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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large Language Models (LLMs) are increasingly used to predict public opinion, but are typically evaluated on structured surveys, which strip away the rich social, cultural, and temporal context of real-world discourse. This misalignment creates a critical evaluation gap. To address this, we introduce MindVote, the first benchmark for public opinion prediction grounded in authentic social media. MindVote consists of 3,918 naturalistic polls from Reddit and Weibo, spanning 23 topics and enriched with detailed contextual metadata. Our evaluation of 15 LLMs on MindVote reveals that general-purpose models outperform the models that fine-tuned on survey, highlighting the importance of in-context reasoning. MindVote provides a robust evaluation framework towards developing more socially intelligent AI.

1 Introduction

A core application for Large Language Models (LLMs) is to predict public opinion distributions, serving as a scalable alternative to costly surveys [20, 2]. However, the prevailing evaluation paradigm relies on traditional structured surveys, a methodology misaligned with context-rich digital environments like social media where opinions are formed and expressed [17]. This reliance on survey-based evaluation creates critical gaps. **Topic Imbalance:** Existing benchmarks are skewed toward formal topics like politics, unlike real-world discourse where entertainment drives 70% of Weibo traffic and Reddit's largest communities focus on gaming [4, 16]. **Cultural Homogeneity:** Benchmarks often use Western-centric questions, testing linguistic translation rather than genuine cultural understanding. **Context Poverty:** Surveys systematically remove platform norms, temporal events, and community discourse—critical signals that prime opinions on social media [6].

To bridge these gaps (shown in Figure 1), we introduce **MindVote**, the first benchmark for public opinion prediction grounded in authentic social media discourse. We construct a dataset of 3,918 polls from Reddit and Weibo, enriched with social context annotations. Using MindVote, we benchmark 15 leading LLMs and find that enhancing a model's capacity for social-context reasoning is more effective than fine-tuning on context-stripped data.

2 Related Work

Previous benchmarks for opinion distribution prediction, such as the U.S.-focused OpinionQA [18] and SubPop [20], and cross-cultural efforts like GlobalOpinionQA [7] and WorldValuesBench [22], primarily rely on structured survey data. While valuable, these benchmarks are limited by demographic-value pairings and a Western-centric lens, abstracting away the naturalistic cultural and social contexts essential for authentic opinion prediction.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

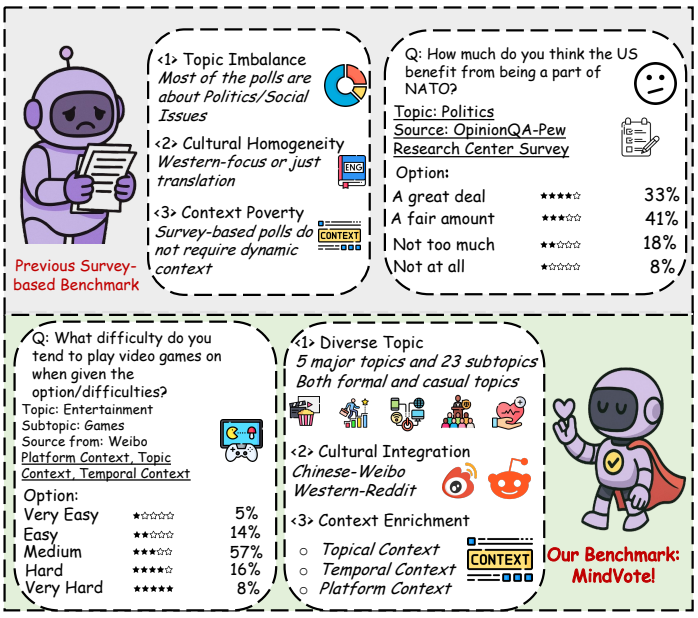

Figure 1: MindVote benchmark addresses three key limitations of previous survey-based approaches. Our benchmark provides diverse topics, cultural integration, and rich contextual metadata, overcoming the topic imbalance, cultural homogeneity, and context poverty of traditional survey datasets.

3 Benchmarking Setup

3.1 Dataset Construction

We constructed MindVote from 7,648 raw polls collected from Reddit and Weibo (2019-2025). A four-stage quality control pipeline involving preprocessing, content filtering for toxicity and duplicates, social relevance validation, and structural filtering yielded 1,959 high-quality core polls. We classified these polls into 5 major and 23 sub-topics and manually annotated them with rich social context (platform norms, topical discourse, and temporal events). Finally, a cross-cultural augmentation process involving machine translation and human validation resulted in our final dataset of 3,918 polls. Detailed construction is in Appendix A.1.

3.2 Experiment Design

The task is to predict the opinion distribution for a given poll question and its associated social context. We evaluate 15 LLMs, including closed-source models (e.g., Claude-3.7-Sonnet [1], GPT-4o [13], Gemini-2.5-Pro [8]), open-source models (e.g., Deepseek-R1 [5], Llama-3-70B [21]), and two survey-specialized models from Suh et al. [20]. We use four primary metrics: 1-Wasserstein Distance (1-Wass.), 1-KL Divergence, Spearman's Rank Correlation, and One-hot Accuracy. Details are in Appendix A.2.

4 Results and Analysis

4.1 Overall Performance and Specialization Pitfall

Our evaluation, summarized in Table 1, shows that closed-source models like o3-medium achieve the highest performance, while Deepseek-R1 leads among open-source models. A critical finding is the **survey-based specialization pitfall**: models fine-tuned on structured survey data (e.g., SubPop-Llama-3-70B) consistently underperform their general-purpose, instruction-tuned counterparts. For instance, SubPop-Llama-2-13B's 1-Wass. score drops by 3.3 percentage points compared to the base Llama-2-13B. This suggests that specializing on context-stripped data degrades a model's ability to generalize to authentic, context-rich 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 1: 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.**

4.2 The Importance of Social Context

We investigate the specialization pitfall by analyzing the role of context. An ablation study confirms that social context is a critical signal; removing it degrades performance across all models, with specialized models suffering the most severe drops. Furthermore, as shown in Figure 2, performance negatively correlates with contextual complexity (e.g., context length, language informality), with specialized models again showing the most brittleness. This highlights their over-reliance on the simple, formal structures found in survey data. Finally, we find that zero-shot contextual priming consistently yields larger performance gains than few-shot learning, demonstrating that the ability to situate a problem in its social environment is more robust than pattern-matching from isolated examples (Figure 3).

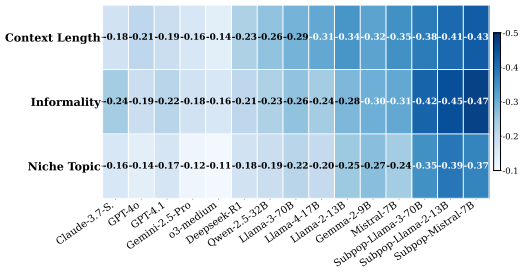

Figure 2: Correlation between model 1-Wass. performance and complexity dimensions. Survey-specialized models show strong brittleness.

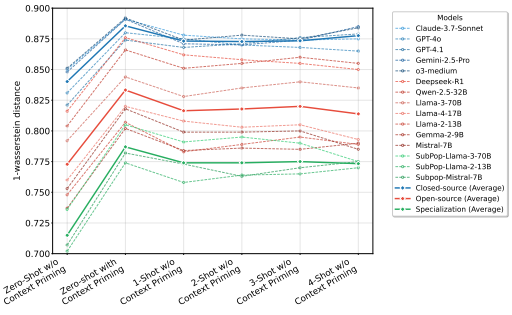

Figure 3: Contextual priming outperforms few-shot learning, highlighting the importance of social context over example-based pattern matching.

Model	w/o Plat.	w/o Topic.	w/o Temp.	No Ctx.
Closed-source Models				
o3-medium	-3.03	-2.31	-1.02	-4.13
Gemini-2.5-Pro	-2.37	-3.30	-0.98	-4.34
Claude-3.7-Sonnet	-4.46	-3.91	-1.39	-5.53
GPT-4o	-5.75	-3.39	-0.87	-4.92
GPT-4.1	-4.92	-3.36	-1.57	-5.30
Open-source Models				
Deepseek-R1	-4.19	-3.55	-1.04	-5.98
Qwen2.5-32B	-4.42	-3.81	-0.88	-6.24
Llama-3-70B	-4.82	-4.78	-2.11	-5.19
Llama-4-17B	-5.80	-5.33	-1.01	-6.08
Llama-2-13B	-4.95	-5.26	-1.37	-5.92
Gemma-2-9B	-5.37	-4.74	-1.26	-6.47
Mistral-7B	-5.92	-4.84	-1.46	-6.55
Specialization Models				
SubPop-Llama-3-70B	-6.82	-6.68	-1.88	-6.95
SubPop-Llama-2-13B	-6.94	-6.11	-2.10	-7.23
SubPop-Mistral-7B	-6.89	-6.41	-2.96	-7.54
Average	-5.12	-4.52	-1.46	-5.91

Table 2: Opinion distiribution prediction performance degradation from the full-context baseline. All scores represent the drop in 1-Wasserstein Distance (%). Abbreviations: Plat. (Platform), Topic. (topical), Temp. (Temporal), Ctx. (Context).

4.3 Error Analysis

We consider the individual poll's 1-Wass. < 0.8 as error. The prediction failures reveals three primary error categories: **Platform Misadaptation** (42.5%), where models misjudge platform-specific user norms (e.g., Reddit's attitude toward monetization); **Cultural Misalignment** (36.6%), where models overlook culture-specific reasoning (e.g., pragmatic spending habits on Weibo); and **Temporal Dislocation** (20.0%), where models fail to account for the real-world pace of change (e.g., the slow adoption of the "X" branding over "Twitter") shown in Table 3.

Error Type	Case Study & Failed Reasoning
Platform Misadaptation	*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 3: Analysis of Claude 3.7 Sonnet prediction errors categorized by failure modes.

5 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

[1] Anthropic. Claude 3.7 sonnet system card, 2025.

[2] Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. *arXiv preprint arXiv:2502.07068*, 2025.

[3] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[4] China Marketing Corp. Weibo marketing in 2024: Essential insights and algorithms of weibo marketing, April 2024. Entertainment-related topics make up at least 70% of the traffic on the platform.

[5] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[6] Evan Doyle and YoungAh Lee. Context, context, context: Priming theory and attitudes towards corporations in social media. *Public Relations Review*, 42(5):913–919, 2016.

[7] Esin DURMUS, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*, 2024.

[8] Google Cloud. Gemini 2.5 pro, 2025.

[9] Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 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*, pages 3197–3207, 2022.

[10] Zexin Lu, Keyang Ding, Yuji Zhang, Jing Li, Baolin Peng, and Lemao Liu. Engage the public: Poll question generation for social media posts. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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 29–40, Online, August 2021. Association for Computational Linguistics.

[11] Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. Benchmarking distributional alignment of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *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)*, pages 24–49, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[12] Hieu Trung Nguyen, Bao Nguyen, Binh Nguyen, and Viet Anh Nguyen. Task-driven layerwise additive activation intervention. *arXiv preprint arXiv:2502.06115*, 2025.

[13] OpenAI. Gpt-4o, 2025.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[15] John Pavlopoulos and Aristidis Likas. Distance from unimodality for the assessment of opinion polarization. *Cognitive Computation*, 15(2):731–738, 2023.

[16] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, 7(2):1–14, 2021. 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.

[17] Paolo Riva, Nicolas Aureli, and Federica Silvestrini. Social influences in the digital era: When do people conform more to a human being or an artificial intelligence? *Acta Psychologica*, 229:103681, 2022.

[18] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[19] Anshumali Shrivastava and Ping Li. In defense of minhash over simhash. In *Artificial intelligence and statistics*, pages 886–894. PMLR, 2014.

[20] Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761*, 2025.

[21] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[22] Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia, May 2024. ELRA and ICCL.

[23] Xiang Zhou, Yixin Nie, and Mohit Bansal. Distributed NLI: Learning to predict human opinion distributions for language reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 972–987, Dublin, Ireland, May 2022. Association for Computational Linguistics.

A Full Benchmark Setup

A.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.

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 [10] 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 [19], 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 [9].
3. **Social Relevance Validation:** Human verification of voting patterns (removing polls ≤ 100 votes for social relevance) and verification of poll and vote existence in all three 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 [18]. We systematically excluded polls with purely categorical options (e.g., preference choices among unordered alternatives). This filtering employed LLM-as-a-judge (DeepSeek-R1 [5]), 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 shown in Figure 4. 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 [3]. 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.

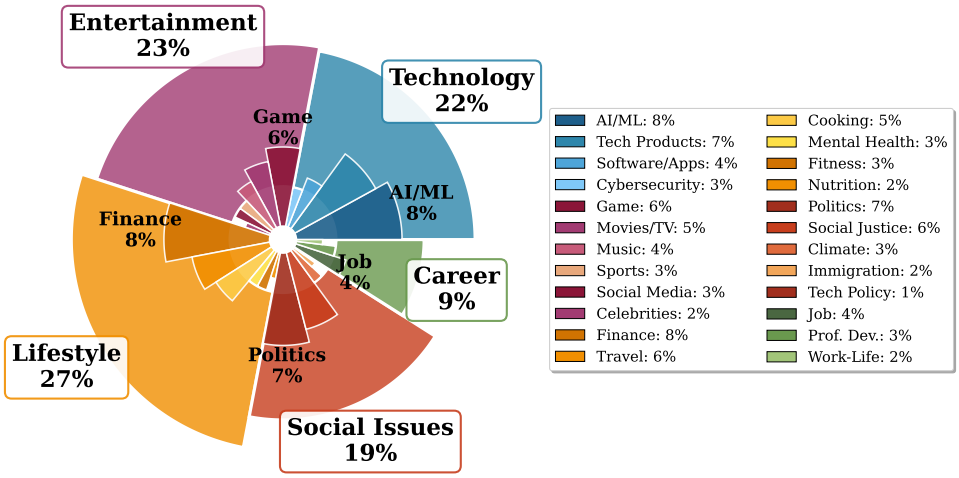

Figure 4: Distribution of topics in the MindVote dataset across five major topics and 23 subtopics. Percentages indicate the relative frequency of each subtopic.

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.
Temporal Context	AI-driven layoff spike fuels job insecurity.

Table 4: An example poll from our MindVote dataset, demonstrating the structure of the question, options, and associated social context provided for model evaluation.

We manually annotate each poll includes rich metadata for social context enrichment: general platform context (e.g., user statistics and user behaviors), topical context (topic-specific discourse patterns), and temporal context for each poll (social events and backgrounds related to the poll when created).

Cross-Cultural Augmentation. We created a parallel bilingual corpus for all three platforms through machine translation with rigorous quality control to demonstrate both linguistic and cultural effects: 10% back-translation validation (BLEU > 35 [14]), 5% native speaker rewriting, and expert review (Fleiss' $\kappa = 0.51$). 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 three layers of social context: general platform context, topical context within that platform, and the specific temporal context at the time the poll was created. To ensure broad accessibility and ease of use, the entire dataset is provided in both CSV and JSON formats. Table 4 shows an example with its simplified metadata keywords.

A.2 Experiment Design

Pipeline. All primary evaluations use a greedy decoding strategy (temperature=0) with default hyperparameter settings under zero-shot with context annotation. 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. [20].

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.)** [18, 11, 20]. 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)** [23, 15], a non-parametric measure of how well the predicted ranking of options matches the true ranking of vote shares; **1 - KL Divergence** [11, 12], which quantifies the information loss when using the model's predicted distribution to approximate the ground truth; and **One-hot Accuracy** [23, 18, 20], 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 [20]. 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.