, K.; Liao, T.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; Lovitt, L.; McCandlish, S.; Sikder, O.; Tamkin, A.; Thamkul, J.; Kaplan, J.; Clark, J.; and Ganguli, D. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. In *First Conference on Language Modeling*.

Elkjær, M. A.; and Wlezien, C. 2024. Estimating public opinion from surveys: the impact of including a "don't know" response option in policy preference questions. *Political science research and methods*, 1–17.

Google Cloud. 2025. Gemini 2.5 Pro.

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*.

Haerpfer, C.; Inglehart, R.; Moreno, A.; Welzel, C.; Kizilova, K.; Diez-Medrano, J.; Lagos, M.; Norris, P.; Ponarin, E.; and Puranen, B. 2022. World values survey wave 7 (2017-2022) cross-national data-set. *(No Title)*.

Han, F.; Yu, X.; Tang, J.; and Ungar, L. 2025. ZeroTuning: Unlocking the Initial Token's Power to Enhance Large Language Models Without Training. *arXiv preprint arXiv:2505.11739*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Kahan, D. M.; Jenkins-Smith, H.; and Braman, D. 2011. Cultural cognition of scientific consensus. *Journal of risk research*, 14(2): 147–174.

Karjus, A.; and Cuskley, C. 2024. Evolving linguistic divergence on polarizing social media. *Humanities and Social Sciences Communications*, 11(1): 1–14.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Krogstad, J.; Passel, J.; and Cohn, D. 2016. Pew Research Center. *US Border Apprehensions Of Families And Unaccompanied Children Jump Dramatically*.

Lees, A.; Tran, V. Q.; Tay, Y.; Sorensen, J.; Gupta, J.; Metzler, D.; and Vasserman, L. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 3197–3207.

Li, X.; Liu, M.; Gao, S.; and Buntine, W. 2023. A survey on out-of-distribution evaluation of neural NLP models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23. ISBN 978-1-956792-03-4.

López-Rabadán, P. 2021. Framing studies evolution in the social media era. Digital advancement and reorientation of the research agenda. *Social Sciences*, 11(1): 9.

Lu, Z.; Ding, K.; Zhang, Y.; Li, J.; Peng, B.; and Liu, L. 2021. Engage the Public: Poll Question Generation for Social Media Posts. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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)*, 29–40. Online: Association for Computational Linguistics.

Markus, H. R.; and Kitayama, S. 2014. Culture and the self: Implications for cognition, emotion, and motivation. In *College student development and academic life*, 264–293. Routledge.

McIntosh, I.; and Wright, S. 2019. Exploring what the notion of 'lived experience' offers for social policy analysis. *Journal of social policy*, 48(3): 449–467.

Meister, N.; Guestrin, C.; and Hashimoto, T. 2025. Benchmarking Distributional Alignment of Large Language Models. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 24–49. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Meta AI. 2025. The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation. Blog post.

Moon, S.; Abdulhai, M.; Kang, M.; Suh, J.; Soedarmadji, W.; Behar, E. K.; and Chan, D. 2024. Virtual Personas for Language Models via an Anthology of Backstories. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 19864–19897. Miami, Florida, USA: Association for Computational Linguistics.

OpenAI. 2025a. GPT-4.1.

OpenAI. 2025b. GPT-4o.

OpenAI. 2025c. OpenAI o3 and o4-mini System Card.

Panickssery, A.; Bowman, S.; and Feng, S. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37: 68772–68802.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Proferes, N.; Jones, N.; Gilbert, S.; Fiesler, C.; and Zimmer, M. 2021. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society*, 7(2): 1–14. Systematic analysis of 727 manuscripts using Reddit as data source, reveals Reddit's rich community structure with user-created and user-moderated subreddits varying considerably in content and norms.

Qi, J. 2024. The Impact of Large Language Models on Social Media Communication. In *Proceedings of the 2024 7th International Conference on Software Engineering and Information Management*, 165–170.

Qian, Y.; Chen, S.; Wu, C.; Yuan, K.; Du, Y.; Jiang, Y.; and Liu, Y. 2025. Large Language Models for Marketing Research: A Survey and New Perspectives.

Qiao, Y. 2023. How Does Weibo's" Super Topics" Enhance the Experience of Fans? *Lecture Notes in Education Psychology and Public Media*, 4: 270–278.

Qu, Y.; and Wang, J. 2024. Performance and biases of Large Language Models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1): 1–13.

Qwen Team. 2024. Qwen2.5: A Party of Foundation Models.

Radivojevic, K.; Clark, N.; and Brenner, P. 2024. Llms among us: Generative ai participating in digital discourse. In *Proceedings of the AAAI Symposium Series*, volume 3, 209–218.

Reveilhac, M.; and Morselli, D. 2024. Augmenting surveys with social media discourse on the workings of democracy from a cross-national perspective. *Frontiers in Political Science*, 6: 1385678.

Rothschild, D. M.; Brand, J.; Schroeder, H.; and Wang, J. 2024. Opportunities and risks of LLMs in survey research. *Available at SSRN*.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 29971–30004. PMLR.

Scarano, S.; Vasudevan, V.; Bagchi, C.; Samory, M.; Yang, J.; and Grabowicz, P. A. 2024. Analyzing and Estimating Support for US Presidential Candidates in Twitter Polls. *arXiv preprint arXiv:2406.03340*.

Shrivastava, A.; and Li, P. 2014. In defense of minhash over simhash. In *Artificial intelligence and statistics*, 886–894. PMLR.

Sinacola, E.; Pachot, A.; and Petit, T. 2025. Llms, Virtual Users, and Bias: Predicting Any Survey Question Without Human Data. *arXiv preprint arXiv:2503.16498*.

Singh, S.; Romanou, A.; Fourrier, C.; Adelani, D. I.; Ngui, J. G.; Vila-Suero, D.; Limkonchotiwat, P.; Marchisio, K.; Leong, W. Q.; Susanto, Y.; Ng, R.; Longpre, S.; Ruder, S.; Ko, W.-Y.; Bosselut, A.; Oh, A.; Martins, A.; Choshen, L.; Ippolito, D.; Ferrante, E.; Fadaee, M.; Ermis, B.; and Hooker, S. 2025. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 18761–18799. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Song, H.; Qi, Z.; Stasko, J.; and Yang, D. 2023. Understanding people's needs in viewing diverse social opinions about controversial topics. In *2023 IEEE 16th Pacific Visualization Symposium (PacificVis)*, 6–10. IEEE.

Suh, J.; Jahanparast, E.; Moon, S.; Kang, M.; and Chang, S. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761*.

Team, Gemma; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Tourangeau, R.; Singer, E.; and Presser, S. 2003. Context effects in attitude surveys: Effects on remote items and impact on predictive validity. *Sociological Methods & Research*, 31(4): 486–513.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, A.; Morgenstern, J.; and Dickerson, J. P. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 1–12.

Zaller, J.; and Feldman, S. 1992. A simple theory of the survey response: Answering questions versus revealing preferences. *American journal of political science*, 579–616.

Zhao, W.; Mondal, D.; Tandon, N.; Dillion, D.; Gray, K.; and Gu, Y. 2024. WorldValuesBench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of LREC-COLING 2024*, 17696–17706. Torino, Italia: ELRA and ICCL.

Zhou, X.; Nie, Y.; and Bansal, M. 2022. Distributed NLI: Learning to Predict Human Opinion Distributions for Language Reasoning. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 972–987. Dublin, Ireland: Association for Computational Linguistics.

Zhu, Y. 2024. Privacy cynicism and diminishing utility of state surveillance: A natural experiment of mandatory location disclosure on China's Weibo. *Big Data & Society*, 11(2): 20539517241242450.