Can LLMs Simulate What the Public Is Thinking About? Evidence from Open-Ended Political Surveys

Anonymous ACL submission

Abstract

We evaluate whether large language models can simulate open-ended political survey responses using a weekly 2024 U.S. election survey paired with contemporaneous news-production data. Across 12 models and multiple context strategies, we compare real and synthetic response distributions using Jensen–Shannon divergence over topic and sentiment labels. News context sharply improves aggregate alignment, suggesting that access to current information is essential for simulating time-sensitive political discourse. However, aggregate gains obscure important failures: context-grounded models overtrack news coverage, diverge on low-frequency topics, and remain uneven across demographic subgroups. These results show that synthetic publics can appear successful under aggregate metrics while failing in ways that matter for open-ended survey measurement.

1 Introduction

Can large language models (LLMs) simulate what the public is thinking about? As LLMs become increasingly capable of producing fluent, personalized text, researchers have begun to explore whether they can serve as synthetic respondents in surveys, experiments, and public opinion research. This possibility is especially tempting for open-ended survey questions, where collecting and coding human responses is costly. If LLMs could approximate the distribution of real responses, they might offer a scalable way to study political attention, test survey instruments, or understand public discourse in rapidly changing information environments.

Open-ended political survey responses, however, are difficult to simulate because they do not simply reflect what information is available. During an election cycle, respondents are exposed to a constantly shifting media environment, but the topics they choose to mention are filtered through attention, memory, personal relevance, partisan interpretation, and social discussion. Media coverage can shape which issues citizens perceive as important (McCombs and Shaw, 1972; Iyengar and Kinder, 1987), but individuals vary in what political information they encounter and attend to in the media (Prior, 2007; Stroud, 2010). Some heavily covered stories may therefore receive little attention from respondents, while other topics may become memorable despite receiving less news coverage. As a result, the distribution of what respondents mention in open-ended answers may not mirror the distribution of what the news covered.

This distinction matters for evaluating LLMs as synthetic publics. A natural way to improve LLM simulations is to provide models with contemporaneous news context. Without such context, models may not know what happened in a given week and are unlikely to reproduce time-sensitive political discourse. But providing news context solves only part of the problem. It tells the model what was covered in the news, not what people noticed, remembered, or chose to discuss. If human responses do not mirror the distribution of news coverage, then a context-grounded LLM that closely follows the news may appear informed while still failing to match the distribution of human responses. In that case, the model is reproducing the media agenda rather than the topics respondents actually chose to discuss.

This question is increasingly practical. LLMs can be used by researchers to generate synthetic survey responses through so-called "silicon sampling", or LLMs can be used by respondents to augment or replace their answers (there is no evidence this is widespread (Gordon et al., 2026), but the threat is real (Westwood, 2025) especially if researchers do not deploy mitigation methods (Veselovsky et al., 2025)). Prior work has evaluated LLMs as proxies for human responses on closed-ended survey

items (Argyle et al., 2023; Bisbee et al., 2024), behavioral experiments (Aher et al., 2023; Horton et al., 2023), public opinion benchmarks (Durmus et al., 2024; Santurkar et al., 2023; Lee et al., 2024; Qu and Wang, 2024; González-Bustamante et al., 2025), and open-ended responses (Hämäläinen et al., 2023; Ma et al., 2025). Other work develops fine-tuned or specialized survey simulators (Kim and Lee, 2026; Cao et al., 2025; Lyman et al., 2025). The evidence remains mixed: model responses can appear plausible, but they may be miscalibrated, low-diversity, sensitive to prompt format, or skewed toward particular subgroups (Parikh et al., 2026; Tjuatja et al., 2024; Dominguez-Olmedo et al., 2024; Boelaert et al., 2025). These findings reinforce broader cautions against treating LLM outputs as replacements for human participants (Dillion et al., 2023; Rothschild et al., 2024; Lin, 2025). Closest to our setting, Ma et al. (2025) study longitudinal German-language open-ended survey simulation. Our contribution pairs a longitudinal 2024 U.S. election survey with contemporaneous news-production data and uses that pairing to test how context changes the response distribution, including whether any observed performance gains persist under temporal, subgroup, and residual-tail diagnostics.

We study this problem using weekly open-ended survey responses collected during the 2024 U.S. election cycle. Respondents were asked to describe the key political events or topics of the previous week in their own words. We compare these real responses to synthetic responses generated by 12 LLMs under multiple context strategies: no context, uniform news context, simulated media-diet context, self-reported media-diet context, and one-week-lagged variants of the context conditions. We label both real and synthetic responses by topic and sentiment, then evaluate whether synthetic responses reproduce real response distributions overall, over time, across demographic subgroups, and in the long tail.

First, do synthetic responses match real responses in the aggregate, after pooling across all respondents and survey waves? Second, do they match week-to-week changes in public attention as the campaign evolves? Third, do they match different demographic groups equally well, or does aggregate fit mask subgroup-specific failures? Finally, do they reproduce the long tail of open-ended responses: the less common and unexpected topics that fall outside the dominant categories?

Our results show that news context is the dominant driver of aggregate alignment: without context, models diverge sharply from real respondents, while contemporaneous news context brings synthetic topic and sentiment distributions much closer to the real response distributions. Yet this improvement does not imply faithful simulation. Because we observe both weekly survey responses and contemporaneous news production, we can show that real respondents do not simply follow the distribution of news coverage. Context-grounded models can become better informed about the week's events while still failing to reproduce the topics humans actually choose to discuss.

Two limitations are especially important. First, synthetic responses continue to diverge in the long tail of less common topics, even though surfacing such topics is a central motivation for asking open-ended questions. Second, aggregate alignment masks subgroup-level mismatches: context-grounded models match some demographic groups better than others, with larger gaps for women, non-white respondents, and respondents without a college degree.

Taken together, these findings suggest that news-grounded LLMs can improve aggregate fit, but should not be treated as substitutes for measured public opinion.

2 Data

Survey Details We conducted a weekly cross-sectional Pollfish survey from February to November 2024 during the U.S. presidential election. Each wave sampled roughly 250 to 500 adult respondents without quotas. Respondents were not longitudinally tracked, so limited repeat participation is possible.

This paper uses the survey's open-ended political-news responses (Q6–Q8), respondent demographics and party identification, and source-consumption fields for self-reported context. The open-ended module asked three questions: "In 1 or 2 sentences: What is your overarching takeaway about politics that best explains the key events or topics of the last week?", "In 1 or 2 sentences: What happened in politics in the last week that best fits with your main takeaway? How so?", and "What is the most memorable thing related to politics you heard last week?". For real responses, we concatenate all available Q6–Q8 text, in this order, into one open-text field before classification.

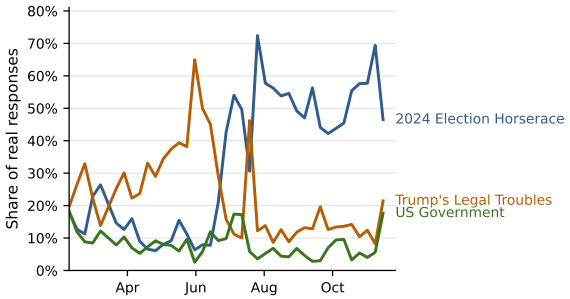

Figure 1: Weekly share of real responses assigned to the three most frequent fixed main-topic labels, excluding *Other* from topic selection. Shares are computed over all classified real responses in each wave.

The broader survey also included structured exposure, attitude, vote-intention, party-feeling, and source-consumption items. Except for the respondent fields used in prompts and context assignment, those structured responses are not used as generation context or evaluation targets in the main experiments; we summarize them only as instrument context in Appendix A.1.

Following prior open-ended synthetic-survey work that visualizes the empirical target distribution (Ma et al., 2025), Figure 1 summarizes the top real-response topics over time. The sharp mid-campaign shift motivates evaluating synthetic responses by survey week and by context timing, rather than only against a static pooled distribution.

Appendix A.1 reports the raw, filtered, and classified row counts used in the analysis.

Data Preprocessing Survey responses were stored in CSV format. We concatenated the available Q6–Q8 open-ended responses into a single text field for classification and for comparison with synthetic responses. Synthetic responses were processed the same way: each model answered the same three open-ended items, and the generated answers were concatenated in the same order before labeling. The main respondent fields used in prompts, subgroup analyses, and poststratification checks were age, gender, education, race, party, urbanicity, and state. Party identification was standardized into Republican, Democratic, and Independent using a two-stage, seven-point party-identification item. For the poststratification checks, we formed demographic cells over these fields and matched them to a contemporary voter file based on 2020 voting data (Ghitza and Gelman, 2013; Wang et al., 2015). For self-reported context experiments, Q20 source-consumption selections were used to weight news sources for each respondent or demographic pool.

News Production Data We obtained the data from the Media Bias Detector project (Wang et al., 2025), which captured periodic homepage snapshots from major U.S. news outlets spanning a range of ideological perspectives. These include the HuffPost, The Guardian, CNN, USA Today, The Washington Post, The New York Times, The Associated Press, The Wall Street Journal, Fox News, and Breitbart among others.[1] The data contained articles automatically classified by topic using a GPT-4o-based annotation pipeline. To ensure reliability, the Media Bias Detector routinely has undergraduate research assistants validate a sample of the topic assignments.

3 Experimental Setup

To compare to our survey dataset, we generated synthetic responses from 12 open-weight models under seven conditions: a no-context baseline, uniform news context, simulated media-diet context, self-reported news context, and one-week-lagged variants of each context condition. For each respondent-model-condition tuple, the model answered the same three open-ended survey items described in Section 2. We then concatenated those answers into the same combined response field used for the real survey data. All combined responses were labeled for topic and sentiment using Qwen3-Next-80B, and validated against GPT-4o. Exact model identifiers and artifact citations are listed in Appendix A.7. Full generation templates and inference settings are reported in Appendices A.2 and A.3. Prompt length and style sensitivity checks are reported in Appendix A.18.

Context Insertion Strategies Each context condition used the same demographic, date, and survey-question template for each of the three open-ended items. When context was enabled, the prompt inserted compact news snippets consisting of publisher, title, and a sentence-bounded context field; full article bodies were not inserted unless they fit within the context limit. We compare four levels of context availability:

- **No context**: the model sees only the respondent demographic, survey date, and current survey question.

[1] Table 6 contains the full publisher list and per publisher article counts.

- **Uniform context**: every respondent in a survey wave receives the same selected articles from the relevant news window.

- **Simulated context**: respondents receive article samples drawn from a simple simulated media diet based on demographic and party fields.

- **Self-reported context**: respondents receive article samples weighted by their reported news-source consumption; in the lagged version, weights come from previous-wave demographic pools rather than same-wave individual reports.

For each non-baseline strategy, the current condition uses the survey-wave news window, operationalized as the seven days ending at the wave date; the lagged condition shifts that window back by one week. Appendix A.6 gives the full selection and formatting algorithms.

Evaluation Metric Following prior synthetic-survey evaluation (Ma et al., 2025), we operationalize response alignment using Jensen-Shannon divergence (JSD) between the empirical label distribution for real responses, p, and synthetic responses, q. With $m = (p + q)/2$, we compute

$$\mathrm{JSD}(p, q) = \frac{1}{2} D_{\mathrm{KL}}(p \| m) + \frac{1}{2} D_{\mathrm{KL}}(q \| m),$$

using base-2 logarithms (Lin, 1991). JSD is bounded in $[0, 1]$, with 0 indicating identical distributions, so lower values indicate closer agreement. We compute JSD at aggregate, weekly, and subgroup levels, and compare generation conditions using Holm-corrected contrasts and bootstrap confidence intervals (Holm, 1979; Efron and Tibshirani, 1994). News context is the primary driver of response quality.

Response Classification We classified each combined open-ended response for sentiment and main topic using `Qwen/Qwen3-Next-80B-A3B-Instruct`. Sentiment was coded as positive, negative, or neutral; main topic was coded into 14 fixed political categories plus *Other*. Responses assigned to *Other* received a second open-ended topic and subtopic label, used only for the long-tail analysis. Full prompts are reported in Appendix A.4.

LLM-based annotation can be competitive for social-science text coding, but prior work emphasizes validation and careful treatment of measurement error (Gilardi et al., 2023; Pangakis et al., 2023; Egami et al., 2023). We use this logic for a different purpose: the labels are an operational measurement layer, not ground truth about public beliefs. The key design choice is consistency. All real and synthetic responses are passed through the same classifier, prompts, and label taxonomy, so the main comparisons ask whether human and model responses induce similar coded distributions under a shared measurement process. This does not remove classifier bias, but it makes relative comparisons across generation conditions more interpretable. We validate the primary classifier against GPT-4o on a separate sample and report agreement diagnostics in Appendix A.5.

4 Results

We report four results. First, news context sharply reduces aggregate divergence between human and synthetic responses. Second, aggregate scores hide important week-to-week failures. Third, demographic gaps remain even when context improves overall fit. Fourth, context changes the residual long tail of responses coded as *Other*.

Aggregate JSD. Main-topic JSD drops from 0.385 without context to 0.087–0.094 with news context across the 12-model balanced comparison (Figure 2, Table 1).[2] At this aggregate level, however, the three context strategies are difficult to distinguish: uniform, simulated, and self-reported context produce similar pooled JSD once averaged over waves and models. This suggests that most of the aggregate gain comes from access to contemporaneous news, not from closely matching individual media diets. Model scale matters less than access to current context. The no-context models have no access to current events and are unsurprisingly poorly aligned with survey answers to time-dependent questions. Model-level aggregate and weekly results are reported in Appendix A.12.

Weekly JSD. Lagging news context by one week has little aggregate effect, but weekly trajectories show that similar averages can mask different temporal errors (Figure 3). This pattern is consistent with the agenda-setting problem introduced in Sec-

[2] Appendix A.14 reports category-level mean absolute error as a complementary diagnostic.

Task	Baseline, no context	Uniform context		Simulated context		Self-reported context	
		Current	Lagged	Current	Lagged	Current	Lagged
Sentiment	0.159 [0.117,0.201]	0.081 [0.043,0.119]	0.081 [0.043,0.119]	0.078 [0.049,0.108]	0.079 [0.049,0.108]	0.079 [0.049,0.109]	0.079 [0.048,0.110]
Main topic	0.385 [0.294,0.476]	0.093 [0.078,0.107]	0.094 [0.077,0.111]	0.089 [0.082,0.097]	0.091 [0.083,0.098]	0.088 [0.077,0.100]	0.087 [0.076,0.097]

Table 1: Pooled aggregate JSD by context strategy. Cells report the mean across model runs with 95% jackknife confidence intervals over model runs.

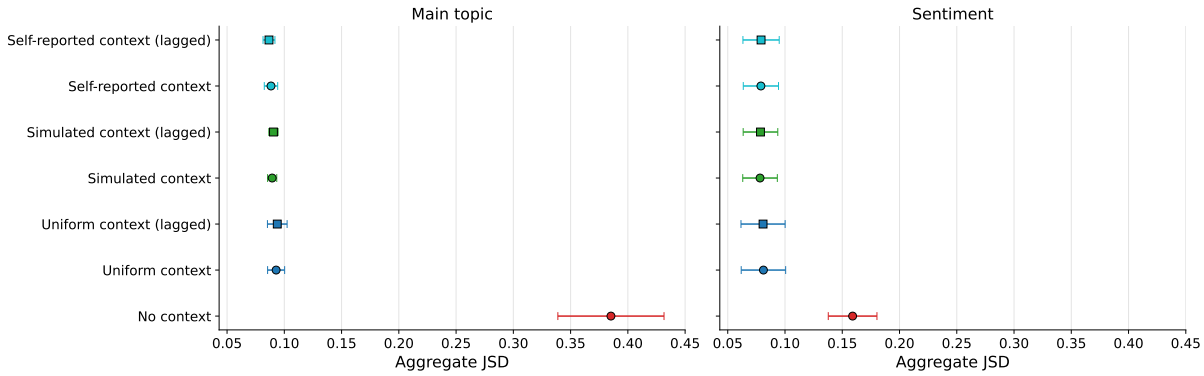

Figure 2: Dot-plot view of aggregate JSD by task and context condition, averaged across 12 model runs; whiskers show standard errors across runs.

tion 2. Models can be conditioned on contemporaneous news, but real responses reflect what respondents noticed and chose to report, not simply what was available in the news. The compact diagnostic in Table 2 shows this mismatch for the largest substantive article–human gaps. Crime is 15.9 percentage points more common in articles than in human responses, while Trump legal coverage is 20.4 points more common in human responses than in the article pool. Context-conditioned responses move toward the article distribution for crime and election horserace coverage, but not for every topic. The full appendix diagnostic adds article shares and all mapped topics (Table 13). We therefore interpret the time-series errors as evidence that context improves factual currency without fully modeling public salience (McCombs and Shaw, 1972).

Demographic Gaps. Demographic gaps persist but vary by context strategy. Across context-conditioned runs, women, non-white respondents, and respondents without a college degree usually have higher main-topic JSD than their comparison groups (Table 3). In this contrast table, stars mark gaps that survive Holm-Bonferroni correction across the pre-specified contrast family; unstarred gaps are descriptive even when their unadjusted intervals exclude zero. The selected subgroup table in the appendix gives the corresponding absolute levels: under self-reported lagged context, for example, JSD remains higher for women than men (0.095 vs. 0.078), for non-white than white respondents (0.096 vs. 0.087), and for respondents 55+ than respondents under 55 (0.110 vs. 0.082; Table 14). Age gaps are context-dependent: uniform context slightly favors respondents over 55, while simulated and self-reported context produce higher divergence for the 55+ group. Partisan gaps are largest under uniform context and shrink under simulated and self-reported context, though Independents remain somewhat less well matched. Full model-by-subgroup tables for each context condition are reported descriptively in Appendix A.16 (Tables 15–21); they show that the subgroup patterns are not driven by a single model. A direct post-stratification check for self-reported lagged context points in the same direction: weighting responses to the target demographic distribution increases main-topic JSD for every model by 0.004–0.025 (Table 10).

Long-Tail Distortion. Long-tail diagnostics reveal a tradeoff in the open-ended residual category. Real respondents are much more likely to be classified as *Other*, but most of those cases are non-substantive: *Other* covers 17.1% of real responses, while the filtered substantive tail covers 5.9% after removing *No answer* and exact fixed-topic overlaps (Table 4). Simulated and self-reported context produce lower *Other* rates (8.2–8.6%), but because

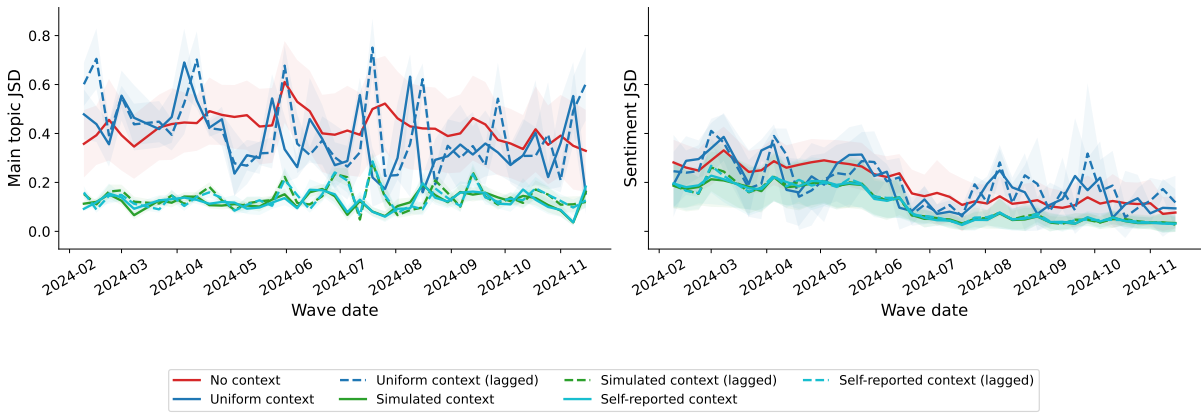

Figure 3: Weekly JSD between real and synthetic response distributions, averaged across 12 model runs by context condition; faint bands show cross-model standard deviation.

Topic	Article - human	Article - no context	Article - current	Article - lagged
	pp.	pp.	pp.	pp.
Trump's Legal Troubles	-20.4 [-21.0, -19.8]	-3.7 [-3.8, -3.5]	-18.7 [-18.8, -18.6]	-18.8 [-18.9, -18.6]
Crime	+15.9 [+15.6, +16.2]	+16.6 [+16.3, +16.8]	+11.2 [+10.9, +11.5]	+11.0 [+10.8, +11.3]
2024 Election Horserace	-9.9 [-10.6, -9.1]	+7.6 [+7.3, +7.9]	-1.9 [-2.2, -1.6]	-0.5 [-0.8, -0.2]
US Economy	+6.5 [+6.2, +6.8]	-8.0 [-8.2, -7.7]	+3.1 [+2.9, +3.3]	+3.0 [+2.8, +3.2]
Israel-Gaza War	+3.4 [+3.1, +3.7]	+6.3 [+6.1, +6.4]	-1.0 [-1.2, -0.8]	-0.9 [-1.1, -0.8]
Education	+3.1 [+2.9, +3.3]	+1.2 [+1.0, +1.3]	+2.9 [+2.7, +3.0]	+2.9 [+2.7, +3.0]

Table 2: Largest substantive article–human topic-share gaps. Cells report article share minus response share in percentage points with label-level bootstrap 95% confidence intervals. Current and lagged columns pool the uniform, simulated, and self-reported context strategies across 12 models.

nearly all of those labels are substantive, their filtered tails are larger than the real one (7.6–8.5% of all responses) and contain roughly 478–534 topics per condition on average. The content shift is not driven mainly by rare named events. The largest gaps are broad residual topics: *Politics* accounts for 27.4% of the filtered real tail but only 13.0–14.0% under simulated and self-reported context, while *Sports* rises from 1.0% in real responses to 11.0–12.6% in those same synthetic tails. Arts, entertainment, and celebrity labels show the same pattern, taking up substantially more synthetic mass than real mass. More event-specific misses still matter, but they are smaller in the filtered tail. TikTok-related labels account for 1.6% of the filtered real tail and at most 0.14% of context-conditioned synthetic tails; mentions of LGBTQ+ policy issues[3] make up 0.36% of the filtered real tail but at most 0.02% of synthetic context-conditioned tails (Figure 8). The rank-frequency curves in Figure 9 show the same pattern. Tail length increases from no context to uniform, simulated, and self-reported context, but longer tails are not necessarily more human-like.

5 Conclusion

This study evaluates whether LLMs can simulate open-ended political survey responses during a fast-moving election cycle. We find that contemporaneous news context substantially improves aggregate alignment between synthetic and real responses. This result is important: models without current information are poorly suited to simulating time-sensitive political discourse, and grounding them in news context makes their topic and sentiment distributions much closer to those of real respondents.

At the same time, our results show why aggregate alignment is an incomplete standard for evaluating synthetic publics. Because our data include both weekly human responses and contemporaneous news production, we can distinguish between matching the news environment and matching what respondents actually chose to mention. Context-grounded models often become better informed about the week's events, but they do not fully reproduce the distribution of human responses and the details that make open-ended survey data valuable.

[3]Here, LGBTQ+ issues refers to policy debates and opinions involving LGBTQ+ communities, not respondent identity.

Contrast	Baseline, no context	Uniform context		Simulated context		Self-reported context	
		Current	Lagged	Current	Lagged	Current	Lagged
Women - Men	+.018 [+.004,+.033]	+.010* [+.005,+.015]	+.011* [+.005,+.017]	+.014* [+.011,+.016]	+.015* [+.012,+.019]	+.016* [+.011,+.020]	+.017* [+.013,+.022]
55+ - under 55	+.013 [-.014,+.041]	-.018* [-.027,-.010]	-.010 [-.021,+.002]	+.029* [+.023,+.035]	+.028* [+.022,+.033]	+.030* [+.022,+.038]	+.028* [+.021,+.035]
Non-white - White	+.016 [-.008,+.041]	+.010* [+.007,+.012]	+.009* [+.006,+.011]	+.009* [+.006,+.012]	+.007* [+.004,+.009]	+.006* [+.004,+.009]	+.009* [+.006,+.011]
No college - College	-.011 [-.031,+.010]	+.019* [+.014,+.023]	+.016* [+.010,+.021]	+.016* [+.012,+.019]	+.015* [+.011,+.020]	+.013* [+.009,+.016]	+.013* [+.010,+.015]
Democrat - Republican	+.016 [-.035,+.067]	-.036* [-.049,-.022]	-.026* [-.035,-.017]	+.012* [+.006,+.018]	+.014* [+.007,+.020]	+.014* [+.008,+.019]	+.010* [+.004,+.016]
Independent - Republican	-.010 [-.080,+.061]	+.039* [+.030,+.048]	+.041* [+.033,+.049]	+.022* [+.014,+.030]	+.019* [+.012,+.026]	+.023* [+.017,+.029]	+.019* [+.010,+.027]
Independent - Democrat	-.026 [-.067,+.015]	+.074* [+.067,+.082]	+.067* [+.059,+.075]	+.010 [+.001,+.019]	+.006 [-.002,+.013]	+.009 [+.001,+.017]	+.008 [-.002,+.019]

Table 3: Pre-specified signed demographic JSD contrasts for Main topic. Positive values mean the first group in the contrast has higher JSD, i.e., is less well matched by the synthetic responses. Cells report mean signed gaps across model runs with 95% model-run jackknife intervals. * marks gaps that survive Holm-Bonferroni correction at $\alpha = 0.05$ within this pre-specified task family.

Condition	Runs	All responses		Within Other		Tail shape			Distance
		Other	Tail	No answer	Fixed topic	Tail	Topics	Top-5	JSD
Real	1	17.1%	5.9%	65.0%	0.8%	34.2%	386	34.9%	–
No context	12	2.5%	2.3%	2.6%	6.5%	90.9%	48.9	74.6%	0.694 [0.582, 0.859]
Uniform context	12	3.7%	3.7%	0.2%	1.7%	98.1%	134.3	46.5%	0.615 [0.563, 0.702]
Uniform context (lagged)	12	3.9%	3.8%	0.2%	1.5%	98.2%	140	47.8%	0.627 [0.565, 0.701]
Simulated context	12	8.6%	8.5%	1.3%	1.0%	97.6%	533.9	35.5%	0.527 [0.481, 0.591]
Simulated context (lagged)	12	8.4%	8.2%	1.3%	1.1%	97.6%	511	36.1%	0.527 [0.475, 0.577]
Self-reported context	12	8.4%	7.9%	0.8%	6.6%	92.7%	480.8	35.0%	0.539 [0.483, 0.592]
Self-reported context (lagged)	12	8.2%	7.6%	0.3%	8.7%	91.0%	477.6	35.3%	0.549 [0.502, 0.593]

Table 4: Exploratory diagnostics for responses classified as *Other*. All-response columns are shares of all responses; within-*Other* columns are shares of responses classified as *Other*. The substantive tail excludes *No answer* and fixed-topic labels, defined as open-ended topics that exactly match fixed main-topic labels. Synthetic rows report means across model runs within each context strategy. Tail JSD compares each model-run filtered-tail topic distribution to the real filtered-tail distribution and is reported as mean [10th, 90th percentile].

They can overtrack news coverage, miss lower-frequency topics, and fit some demographic groups better than others.

These findings suggest a more cautious role for LLMs in survey research. LLMs may be useful for analyzing open-ended responses, stress-testing survey instruments, or studying how synthetic responses differ from measured public opinion. However, they should not be treated as substitutes for human respondents based only on aggregate distributional fit. Open-ended survey data are valuable because they reveal what people select from a complex information environment. Evaluating synthetic publics therefore requires tests that preserve this structure: temporal variation, subgroup heterogeneity, and the long tail of less common topics.

6 Limitations

Our findings should be interpreted within several limits. First, the empirical setting is a single U.S. election-year survey with a fixed set of open-ended prompts. The 2024 election provides a useful stress test because political attention shifted rapidly, but the results may not generalize to lower-salience periods, non-political domains, or surveys with different question wording.

Second, the survey data come from weekly cross-sectional Pollfish samples rather than a probability sample or longitudinal panel. Respondents were not tracked over time, and limited repeat participation is possible. Our goal is therefore not to estimate population-level public opinion during the election, but to compare real and synthetic response distributions under the same observed survey design. Subgroup analyses are also limited by the demographic variables collected and by sample sizes within each subgroup.

Third, our news context is only an approximation of the information environment respondents may have encountered. The Media Bias Detector corpus provides broad coverage of contemporane-

ous news production across major outlets, but it does not capture all political information sources, including social media, local news, interpersonal discussion, campaign communication, or algorithmically personalized feeds (Haider et al., 2025). The self-reported media-diet condition uses respondents' reported source consumption, but self-reports are noisy and cannot recover individual exposure. The news data also cannot identify the mechanisms that lead respondents to mention some stories rather than others. We observe that open-ended responses diverge from news production, but cannot distinguish among attention, memory, personal relevance, social discussion, or unobserved exposure.

Fourth, our experiments cover a finite set of models, prompts, sampling settings, and context-selection strategies. We do not claim to identify the minimum achievable divergence between synthetic and real responses. Better prompts, richer behavioral modeling, different retrieval sources, fine-tuning, or agent-based simulations could reduce some of the gaps we document. Our claim is narrower: the context strategies we test substantially improve aggregate alignment but still leave failures in temporal dynamics, subgroup fit, and low-frequency topics.

Fifth, our evaluation depends on automated topic and sentiment labels. These labels provide a common measurement framework for comparing responses, but they are not ground truth about respondent beliefs. We apply the same classifier and label taxonomy to real and synthetic responses and validate the primary classifier against GPT-4o, which supports relative comparisons across conditions. Still, classification error may affect absolute JSD levels, and the long-tail analysis is especially sensitive to the quality of secondary open-ended topic labels.

7 Ethical Considerations

This study uses survey responses from adult participants recruited through Pollfish. See Appendix A.1 for specific details regarding the survey instrument and human subjects. The main ethical risk in this work concerns respondent privacy. Open-ended political responses can contain sensitive opinions, personal experiences, or identifying details. We therefore do not release raw respondent text in identifiable form. Public artifacts will be limited to code, model prompts, derived annotations, aggregate statistics, and documentation sufficient to reproduce the analyses without exposing individual respondents' free-text answers. Any released examples should be paraphrased or otherwise screened to reduce re-identification risk.

A second risk concerns the misuse of synthetic political responses. Our results should not be used to justify replacing human survey data with LLM-generated responses. The central finding of the paper is that context-grounded models can improve aggregate fit while still failing in ways that matter for open-ended measurement, including overtracking news coverage, uneven subgroup fit, and weak recovery of low-frequency topics. We therefore still caution against using synthetic responses as substitutes for measured public opinion, especially for subgroup analysis, political forecasting, or claims about what particular populations believe.

The experiments also generate synthetic political text. Although these outputs are used only for measurement and comparison, generated responses about politics could be misleading if presented outside the experimental context. We treat synthetic outputs as research artifacts, not as evidence of actual public opinion. Any release of synthetic responses should clearly mark them as model-generated and should avoid presenting them as human survey data.

Finally, our analysis relies on automated topic and sentiment labels. These labels are a measurement tool for comparing response distributions, not ground truth about respondents' beliefs. We apply the same labeling pipeline to real and synthetic responses to support relative comparisons, but we do not use the labels to make individual-level claims about respondents. This is particularly important for political opinion data, where misclassification could incorrectly attribute views or concerns to individuals or demographic groups.

References

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4):401–416.

Julien Boelaert, Samuel Coavoux, Étienne Ollion, Ivaylo Petev, and Patrick Präg. 2025. Machine bias. how do generative language models answer opinion polls?1. *Sociological Methods & Research*, 54(3):1156–1196.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025. Specializing large language models to simulate survey response distributions for global populations. In *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 3141–3154, Albuquerque, New Mexico. Association for Computational Linguistics.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 45850–45878. Curran Associates, Inc.

Esin Durmus, Karina Nguyen, Thomas I. 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. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2023. Using imperfect surrogates for downstream inference: design-based supervised learning for social science applications of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Gemma Team. 2025. Gemma 3 4B IT Model Card. Hugging Face model card. Accessed 2026-05-24.

Yair Ghitza and Andrew Gelman. 2013. Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3):762–776.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Bastián González-Bustamante, Nando Verelst, and Carla Cisternas. 2025. Emulating public opinion: A proof-of-concept of ai-generated synthetic survey responses for the chilean case.

Andrew Gordon, David Rothschild, Felipe M Affonso, Justin Sulik, David Hauser, Karine Pepin, and Simon Jones. 2026. Ai agent prevalence and data quality across multiple online sample providers.

Samar Haider, Amir Tohidi, Jenny S Wang, Timothy Dörr, David M Rothschild, Chris Callison-Burch, and Duncan J Watts. 2025. The media bias detector: A framework for annotating and analyzing the news at scale. *arXiv preprint arXiv:2509.25649*.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

John J Horton, Apostolos Filippas, and Benjamin S Manning. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research.

Shanto Iyengar and Donald R. Kinder. 1987. *News That Matters: Television and American Opinion*. University of Chicago Press, Chicago.

Junsol Kim and Byungkyu Lee. 2026. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *Preprint*, arXiv:2305.09620.

Sanguk Lee, Tai-Quan Peng, Matthew H. Goldberg, Seth A. Rosenthal, John E. Kotcher, Edward W. Maibach, and Anthony Leiserowitz. 2024. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):e0000429.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Zhicheng Lin. 2025. Six fallacies in substituting large language models for human participants. *Advances in Methods and Practices in Psychological Science*, 8(3):25152459251357566.

Alex Lyman, Bryce Hepner, Lisa P. Argyle, Ethan C. Busby, Joshua R. Gubler, and David Wingate. 2025. Balancing large language model alignment and algorithmic fidelity in social science research. *Sociological Methods & Research*, 54(3):1110–1155.

Bolei Ma, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, and Matthias Aßenmacher. 2025. Algorithmic fidelity of large language models in generating synthetic German public opinions: A case study. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1785–1809, Vienna, Austria. Association for Computational Linguistics.

Maxwell E. McCombs and Donald L. Shaw. 1972. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2):176–187.

Meta AI. 2024. Llama 3.2 1B and 3B Instruct Model Cards. Hugging Face model cards. Includes `meta-llama/Llama-3.2-1B-Instruct` and `meta-llama/Llama-3.2-3B-Instruct`; accessed 2026-05-24.

Microsoft. 2024. Phi-3.5-mini-instruct Model Card. Hugging Face model card. Accessed 2026-05-24.

Mistral AI. 2025. Mistral Small 3.2 24B Instruct 2506 Model Card. Hugging Face model card. Accessed 2026-05-24.

OpenAI. 2024. GPT-4o System Card. OpenAI system card. Accessed 2026-05-24.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *Preprint*, arXiv:2508.10925.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *Preprint*, arXiv:2306.00176.

Riya Parikh, Sarah H. Cen, and Chara Podimata. 2026. Do llms track public opinion? a multi-model study of favorability predictions in the 2024 u.s. presidential election. *Preprint*, arXiv:2602.06302.

Markus Prior. 2007. *Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections*. Cambridge University Press, Cambridge.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1).

Qwen Team. 2024. Qwen2.5 Instruct Model Cards. Hugging Face model cards. Includes `Qwen/Qwen2.5-1.5B-Instruct` and `Qwen/Qwen2.5-7B-Instruct`; accessed 2026-05-24.

Qwen Team. 2025a. Qwen3 Model Cards. Hugging Face model cards. Includes `Qwen/Qwen3-1.7B`, `Qwen/Qwen3-4B`, `Qwen/Qwen3-8B`, and `Qwen/Qwen3-32B`; accessed 2026-05-24.

Qwen Team. 2025b. Qwen3-Next-80B-A3B-Instruct Model Card. Hugging Face model card. Accessed 2026-05-24.

David M Rothschild, James Brand, Hope Schroeder, and Jenny Wang. 2024. Opportunities and risks of llms in survey research. *SSRN*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Natalie Jomini Stroud. 2010. Polarization and partisan selective exposure. *Journal of Communication*, 60(3):556–576.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Veniamin Veselovsky, Manoel Horta Ribeiro, Philip J. Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2025. Prevalence and prevention of large language model use in crowd work. *Commun. ACM*, 68(3):42–47.

Jenny S Wang, Samar Haider, Amir Tohidi, Anushkaa Gupta, Yuxuan Zhang, Chris Callison-Burch, David Rothschild, and Duncan J Watts. 2025. Media bias detector: Designing and implementing a tool for real-time selection and framing bias analysis in news coverage. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991.

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

A Supplementary Methods and Results

This appendix first documents the survey, prompt, context, model, and release details needed to reproduce the experiments, then reports supplemental diagnostics.

A.1 Survey Instrument and Human-Subjects Documentation

Survey respondents were recruited through Pollfish for single-session survey administrations. The Pollfish collection protocol received an IRB exempt determination. Eligible respondents were adults aged 18 or older; minors and cognitively impaired

populations were not targeted. Consent and participation notices were administered through the Pollfish survey flow, and participation was voluntary.

Pollfish managed participant incentives through its standard platform reward mechanisms, such as in-app credits or other digital incentives provided after survey completion. The research team did not directly pay respondents, provide gift cards or cash, collect Social Security numbers, or maintain a financial relationship with participants. For this survey, the research team paid Pollfish approximately $1.65 per completed respondent; the exact reward received by each respondent was determined by Pollfish and was not controlled by the research team.

Raw open-ended responses were treated as respondent data and are not released in identifiable form. Public artifacts are limited to code, derived annotations, aggregate statistics, and documentation needed to reproduce the analyses without exposing raw respondent text.

The broader survey instrument also included structured exposure, attitude, vote-intention, party-feeling, and source-consumption items. Except for respondent demographics, party identification, and source-consumption fields, these structured responses were not used as generation context or evaluation targets in the main experiments.

The raw Pollfish exports used here contain 19,250 respondent-wave rows across 40 waves, from February 9 to November 15, 2024. We dropped 382 rows for which the concatenated Q6–Q8 open-text field was blank or no more than five characters, leaving 18,868 respondent-level open-text rows in both the processed survey file and the media-diet version used for self-reported context. Because each retained respondent had three possible open-ended fields, this corresponds to 56,604 possible respondent-question text fields, 40,746 of which were nonempty before concatenation. The real-response classification file contains the same 18,868 rows, with no missing main-topic or sentiment labels.

Figure 4 provides descriptive context for the real survey responses and is not used as a synthetic-response evaluation metric.

A.2 Generation Prompts

The following prompt boxes reproduce the generation templates used in the final scripts, with LaTeX line wrapping and placeholder notation added for readability. The user message is shared across the default and ablation conditions; the ablations replace only the system message. The resulting chat messages were rendered with each model's Hugging Face chat template before inference.

As stated in Section 2, after setting the context, each generation prompt was used to generate responses to Questions 6–8.

Prompt: Core Generation

```
System: You are simulating a survey
respondent. Answer the survey question as
the described person would naturally write
- keep it casual, brief, and in their
voice. Do not be overly articulate or
polished. Just answer the question
directly, nothing else.

User: You are the following person taking
a political survey:
- Age: <age>
- Gender: <gender>
- Race/Ethnicity: <race>
- Political affiliation: <party>
- Education: <education>
- State: <state>

Today's date is <wave date>.

<Optional news-context block, if context
is enabled.>

Survey question: <question text>

Your response (1-2 sentences):

<Optional word-limit line: Respond in <N>
words or fewer.>
```

Prompt: Mobile Style Ablation

```
System: You are simulating a survey
respondent typing on their phone. Keep it
very short - one sentence or a brief
fragment. Use casual language, skip
capitalization and punctuation when
natural. Don't be articulate or polished.
Type like someone rushing through a survey
on mobile. Typos and incomplete thoughts
are fine. Just answer the question,
nothing else.

User: Same persona, date, optional news
context, and survey-question template as
the core prompt.
```

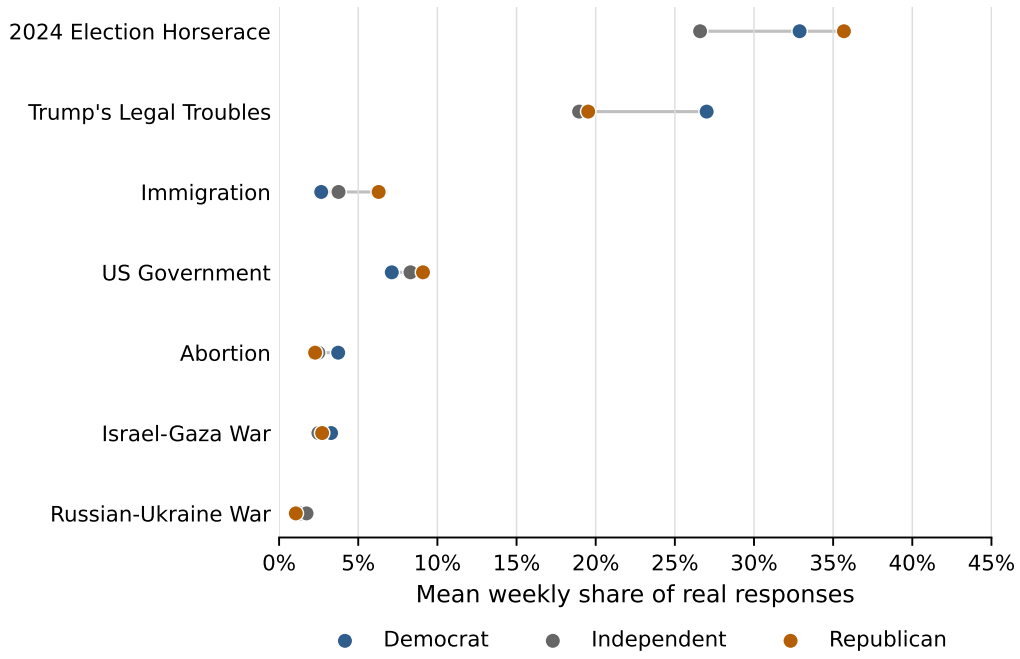

Figure 4: Real-response topics with the largest party gaps. Topics are selected by the mean weekly max–min spread across Democrat, Independent, and Republican respondents, after requiring at least 1% pooled share and excluding *Other*. Points show each party's mean weekly response share.

Prompt: Terse Style Ablation

```
System: You are simulating a survey
respondent who gives minimal effort
responses. Answer in one short sentence or
a few words. Be blunt and brief. Do not
elaborate or explain. Just answer the
question directly, nothing else.

User: Same persona, date, optional news
context, and survey-question template as
the core prompt.
```

Prompt: Casual Style Ablation

```
System: You are simulating a survey
respondent. Write casually like you're
texting a friend about politics. Keep it
brief and informal. Don't use fancy
vocabulary or complex sentences. Just
answer the question directly, nothing
else.

User: Same persona, date, optional news
context, and survey-question template as
the core prompt.
```

A.3 Generation Conditions

The main experiment crossed each of the 12 generation models with seven prompt conditions: baseline/no context; uniform, simulated, and self-reported context using the current survey-wave window; and the same three context strategies using a one-week-lagged window. Appendix A.6 defines the context-selection rules; this subsection records the inference settings shared by the generation runs.

We generated one response to each of the three open-ended survey questions for each respondent and stored the raw completions as JSONL. Main runs used vLLM batched inference with Hugging Face chat templates, bfloat16 weights, one GPU per model run, tensor parallel size 1, and 90% vLLM GPU-memory utilization. Sampling used temperature 0.7, top-p 0.9, and a maximum of 150 new tokens per question. We did not apply a word cap in the main runs; the length ablations added only a prompt instruction capping responses at the requested word limit. Qwen3 generation used the tokenizer option `enable_thinking=False`. The news-context sampler used seed 42; no separate decoding seed was set.

Generated responses were not parsed before classification. We stripped leading and trailing whitespace, preserved the free-text answer for each question, and concatenated the three fields for downstream labeling. Failed model loads or interrupted Slurm jobs exited without imputation; reruns used the resume or overwrite behavior in the launch scripts, and only completed JSONL records entered analysis.

A.4 Classification Prompts

The first classification pass used separate fixed-label prompts for the reported sentiment and main-topic labels. The fixed-label pass requested a single numeric answer; it did not request confidence scores or JSON output. The second pass was applied only to responses classified as *Other* in the main-topic task and requested an open-ended `topic | subtopic` label. Braced fields denote row-specific substitutions.

Prompt: Sentiment Classification

```
System: You are a helpful assistant.

User: Analyze the overall sentiment of
this text about current events.

Text: "{text}"

Sentiment:
1. Positive
2. Negative
3. Neutral

Answer (number only):
```

Prompt: Main-Topic Classification

```
System: You are a helpful assistant.

User: Identify the main topic discussed in
this text:

1. US Economy
2. Israel-Gaza War
3. Civil Liberties (e.g., Voting Rights)
4. Crime
5. Gun Safety
6. Abortion
7. Immigration
8. Russian-Ukraine War
9. Environment
10. Taxes
11. Education
12. Trump's Legal Troubles
13. 2024 Election Horserace (i.e., who is
going to win)
14. US Government (e.g., Congress, Supreme
Court, President, etc.)
15. Other

Text: "{text}"

Answer (number only):
```

Prompt: Open-Ended Other Labels

```
System: You are labeling open-ended survey
responses for social science research.
Use concise, neutral labels. Do not infer
beliefs that are not stated.

User: A previous 15-way classifier
labeled this response as main-topic
"Other". Assign an open-ended topic and
subtopic for what the response is about.

The goal is to preserve the long tail. Do
not choose from a fixed taxonomy. Create
concise natural labels from the response
text.

Original main-topic boundaries, for
context only:
- US Economy
- Israel-Gaza War
- Civil Liberties / Voting Rights
- Crime
- Gun Safety
- Abortion / Reproductive Rights
- Immigration
- Russia-Ukraine War
- Environment
- Taxes
- Education
- Trump's Legal Troubles
- 2024 Election Horserace
- US Government

Article-corpus domains that may appear in
Other, also for context only:
- Arts and Entertainment
- Celebrity
- Sports - Baseball
- Sports - Basketball
- Sports - Football

Guidelines:
- topic must be at most 2 words.
- subtopic must be at most 2 words.
- Prefer the respondent's specific subject
over a broad bucket.
- Reuse an obvious common label when the
text clearly matches it, but do not force
rare items into common labels.
- If the response is a non-answer, use
topic "No answer" and subtopic "No topic".
- If multiple topics appear, label the
dominant topic.
- Answer with topic and subtopic only,
separated by a vertical bar.

Survey wave date: {wave_date}

Response text: "{text}"

Answer format: topic | subtopic
```

A.5 Label Validation

The experiments evaluate model performance under a fixed classification scheme. To validate this scheme, we relabeled the original survey responses with GPT-4o and compared those labels with the Qwen classifier used in the main analysis. Table 5

Task	n	Accuracy	κ	Macro F1
Main topic	4,168	78.9%	0.762	0.797
Sentiment	4,168	76.5%	0.617	0.771

Table 5: Agreement between the primary Qwen3-Next-80B classifier and GPT-4o relabeling on the validation sample.

reports agreement above 0.7 on the assigned topic and above 0.6 on sentiment.

These agreement levels do not make the classifiers interchangeable, and they do not turn the labels into ground truth about respondent beliefs. They do, however, indicate that the main classifier is not behaving like a noisy or erratic labeler. The Qwen model assigned negative sentiment labels more often than GPT-4o, but it did so consistently across the comparison set. Because all experimental conditions are evaluated with the same classifier, the labels support relative comparisons among model conditions while remaining imperfect absolute measurements.

A.6 Context Creation Algorithms

Let d be the survey wave date, r a respondent, k the requested number of articles, and $\ell \in \{0, 1\}$ the week lag. The main context runs set $k = 5$. The seven generation conditions map to this notation as follows: the baseline sets the selected article list to empty; uniform context uses the UNIFORM selector; simulated context uses the SIMULATED media-diet selector; self-reported context uses the SELFREPORTED media-diet selector; and current versus lagged variants set $\ell = 0$ or $\ell = 1$, respectively.

Context creation proceeds in three steps. First, we restrict the article pool to the relevant survey-wave window: $\ell = 0$ uses $[d-7, d)$ and $\ell = 1$ uses $[d - 14, d - 7)$. For self-reported media diets, lagging also changes source weights to previous-wave demographic pools. Second, we select articles either at the wave level or respondent level. Uniform and attention-weighted contexts select one shared set of articles for all respondents in a wave, while simulated and self-reported media-diet contexts draw respondent-specific article sets using source weights and sampled article counts. Topic-based selectors choose one article per high-priority topic with a source-diversity preference; media-diet selectors sample sources and articles without replacement. Third, we format the selected articles as compact prompt context using publisher, title, and a stored sentence-bounded context field. At prompt-formatting time, context fields longer than 400 characters are truncated at a sentence boundary.

Publisher	Articles	Share	Topics
Fox News	5,966	6.9%	18
USA Today	5,770	6.7%	18
New York Post	5,604	6.5%	18
Chicago Tribune	5,498	6.4%	18
The Guardian	5,424	6.3%	18
Breitbart	5,360	6.2%	18
Associated Press	4,858	5.7%	18
Washington Post	4,719	5.5%	18
Huffington Post	4,636	5.4%	18
New York Times	4,584	5.3%	18
Newsmax	4,395	5.1%	18
CNN	4,288	5.0%	18
Reuters	3,564	4.1%	18
Bloomberg	3,553	4.1%	18
Los Angeles Times	3,433	4.0%	18
BBC	2,780	3.2%	18
Daily Wire	2,746	3.2%	18
Wall Street Journal	2,619	3.0%	18
Financial Times	2,487	2.9%	18
CNBC	1,968	2.3%	17
Star Tribune	1,644	1.9%	18

Table 6: Publishers in the news-context article pool used for response generation. Counts are unique normalized articles eligible for selection in any current or one-week-lagged survey window.

The algorithms below give the full selection procedure. The source pool for these selectors is summarized before the algorithm blocks; counts are generated from the normalized article parquet and the final survey wave dates by the reproducible publisher-summary script.

- The normalized article file contains 217,273 unique articles; the current and one-week-lagged survey windows contain 85,896 unique context-eligible articles.
- The context-eligible pool spans 2024-01-26–2024-11-15 across 40 survey waves and 80 current or lagged article windows.
- The eligible pool covers 21 publishers and 18 normalized topics; weekly windows average 2052.4 articles (range 1,629–2,380).
- Every current or lagged weekly article window includes 21 publishers in the candidate article pool.

Algorithm 1 End-to-end context creation and prompting

Require: Article corpus A, respondents R, human topic labels H, mode m, size k, lag ℓ, seed s_0

for all $r \in R$ **do**
  $d \leftarrow \text{wave_date}(r)$
  $b \leftarrow d - 7(\ell + 1)$; $e \leftarrow d - 7\ell$
  $W \leftarrow \{a \in A : b \leq \text{date}(a) < e\}$ ▷ $\ell = 0$: current week; $\ell = 1$: prior week
  if $m = \textsc{None}$ **then**
    $S \leftarrow [\,]$
  else if $m = \textsc{Uniform}$ **then**
    $S \leftarrow \text{Uniform}(W, d, k)$
  else if $m = \textsc{Attention}$ **then**
    $S \leftarrow \text{Attention}(W, H, d, k)$
  else
    $(w, n, g) \leftarrow \text{MediaDiet}(m, W, R, r, d, \ell, s_0)$
    $S \leftarrow \text{SampleBySource}(W, w, \min(k, n), g)$
  end if
  $C \leftarrow \text{Format}(S)$ ▷ publisher, title, context text
  $P \leftarrow \text{Persona}(r, d) + C + \text{Question}$
  $y_r \leftarrow \text{Generate}(P)$
end for

Algorithm 2 Topic-based context selectors

procedure $\textsc{Uniform}(W, d, k)$
  $T \leftarrow \text{topk_topics_by_count}(d, k)$
  return $\text{PickTopics}(W, T, k)$
end procedure

procedure $\textsc{Attention}(W, H, d, k)$
  $q_d \leftarrow \text{EMA}(\text{topic_shares}(H), 0.3, d)$
  $T \leftarrow \text{topk}(q_d, k)$
  return $\text{PickTopics}(W, T, k)$
end procedure

procedure $\textsc{PickTopics}(W, T, k)$
  $S \leftarrow [\,]$; $U \leftarrow \emptyset$
  for all $t \in T$ **do**
    $P \leftarrow \{a \in W : \text{topic}(a) = t\}$
    $Q \leftarrow \{a \in P : \text{source}(a) \notin U\}$
    if $Q \neq \emptyset$ **then**
      $P \leftarrow Q$
    end if
    if $P \neq \emptyset$ **then**
      $a^\star \leftarrow$ top-ranked article in P
      append $a^\star$ to S; add $\text{source}(a^\star)$ to U
    end if
    if $\text{len}(S) = k$ **then**
      return S
    end if
  end for
  return S
end procedure

A.6.1 Publisher Sampling Support

The normalized context pool contains 21 publishers (Table 6), and all 21 appear in every current and one-week-lagged survey-week article window used by the experiments. Thus, for stochastic per-respondent context strategies, all 21 publishers are in the weekly sampling support before source weights are applied. In simulated context, party-specific multipliers are applied to named political sources and remaining sources receive a neutral weight, with AP and Reuters treated as wire services. In self-reported context, Pollfish-mapped sources receive respondent-specific or previous-pool-specific weights, unmapped publishers remain eligible through a low floor weight, and AP and Reuters are again assigned neutral wire-service weight. Uniform and attention-weighted context do not sample publishers directly; they select articles by weekly topic priority while preferring publisher diversity, so all publishers are eligible at the article-pool stage but at most k articles are inserted into each prompt.

A.7 Model Artifacts

This subsection lists the model artifacts used for synthetic response generation, primary classification, and label validation. Table 7 gives the exact checkpoint or endpoint identifiers and the citation associated with each artifact family. The open-weight generation models span approximately 1B to 32B parameters: Llama 3.2 1B and 3B, Qwen2.5 1.5B and 7B, Qwen3 1.7B, 4B, 8B, and 32B, Phi-3.5-mini at approximately 3.8B, Gemma 3 4B, GPT-OSS 20B, and Mistral Small 3.2 24B. The primary classifier, Qwen3-Next-80B-A3B, is an 80B-parameter sparse model with approximately 3B activated parameters; GPT-4o is a proprietary validation endpoint whose parameter count is not publicly reported.

A.8 Computing Resources

Synthetic response generation and response classification were run as Slurm jobs on a GPU cluster. Generation jobs were run using one NVIDIA DGX-B200 GPU, 8 CPU cores, 128GB RAM. Classification jobs used `Qwen/Qwen3-Next-80B-A3B-Instruct` with tensor parallelism over two B200 GPUs, requesting 8 CPU cores, 128GB RAM.

The completed Slurm logs indicate an estimated total compute budget of approximately 105 B200 GPU-hours across generation, classification, ablation, and long-tail labeling jobs.

Downstream divergence calculations, significance tests, and figure generation were run as local (i.e. personal laptop) CPU analyses over the generated and classified artifacts.

Role / use	Model artifacts	Citation
Generation	`meta-llama/Llama-3.2-1B-Instruct`; `meta-llama/Llama-3.2-3B-Instruct`	(Meta AI, 2024)
Generation	`Qwen/Qwen3-1.7B`; `Qwen/Qwen3-4B`; `Qwen/Qwen3-8B`; `Qwen/Qwen3-32B`	(Qwen Team, 2025a)
Generation	`Qwen/Qwen2.5-1.5B-Instruct`; `Qwen/Qwen2.5-7B-Instruct`	(Qwen Team, 2024)
Generation	`microsoft/Phi-3.5-mini-instruct`	(Microsoft, 2024)
Generation	`google/gemma-3-4b-it`	(Gemma Team, 2025)
Generation	`mistralai/Mistral-Small-3.2-24B-Instruct-2506`	(Mistral AI, 2025)
Generation	`openai/gpt-oss-20b`	(OpenAI, 2025)
Classification	`Qwen/Qwen3-Next-80B-A3B-Instruct`; primary topic, sentiment, and other-topic labeling	(Qwen Team, 2025b)
Validation	`gpt-4o`; relabeling validation sample	(OpenAI, 2024)

Table 7: Model artifacts used in generation, classification, and validation. Exact checkpoint or endpoint identifiers are reported where models were loaded from Hugging Face or OpenAI model endpoints.

Algorithm 3 Media-diet weights and article counts

procedure MEDIADIET(m, W, R, r, d, ℓ, s_0)
  $g \leftarrow \mathrm{hash}(\mathrm{id}(r), d, s_0)$
  if $m =$ SIMULATED **then**
    $\mu \leftarrow 5$
    if $\mathrm{edu}(r) =$ COLLEGE **then**
      $\mu \leftarrow \mu + 1$
    end if
    if $\mathrm{edu}(r) =$ NOCOLLEGE **then**
      $\mu \leftarrow \mu - 1$
    end if
    if $\mathrm{age_lower}(r) \geq 45$ **then**
      $\mu \leftarrow \mu + 1$
    end if
    $n \leftarrow \mathrm{clip}(\mathrm{Poisson}(\max(\mu, 1)), 1, 10)$
    $w \leftarrow \mathrm{party_source_weights}(r, W)$
    set wire-service weights in w to 1.0
  else if $m =$ SELFREPORTED **then**
    if $\ell = 0$ **then**
      $x \leftarrow$ source reports and discussion score for r
    else
      $G \leftarrow \mathrm{prev_pool}(R, r, d)$ with demographic fallback
      $x \leftarrow$ source rates and mean discussion in G
    end if
    $n \leftarrow \mathrm{clip}(\mathrm{Poisson}(\mathrm{disc_volume}(x)), 1, 10)$
    for all $s \in \mathrm{sources}(W)$ **do**
      $w_s \leftarrow 1.0$ if s is a wire service
      otherwise $w_s \leftarrow 0.2 + 2.8\,\mathrm{rate}(x, s)$
    end for
  end if
  return (w, n, g)
end procedure

Algorithm 4 Source-weighted article sampling

procedure SAMPLEBYSOURCE(W, w, n, g)
  group W into non-empty source pools B_s
  $p_s \leftarrow w_s / \sum_{u: B_u \neq \emptyset} w_u$
  $S \leftarrow [\,]$; $I \leftarrow \emptyset$
  for $i = 1, \ldots, 3n$ **do**
    if $\mathrm{len}(S) = n$ **then**
      return S
    end if
    sample source $s \sim \mathrm{Categorical}(p)$ using seed g
    $P \leftarrow \{a \in B_s : \mathrm{id}(a) \notin I\}$
    if $P \neq \emptyset$ **then**
      sample article $a \sim \mathrm{Uniform}(P)$ using seed g
      append a to S; add $\mathrm{id}(a)$ to I
    end if
  end for
  return S
end procedure

A.9 Software

The code supplement included with this submission contains the analysis, generation, classification, and figure-generation scripts required to reproduce the reported results from the released artifacts. The Python environment is specified by `pyproject.toml` and `uv.lock` and can be recreated with `uv sync`. Cluster submission wrappers are not included in the anonymous supplement, but the executable scripts and command-line interfaces used by those wrappers are included.

The code supplement included with this submission contains the analysis, generation, classification, and figure-generation scripts used for the reported experiments. The Python environment is specified by `pyproject.toml` and `uv.lock` and can be recreated with `uv sync -all-extras`. Cluster submission wrappers are not included in the anonymous supplement, but the executable scripts and command-line interfaces used by those wrappers are included.

Because the anonymous data supplement excludes raw survey text, raw article text, row-level open-ended *Other* labels, and raw LLM generations, some upstream steps and text-dependent diagnostics cannot be fully rerun from the anonymous artifacts alone. The released code and data do support reproduction of the label-based analyses and aggregate summaries included in the supplement.

A.10 Data Release

The data supplement attached to this submission contains derived topic and sentiment labels for each

model and condition tested, article-count metadata, aggregate long-tail summaries, validation samples, and poststratification inputs used in the reported analyses. It does not include raw Pollfish exports, raw respondent free text, raw article text, or row-level open-ended *Other* labels, because these artifacts may contain sensitive or identifying text. Raw LLM generations are also omitted from the anonymous supplement because of their size; after review, we plan to provide access to generated synthetic text through a separate release channel when storage and documentation constraints are resolved.

A.11 Artifact Licenses and Intended Use

We used model checkpoints and endpoints under their associated provider licenses or terms of use and cite the model-card or system-card sources in Appendix A.7. The Pollfish survey data were collected for research under the IRB-exempt protocol described in Appendix A.1, and raw respondent text is not redistributed. The news-context corpus was used only to construct research prompts and aggregate article-topic diagnostics; raw article text is not redistributed in the anonymous supplement. The released code and derived data are intended to support reproduction and research evaluation, not individual-level profiling, public-opinion forecasting, or presentation of synthetic responses as human survey data.

A.12 Model-Level Core Results

This subsection gives the model-level version of the core JSD results. Tables 8 and 9 report aggregate sentiment and main-topic JSD by model and context strategy, while Figures 5, 6, and 7 show the corresponding weekly and aggregate model-level patterns.

A.13 Poststratification Check

For each real or synthetic condition c, we merge responses to poststratification cells, drop unmatched or unlabeled rows, estimate the weighted topic distribution, and compute JSD against the weighted real distribution.

$$\hat{p}_{c,k}^{(w)} = \frac{\sum_{i \in \mathcal{M}_c} w_i \mathbf{1}(y_i = k)}{\sum_{i \in \mathcal{M}_c} w_i},$$
$$\mathrm{JSD}_c = D_{\mathrm{JS}}\left(\hat{\mathbf{p}}_{\mathrm{real}}^{(w)}, \hat{\mathbf{p}}_c^{(w)}\right).$$

For Table 10, poststratification matching retained 95.5% of labeled real rows and, rounded to one decimal place, 95.5% of labeled synthetic rows in every model condition.

A.14 Category-Level Error

This subsection reports category-level mean absolute error as a complement to distribution-level JSD. Table 11 reports main-topic errors, and Table 12 reports sentiment errors.

A.15 Media-Supply Diagnostic

Table 13 reports article-topic shares and article-minus-response gaps after mapping article topics into the main-topic label space.

A.16 Demographic Diagnostics

This subsection expands the main-text demographic contrast analysis. Table 14 reports selected subgroup JSD by context condition, and Tables 15–21 report model-level subgroup JSD for each context condition.

A.17 Long-Tail Diagnostics

This subsection supports the main long-tail analysis with composition tables and distributional figures. Tables 22–25 decompose responses classified as *Other*, separate non-answers from substantive tail labels, summarize tail concentration, and list the most frequent open-ended tail topics; Figures 8 and 9 show the top filtered tail topics and the rank-frequency shape of the tail.

A.18 Ablation Diagnostics

This subsection reports sensitivity checks for response length and response style under the self-reported lagged context condition. Table 26 reports JSD under word-limit prompts, Table 27 reports JSD under style-prompt variants, and Figure 10 summarizes the mean JSD patterns.

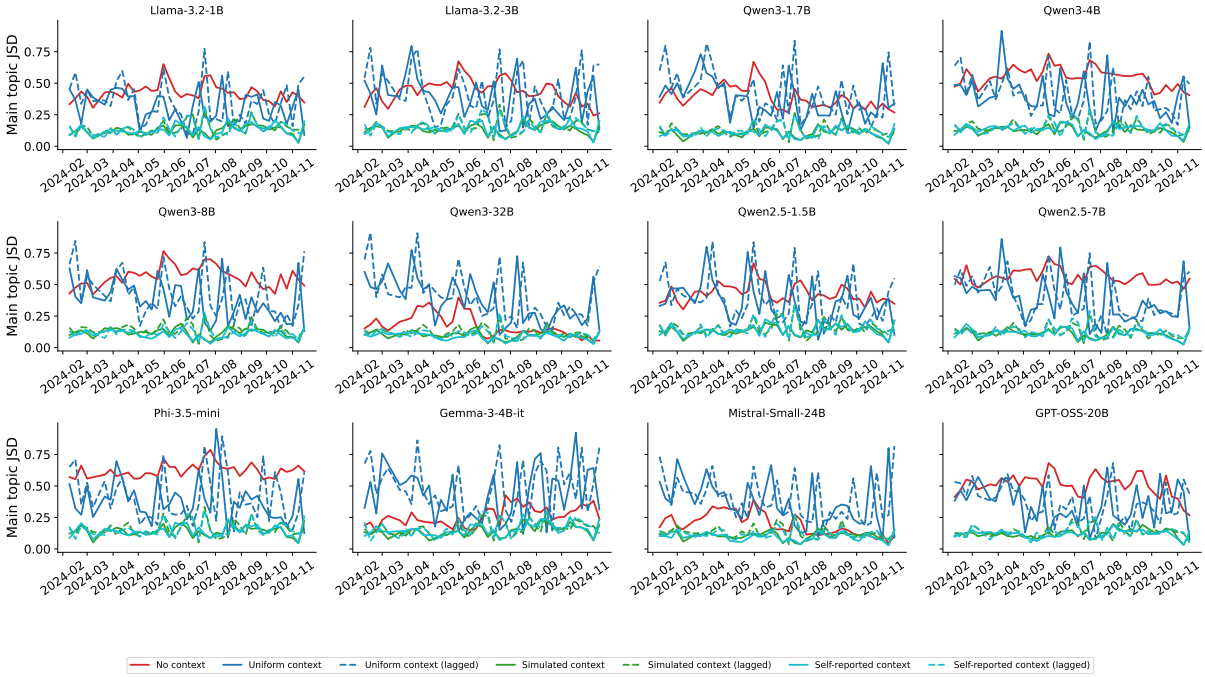

Figure 5: Model-level weekly main-topic JSD by context condition. Each panel shows one generation model.

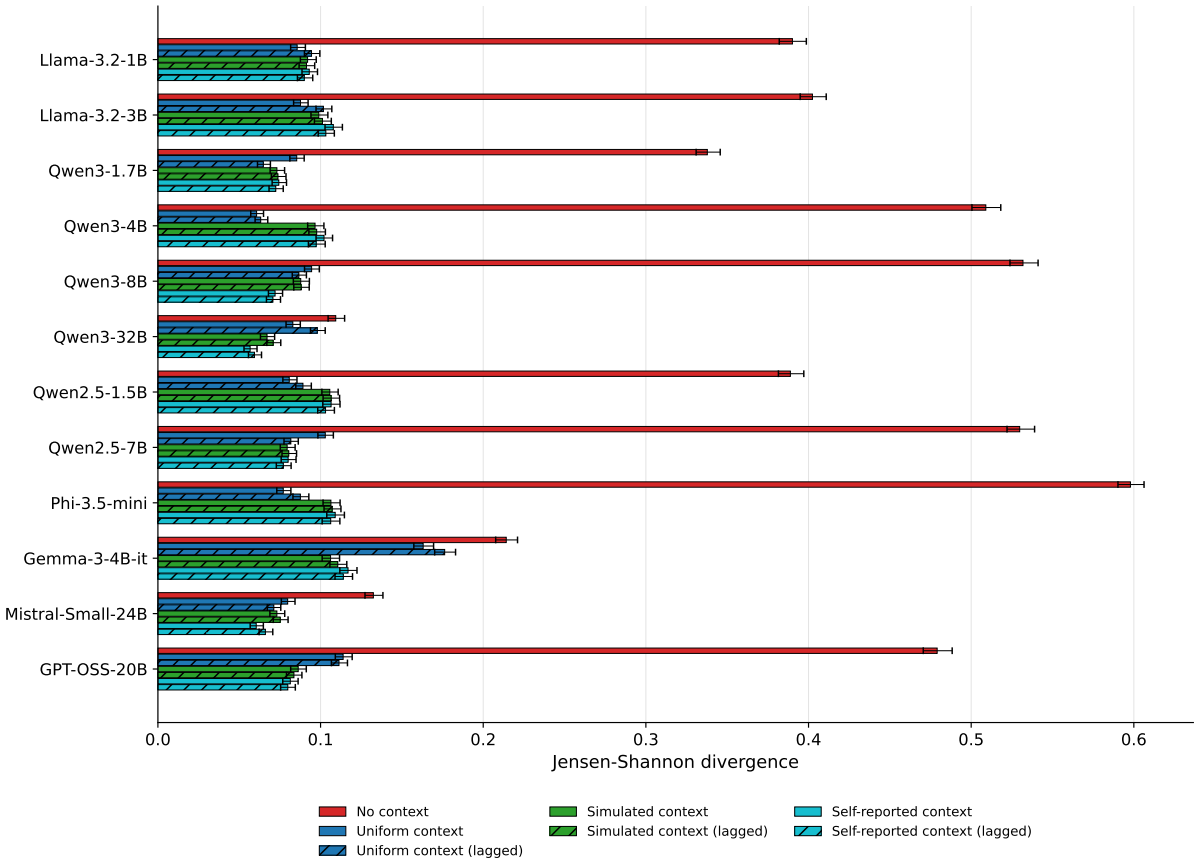

Figure 6: Model-level main-topic JSD by context condition. Bars show aggregate JSD against real responses; whiskers show bootstrap 95% confidence intervals.

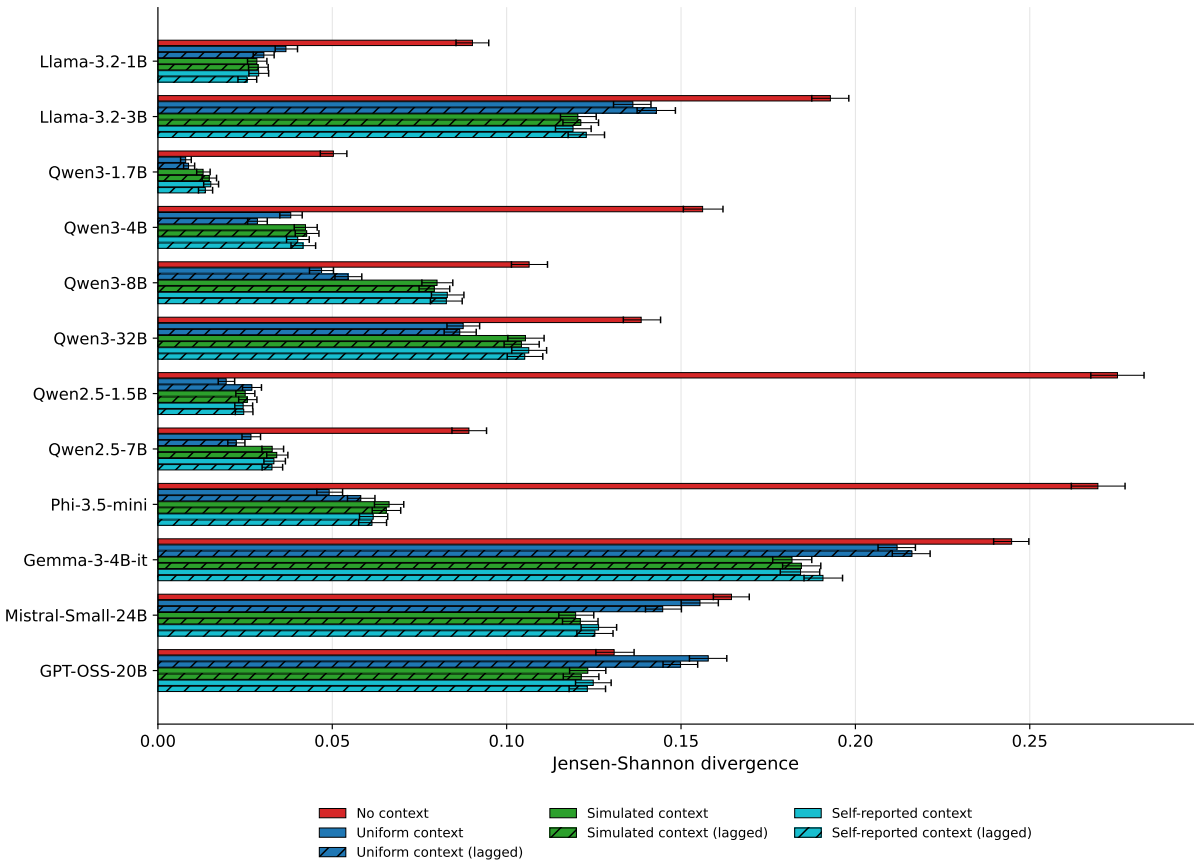

Figure 7: Model-level sentiment JSD by context condition. Bars show aggregate JSD against real responses; whiskers show bootstrap 95% confidence intervals.

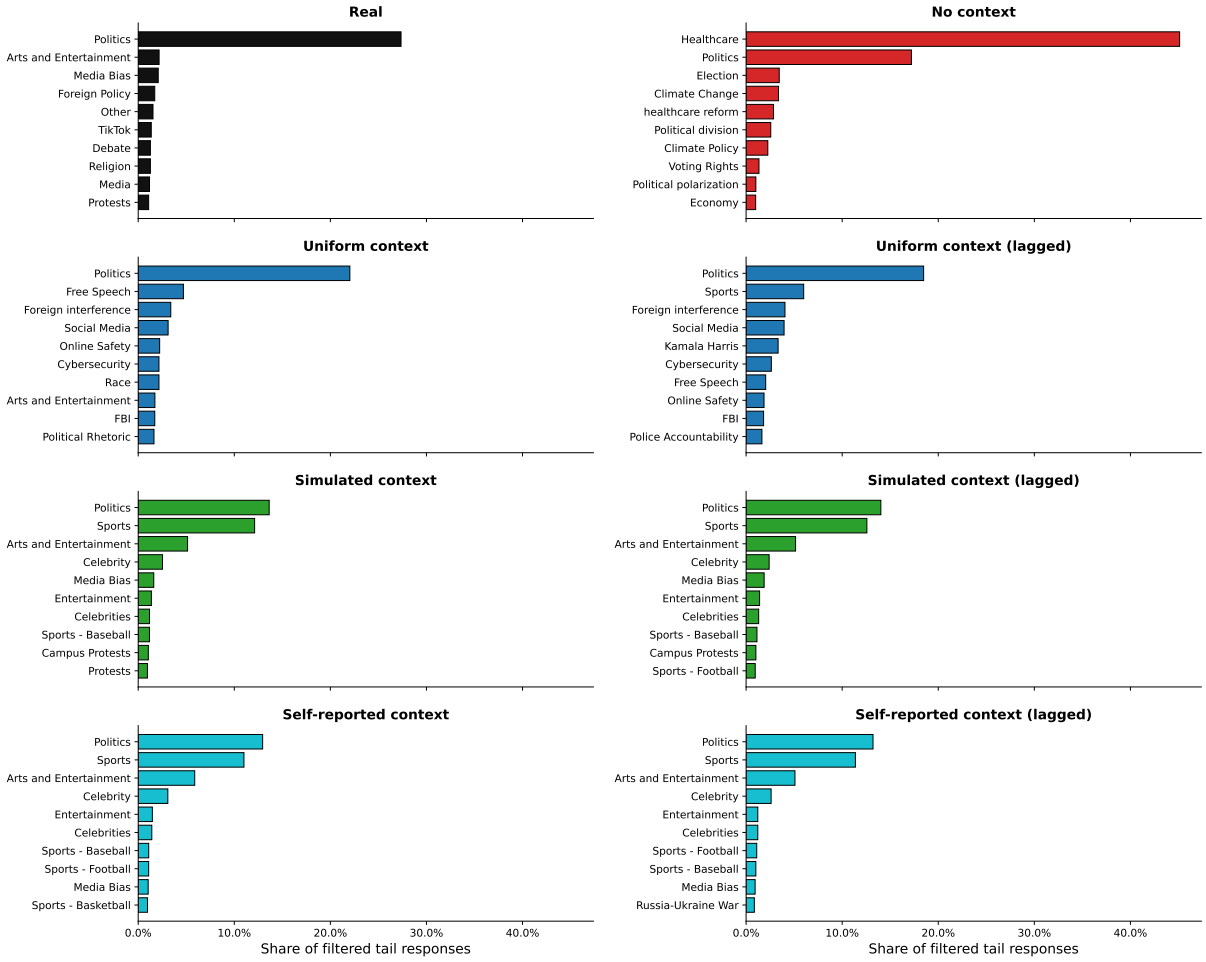

Figure 8: Top-10 open-ended topic shares among the remaining substantive tail after excluding *No answer* and exact matches to fixed main-topic labels. Topics are sorted within each condition by share.

Model	Baseline, no context	Uniform context		Simulated context		Self-reported context	
		Current	Lagged	Current	Lagged	Current	Lagged
Llama-3.2-1B	0.090 [0.085,0.095]	0.037 [0.034,0.040]	0.030 [0.027,0.033]	0.028 [0.026,0.031]	0.029 [0.026,0.032]	0.029 [0.026,0.032]	0.026 [0.023,0.028]
Llama-3.2-3B	0.193 [0.187,0.198]	0.136 [0.131,0.141]	0.143 [0.137,0.148]	0.120 [0.115,0.126]	0.121 [0.116,0.126]	0.119 [0.114,0.124]	0.123 [0.118,0.128]
Qwen3-1.7B	0.050 [0.047,0.054]	0.008 [0.007,0.010]	0.009 [0.007,0.011]	0.013 [0.011,0.015]	0.015 [0.013,0.017]	0.015 [0.013,0.017]	0.014 [0.012,0.016]
Qwen3-4B	0.156 [0.151,0.162]	0.038 [0.035,0.041]	0.029 [0.026,0.031]	0.042 [0.039,0.046]	0.043 [0.039,0.046]	0.040 [0.037,0.043]	0.042 [0.038,0.045]
Qwen3-8B	0.106 [0.101,0.112]	0.047 [0.043,0.050]	0.055 [0.051,0.058]	0.080 [0.076,0.085]	0.079 [0.075,0.084]	0.083 [0.078,0.088]	0.083 [0.078,0.087]
Qwen3-32B	0.139 [0.133,0.144]	0.088 [0.083,0.092]	0.087 [0.082,0.091]	0.105 [0.100,0.111]	0.104 [0.099,0.109]	0.106 [0.101,0.111]	0.105 [0.100,0.110]
Qwen2.5-1.5B	0.275 [0.268,0.283]	0.020 [0.017,0.022]	0.027 [0.024,0.030]	0.025 [0.022,0.028]	0.026 [0.023,0.028]	0.024 [0.022,0.027]	0.025 [0.022,0.027]
Qwen2.5-7B	0.089 [0.084,0.094]	0.027 [0.024,0.029]	0.022 [0.020,0.025]	0.033 [0.030,0.036]	0.034 [0.031,0.037]	0.033 [0.030,0.037]	0.033 [0.030,0.036]
Phi-3.5-mini	0.270 [0.262,0.277]	0.049 [0.046,0.053]	0.058 [0.054,0.062]	0.066 [0.062,0.070]	0.065 [0.061,0.070]	0.062 [0.058,0.066]	0.061 [0.058,0.066]
Gemma-3-4B-it	0.245 [0.240,0.250]	0.212 [0.206,0.217]	0.216 [0.211,0.221]	0.182 [0.176,0.187]	0.185 [0.179,0.190]	0.184 [0.178,0.190]	0.191 [0.185,0.196]
Mistral-Small-24B	0.164 [0.159,0.170]	0.155 [0.150,0.161]	0.145 [0.140,0.150]	0.120 [0.115,0.125]	0.121 [0.116,0.126]	0.126 [0.121,0.132]	0.125 [0.120,0.131]
GPT-OSS-20B	0.131 [0.126,0.137]	0.158 [0.152,0.163]	0.150 [0.145,0.155]	0.123 [0.118,0.128]	0.121 [0.116,0.126]	0.125 [0.120,0.130]	0.123 [0.118,0.128]

Table 8: Aggregate Sentiment JSD by model and context strategy. Cells report bootstrap 95% confidence intervals from the label-level resampling in the divergence computation.

Model	Baseline, no context	Uniform context		Simulated context		Self-reported context	
		Current	Lagged	Current	Lagged	Current	Lagged
Llama-3.2-1B	0.390 [0.382,0.399]	0.086 [0.082,0.091]	0.094 [0.090,0.100]	0.092 [0.088,0.097]	0.091 [0.087,0.096]	0.093 [0.089,0.098]	0.090 [0.086,0.095]
Llama-3.2-3B	0.402 [0.395,0.411]	0.088 [0.083,0.092]	0.102 [0.097,0.107]	0.099 [0.094,0.104]	0.101 [0.096,0.107]	0.108 [0.103,0.113]	0.103 [0.099,0.108]
Qwen3-1.7B	0.338 [0.331,0.346]	0.085 [0.081,0.090]	0.065 [0.061,0.069]	0.073 [0.069,0.078]	0.074 [0.070,0.079]	0.074 [0.070,0.079]	0.072 [0.068,0.077]
Qwen3-4B	0.509 [0.500,0.518]	0.061 [0.057,0.065]	0.063 [0.060,0.068]	0.097 [0.092,0.102]	0.098 [0.093,0.103]	0.102 [0.097,0.107]	0.097 [0.093,0.103]
Qwen3-8B	0.532 [0.524,0.541]	0.094 [0.090,0.099]	0.087 [0.083,0.091]	0.088 [0.083,0.093]	0.088 [0.084,0.093]	0.072 [0.068,0.077]	0.071 [0.067,0.075]
Qwen3-32B	0.109 [0.105,0.115]	0.083 [0.079,0.088]	0.098 [0.094,0.103]	0.067 [0.063,0.072]	0.071 [0.067,0.076]	0.057 [0.053,0.061]	0.059 [0.056,0.064]
Qwen2.5-1.5B	0.389 [0.381,0.397]	0.081 [0.077,0.085]	0.089 [0.085,0.094]	0.106 [0.101,0.111]	0.106 [0.102,0.112]	0.106 [0.101,0.112]	0.103 [0.098,0.109]
Qwen2.5-7B	0.530 [0.522,0.539]	0.103 [0.098,0.108]	0.082 [0.078,0.086]	0.079 [0.075,0.084]	0.080 [0.076,0.085]	0.080 [0.076,0.085]	0.077 [0.073,0.082]
Phi-3.5-mini	0.598 [0.590,0.606]	0.077 [0.073,0.082]	0.088 [0.083,0.093]	0.106 [0.102,0.112]	0.107 [0.102,0.113]	0.109 [0.104,0.115]	0.106 [0.101,0.112]
Gemma-3-4B-it	0.214 [0.208,0.221]	0.163 [0.157,0.169]	0.176 [0.170,0.183]	0.106 [0.101,0.112]	0.111 [0.106,0.116]	0.117 [0.112,0.122]	0.114 [0.109,0.120]
Mistral-Small-24B	0.132 [0.127,0.138]	0.080 [0.076,0.084]	0.071 [0.067,0.076]	0.073 [0.069,0.078]	0.075 [0.071,0.080]	0.060 [0.057,0.065]	0.066 [0.062,0.071]
GPT-OSS-20B	0.479 [0.471,0.488]	0.114 [0.109,0.119]	0.111 [0.107,0.117]	0.086 [0.082,0.091]	0.084 [0.079,0.089]	0.081 [0.077,0.086]	0.080 [0.075,0.084]

Table 9: Aggregate Main topic JSD by model and context strategy. Cells report bootstrap 95% confidence intervals from the label-level resampling in the divergence computation.

Model	Raw JSD	Poststrat. JSD	Δ JSD
Llama-3.2-1B	0.090	0.109	+0.019
Llama-3.2-3B	0.103	0.122	+0.019
Qwen3-1.7B	0.072	0.087	+0.014
Qwen3-4B	0.097	0.112	+0.015
Qwen3-8B	0.071	0.083	+0.012
Qwen3-32B	0.059	0.068	+0.009
Qwen2.5-1.5B	0.103	0.128	+0.025
Qwen2.5-7B	0.077	0.088	+0.011
Phi-3.5-mini	0.106	0.126	+0.020
Gemma-3-4B-it	0.114	0.133	+0.019
Mistral-Small-24B	0.066	0.076	+0.010
GPT-OSS-20B	0.080	0.084	+0.004

Table 10: Direct poststratification-weighted Main-topic JSD check for self-reported lagged context.

Category	Baseline, no context	Uniform context		Simulated context		Self-reported context	
		Current	Lagged	Current	Lagged	Current	Lagged
US Economy	15.0	2.0	3.0	4.8	4.7	3.6	3.1
Israel-Gaza War	2.8	2.5	2.5	5.4	5.5	6.3	6.5
Civil Liberties	5.1	0.6	0.4	1.6	1.8	1.7	1.9
Crime	0.7	4.9	5.6	4.6	4.5	4.7	4.4
Gun Safety	0.9	0.5	0.9	1.1	0.9	1.1	1.2
Abortion	5.5	1.8	1.3	0.5	0.6	0.7	0.7
Immigration	2.5	1.9	1.5	3.0	3.0	2.3	2.3
Russian-Ukraine War	1.2	1.1	0.9	3.8	4.0	4.0	3.9
Environment	8.1	0.5	0.5	1.9	2.0	2.0	2.0
Taxes	7.7	0.9	0.9	0.7	0.9	0.6	0.5
Education	2.1	0.4	0.4	0.6	0.6	0.6	0.6
Trump's Legal Troubles	20.0	10.1	9.5	8.6	8.1	7.8	7.3
2024 Election Horserace	22.1	6.4	6.8	11.6	12.0	11.7	11.6
US Government	12.5	2.6	2.3	1.1	1.1	1.1	1.1
Other	14.6	13.4	13.2	8.5	8.7	8.7	8.9
Mean across categories	8.1	3.3	3.3	3.9	3.9	3.8	3.7

Table 11: Category-level absolute error for Main topic. Cells report mean absolute percentage-point difference between real and synthetic category shares, averaged across the 12 balanced model runs within each context strategy.

Category	Baseline, no context	Uniform context		Simulated context		Self-reported context	
		Current	Lagged	Current	Lagged	Current	Lagged
Positive	19.6	5.4	5.5	7.3	7.4	7.4	7.3
Negative	23.0	18.8	19.1	17.2	17.3	17.3	17.4
Neutral	25.3	21.4	21.3	21.7	21.8	21.7	21.7
Mean across categories	22.6	15.2	15.3	15.4	15.5	15.5	15.4

Table 12: Category-level absolute error for Sentiment. Cells report mean absolute percentage-point difference between real and synthetic category shares, averaged across the 12 balanced model runs within each context strategy.

Topic	Article share %	Article - human pp.	Article - no context pp.	Article - current pp.	Article - lagged pp.
Trump's Legal Troubles	2.4	-20.4 [-21.0, -19.8]	-3.7 [-3.8, -3.5]	-18.7 [-18.8, -18.6]	-18.8 [-18.9, -18.6]
Crime	16.7	+15.9 [+15.6, +16.2]	+16.6 [+16.3, +16.8]	+11.2 [+10.9, +11.5]	+11.0 [+10.8, +11.3]
2024 Election Horserace	24.3	-9.9 [-10.6, -9.1]	+7.6 [+7.3, +7.9]	-1.9 [-2.2, -1.6]	-0.5 [-0.8, -0.2]
US Government	0.0	-7.9 [-8.3, -7.6]	-20.0 [-20.2, -19.8]	-8.8 [-8.8, -8.7]	-8.7 [-8.8, -8.7]
US Economy	8.7	+6.5 [+6.2, +6.8]	-8.0 [-8.2, -7.7]	+3.1 [+2.9, +3.3]	+3.0 [+2.8, +3.2]
Other	22.0	+4.9 [+4.2, +5.5]	+19.5 [+19.2, +19.8]	+15.1 [+14.8, +15.4]	+15.2 [+14.9, +15.5]
Israel-Gaza War	6.3	+3.4 [+3.1, +3.7]	+6.3 [+6.1, +6.4]	-1.0 [-1.2, -0.8]	-0.9 [-1.1, -0.8]
Education	3.6	+3.1 [+2.9, +3.3]	+1.2 [+1.0, +1.3]	+2.9 [+2.7, +3.0]	+2.9 [+2.7, +3.0]
Russian-Ukraine War	3.9	+2.7 [+2.5, +2.9]	+3.6 [+3.5, +3.8]	+0.5 [+0.3, +0.6]	+0.4 [+0.2, +0.5]
Environment	3.2	+2.7 [+2.5, +2.9]	-5.4 [-5.6, -5.2]	+1.6 [+1.4, +1.7]	+1.5 [+1.4, +1.6]
Abortion	2.2	-0.7 [-1.0, -0.4]	-4.6 [-4.7, -4.5]	-0.4 [-0.5, -0.3]	-0.8 [-0.9, -0.7]
Taxes	0.8	-0.2 [-0.4, -0.1]	-7.8 [-7.9, -7.7]	-0.3 [-0.4, -0.3]	-0.4 [-0.4, -0.3]
Immigration	4.0	-0.2 [-0.5, +0.2]	-0.5 [-0.7, -0.4]	-1.4 [-1.5, -1.2]	-1.7 [-1.9, -1.6]
Civil Liberties	1.2	+0.0 [-0.1, +0.2]	-4.7 [-4.9, -4.6]	-1.2 [-1.3, -1.1]	-1.2 [-1.3, -1.1]
Gun Safety	0.7	+0.0 [-0.1, +0.2]	-0.1 [-0.2, -0.0]	-0.6 [-0.7, -0.6]	-0.9 [-1.0, -0.9]

Table 13: Descriptive topic-label diagnostic for media supply and public salience. Article topics are mapped into the main-topic label space. Difference columns report article share minus response share in percentage points with label-level bootstrap 95% confidence intervals. Synthetic columns pool the balanced 12-model comparison across the uniform, simulated, and self-reported context strategies. Rows are sorted by the absolute article–human gap.

Demographic group	Baseline, no context	Uniform context		Simulated context		Self-reported context	
		Current	Lagged	Current	Lagged	Current	Lagged
Party: Democrat	0.420 [0.317,0.523]	0.078 [0.065,0.091]	0.082 [0.065,0.099]	0.097 [0.086,0.107]	0.099 [0.088,0.109]	0.097 [0.083,0.111]	0.094 [0.081,0.107]
Party: Republican	0.404 [0.305,0.503]	0.114 [0.094,0.133]	0.108 [0.091,0.125]	0.085 [0.078,0.092]	0.085 [0.079,0.092]	0.083 [0.073,0.094]	0.083 [0.074,0.092]
Party: Independent	0.394 [0.318,0.471]	0.153 [0.137,0.169]	0.149 [0.130,0.168]	0.107 [0.099,0.114]	0.104 [0.097,0.112]	0.106 [0.097,0.115]	0.102 [0.095,0.109]
Gender: Women	0.395 [0.304,0.487]	0.098 [0.082,0.114]	0.100 [0.081,0.119]	0.096 [0.088,0.105]	0.099 [0.089,0.108]	0.096 [0.083,0.110]	0.095 [0.084,0.107]
Gender: Men	0.377 [0.288,0.466]	0.088 [0.075,0.102]	0.089 [0.075,0.103]	0.083 [0.075,0.090]	0.083 [0.077,0.090]	0.081 [0.071,0.090]	0.078 [0.070,0.087]
Race: White	0.389 [0.297,0.481]	0.092 [0.077,0.106]	0.093 [0.075,0.110]	0.090 [0.082,0.098]	0.092 [0.083,0.100]	0.090 [0.078,0.102]	0.087 [0.077,0.098]
Race: Non-white	0.405 [0.311,0.500]	0.101 [0.085,0.117]	0.102 [0.085,0.118]	0.100 [0.091,0.108]	0.098 [0.090,0.106]	0.096 [0.085,0.107]	0.096 [0.085,0.107]
Age: under 55	0.390 [0.306,0.474]	0.112 [0.099,0.126]	0.108 [0.093,0.122]	0.083 [0.077,0.089]	0.086 [0.079,0.092]	0.082 [0.073,0.091]	0.082 [0.073,0.090]
Age: 55+	0.404 [0.303,0.504]	0.094 [0.077,0.112]	0.098 [0.078,0.119]	0.112 [0.101,0.123]	0.113 [0.103,0.124]	0.112 [0.097,0.128]	0.110 [0.096,0.123]
Education: College	0.395 [0.305,0.485]	0.085 [0.070,0.101]	0.088 [0.069,0.106]	0.083 [0.075,0.091]	0.084 [0.075,0.093]	0.083 [0.071,0.095]	0.082 [0.071,0.092]
Education: No college	0.384 [0.293,0.475]	0.104 [0.090,0.118]	0.103 [0.088,0.119]	0.099 [0.091,0.107]	0.100 [0.092,0.107]	0.096 [0.085,0.107]	0.094 [0.084,0.105]

Table 14: Selected demographic subgroup JSD by context condition for Main topic. Cells report the mean across model runs with 95% jackknife confidence intervals. Collapsed age and race rows average their source subgroups using survey-count weights before pooling.

Model	Party			Age						Gender		Education		Race			
	Democrat	Republican	Independent	18-24	25-34	35-44	45-54	55-64	65+	Female	Male	College	No College	Caucasian	African-American	Asian	Hispanic
Llama-3.2-1B	0.402	0.390	0.373	0.287	0.347	0.363	0.419	0.436	0.426	0.397	0.383	0.390	0.393	0.398	0.374	0.371	0.387
Llama-3.2-3B	0.417	0.401	0.448	0.347	0.363	0.373	0.426	0.446	0.415	0.420	0.384	0.386	0.423	0.406	0.408	0.402	0.485
Qwen3-1.7B	0.333	0.463	0.293	0.312	0.318	0.321	0.363	0.387	0.347	0.338	0.340	0.347	0.341	0.359	0.286	0.296	0.352
Qwen3-4B	0.649	0.435	0.660	0.456	0.462	0.531	0.574	0.584	0.412	0.535	0.480	0.502	0.525	0.499	0.611	0.536	0.537
Qwen3-8B	0.567	0.617	0.440	0.459	0.502	0.509	0.557	0.573	0.546	0.535	0.534	0.573	0.508	0.534	0.620	0.604	0.563
Qwen3-32B	0.115	0.126	0.177	0.146	0.126	0.117	0.126	0.112	0.079	0.125	0.098	0.118	0.124	0.104	0.170	0.134	0.121
Qwen2.5-1.5B	0.414	0.388	0.352	0.366	0.372	0.384	0.416	0.413	0.384	0.402	0.375	0.390	0.391	0.401	0.373	0.412	0.390
Qwen2.5-7B	0.570	0.541	0.455	0.559	0.538	0.492	0.508	0.556	0.603	0.547	0.510	0.533	0.530	0.530	0.541	0.580	0.602
Phi-3.5-mini	0.624	0.689	0.526	0.606	0.607	0.588	0.618	0.622	0.603	0.597	0.604	0.615	0.583	0.613	0.548	0.682	0.579
Gemma-3-4B-it	0.222	0.256	0.283	0.319	0.307	0.292	0.241	0.151	0.270	0.207	0.244	0.262	0.173	0.221	0.215	0.277	0.145
Mistral-Small-24B	0.161	0.107	0.237	0.206	0.151	0.151	0.163	0.136	0.087	0.142	0.122	0.143	0.132	0.128	0.166	0.169	0.178
GPT-OSS-20B	0.563	0.432	0.484	0.424	0.467	0.501	0.510	0.493	0.444	0.499	0.454	0.479	0.486	0.478	0.506	0.526	0.518

Table 15: Model-level subgroup JSD for Main topic under No context. Cells report subgroup point estimates.

Model	Party			Age						Gender		Education		Race			
	Democrat	Republican	Independent	18-24	25-34	35-44	45-54	55-64	65+	Female	Male	College	No College	Caucasian	African-American	Asian	Hispanic
Llama-3.2-1B	0.075	0.095	0.133	0.124	0.115	0.101	0.096	0.073	0.118	0.092	0.080	0.083	0.089	0.083	0.103	0.083	0.092
Llama-3.2-3B	0.074	0.097	0.149	0.118	0.112	0.105	0.095	0.078	0.152	0.093	0.085	0.083	0.094	0.088	0.095	0.085	0.077
Qwen3-1.7B	0.075	0.104	0.137	0.169	0.123	0.108	0.092	0.070	0.095	0.089	0.082	0.076	0.098	0.084	0.092	0.091	0.094
Qwen3-4B	0.052	0.088	0.128	0.108	0.079	0.078	0.067	0.053	0.086	0.066	0.057	0.053	0.073	0.061	0.064	0.055	0.073
Qwen3-8B	0.075	0.119	0.164	0.129	0.119	0.114	0.111	0.092	0.102	0.097	0.093	0.086	0.106	0.095	0.102	0.104	0.116
Qwen3-32B	0.062	0.105	0.151	0.146	0.120	0.105	0.082	0.082	0.068	0.084	0.082	0.071	0.100	0.082	0.090	0.076	0.108
Qwen2.5-1.5B	0.074	0.085	0.135	0.114	0.106	0.097	0.085	0.076	0.130	0.088	0.073	0.076	0.089	0.079	0.091	0.083	0.089
Qwen2.5-7B	0.085	0.119	0.166	0.169	0.132	0.124	0.117	0.093	0.101	0.110	0.096	0.092	0.118	0.100	0.115	0.095	0.144
Phi-3.5-mini	0.068	0.100	0.133	0.099	0.097	0.083	0.093	0.078	0.107	0.085	0.069	0.068	0.091	0.076	0.096	0.073	0.096
Gemma-3-4B-it	0.147	0.178	0.224	0.163	0.164	0.171	0.168	0.154	0.289	0.179	0.147	0.162	0.168	0.162	0.183	0.183	0.160
Mistral-Small-24B	0.069	0.087	0.129	0.135	0.107	0.098	0.086	0.071	0.117	0.081	0.080	0.069	0.096	0.078	0.095	0.071	0.088
GPT-OSS-20B	0.082	0.187	0.182	0.159	0.149	0.143	0.114	0.098	0.133	0.114	0.115	0.105	0.127	0.114	0.125	0.113	0.093

Table 16: Model-level subgroup JSD for Main topic under Uniform context. Cells report subgroup point estimates.

Model	Party			Age						Gender		Education		Race			
	Democrat	Republican	Independent	18-24	25-34	35-44	45-54	55-64	65+	Female	Male	College	No College	Caucasian	African-American	Asian	Hispanic
Llama-3.2-1B	0.083	0.105	0.136	0.130	0.106	0.095	0.101	0.085	0.159	0.101	0.088	0.092	0.099	0.092	0.108	0.113	0.089
Llama-3.2-3B	0.094	0.108	0.139	0.107	0.092	0.115	0.104	0.098	0.203	0.113	0.092	0.097	0.110	0.102	0.101	0.105	0.096
Qwen3-1.7B	0.055	0.080	0.116	0.127	0.102	0.087	0.064	0.061	0.039	0.068	0.066	0.056	0.077	0.065	0.073	0.075	0.080
Qwen3-4B	0.051	0.076	0.131	0.094	0.076	0.076	0.072	0.063	0.093	0.066	0.063	0.058	0.072	0.061	0.072	0.070	0.072
Qwen3-8B	0.072	0.102	0.145	0.124	0.102	0.104	0.095	0.082	0.110	0.091	0.084	0.077	0.102	0.085	0.098	0.089	0.111
Qwen3-32B	0.082	0.109	0.172	0.182	0.138	0.121	0.106	0.091	0.057	0.099	0.098	0.085	0.116	0.097	0.101	0.089	0.121
Qwen2.5-1.5B	0.083	0.095	0.133	0.147	0.105	0.098	0.095	0.082	0.130	0.098	0.079	0.086	0.095	0.087	0.098	0.086	0.108
Qwen2.5-7B	0.065	0.103	0.140	0.116	0.105	0.091	0.097	0.077	0.082	0.088	0.076	0.072	0.097	0.080	0.089	0.082	0.086
Phi-3.5-mini	0.073	0.117	0.136	0.084	0.089	0.097	0.101	0.086	0.143	0.096	0.080	0.081	0.097	0.085	0.103	0.103	0.079
Gemma-3-4B-it	0.169	0.176	0.240	0.176	0.170	0.173	0.175	0.166	0.306	0.193	0.158	0.179	0.177	0.177	0.180	0.213	0.148
Mistral-Small-24B	0.065	0.074	0.121	0.113	0.083	0.084	0.077	0.063	0.138	0.073	0.073	0.063	0.085	0.069	0.087	0.068	0.080
GPT-OSS-20B	0.094	0.154	0.179	0.149	0.144	0.138	0.109	0.100	0.132	0.113	0.112	0.110	0.116	0.111	0.120	0.121	0.118

Table 17: Model-level subgroup JSD for Main topic under Uniform context (lagged). Cells report subgroup point estimates.

Model	Party			Age						Gender		Education		Race			
	Democrat	Republican	Independent	18-24	25-34	35-44	45-54	55-64	65+	Female	Male	College	No College	Caucasian	African-American	Asian	Hispanic
Llama-3.2-1B	0.098	0.095	0.091	0.054	0.071	0.080	0.105	0.117	0.127	0.096	0.089	0.088	0.099	0.094	0.098	0.104	0.092
Llama-3.2-3B	0.102	0.096	0.116	0.063	0.077	0.082	0.117	0.128	0.126	0.109	0.089	0.091	0.109	0.101	0.112	0.103	0.096
Qwen3-1.7B	0.080	0.074	0.079	0.046	0.057	0.062	0.087	0.096	0.093	0.077	0.070	0.067	0.083	0.074	0.085	0.071	0.092
Qwen3-4B	0.107	0.087	0.121	0.064	0.086	0.081	0.118	0.119	0.109	0.106	0.087	0.087	0.109	0.096	0.112	0.099	0.125
Qwen3-8B	0.093	0.079	0.123	0.052	0.070	0.075	0.107	0.110	0.113	0.097	0.078	0.077	0.102	0.087	0.107	0.094	0.110
Qwen3-32B	0.068	0.064	0.098	0.051	0.059	0.061	0.087	0.080	0.081	0.073	0.061	0.058	0.079	0.068	0.077	0.076	0.073
Qwen2.5-1.5B	0.118	0.101	0.105	0.062	0.079	0.088	0.116	0.139	0.136	0.111	0.102	0.100	0.114	0.109	0.110	0.109	0.121
Qwen2.5-7B	0.087	0.072	0.103	0.052	0.066	0.075	0.096	0.098	0.089	0.087	0.072	0.071	0.091	0.081	0.088	0.101	0.082
Phi-3.5-mini	0.126	0.091	0.118	0.073	0.087	0.086	0.124	0.133	0.138	0.114	0.099	0.098	0.118	0.105	0.132	0.121	0.112
Gemma-3-4B-it	0.117	0.097	0.111	0.072	0.093	0.091	0.113	0.124	0.165	0.115	0.097	0.102	0.112	0.109	0.114	0.124	0.088
Mistral-Small-24B	0.071	0.073	0.098	0.049	0.061	0.065	0.091	0.090	0.095	0.080	0.067	0.069	0.079	0.075	0.078	0.084	0.079
GPT-OSS-20B	0.094	0.088	0.114	0.073	0.073	0.086	0.099	0.099	0.105	0.091	0.082	0.086	0.089	0.087	0.093	0.103	0.089

Table 18: Model-level subgroup JSD for Main topic under Simulated context. Cells report subgroup point estimates.

Model	Party			Age						Gender		Education		Race			
	Democrat	Republican	Independent	18-24	25-34	35-44	45-54	55-64	65+	Female	Male	College	No College	Caucasian	African-American	Asian	Hispanic
Llama-3.2-1B	0.099	0.091	0.091	0.058	0.073	0.073	0.118	0.115	0.132	0.100	0.083	0.085	0.100	0.093	0.098	0.092	0.100
Llama-3.2-3B	0.106	0.098	0.110	0.068	0.081	0.086	0.125	0.124	0.141	0.113	0.089	0.094	0.111	0.102	0.115	0.100	0.110
Qwen3-1.7B	0.082	0.073	0.077	0.042	0.059	0.064	0.084	0.098	0.104	0.078	0.072	0.069	0.081	0.074	0.080	0.091	0.087
Qwen3-4B	0.107	0.090	0.116	0.063	0.084	0.094	0.117	0.119	0.114	0.106	0.090	0.087	0.112	0.098	0.107	0.097	0.122
Qwen3-8B	0.093	0.082	0.106	0.054	0.068	0.074	0.110	0.110	0.120	0.099	0.076	0.076	0.104	0.089	0.101	0.077	0.105
Qwen3-32B	0.073	0.066	0.096	0.051	0.062	0.069	0.092	0.084	0.084	0.076	0.067	0.061	0.084	0.072	0.080	0.070	0.078
Qwen2.5-1.5B	0.118	0.102	0.105	0.060	0.089	0.090	0.127	0.132	0.142	0.114	0.100	0.102	0.113	0.110	0.109	0.115	0.114
Qwen2.5-7B	0.089	0.071	0.109	0.059	0.069	0.072	0.102	0.100	0.091	0.088	0.073	0.072	0.092	0.081	0.095	0.077	0.091
Phi-3.5-mini	0.130	0.086	0.123	0.064	0.082	0.094	0.128	0.133	0.138	0.118	0.096	0.103	0.114	0.107	0.123	0.108	0.115
Gemma-3-4B-it	0.124	0.098	0.118	0.072	0.087	0.098	0.125	0.130	0.176	0.122	0.098	0.111	0.112	0.114	0.115	0.122	0.103
Mistral-Small-24B	0.072	0.077	0.094	0.052	0.060	0.070	0.098	0.091	0.096	0.081	0.071	0.069	0.084	0.078	0.078	0.070	0.086
GPT-OSS-20B	0.091	0.086	0.106	0.070	0.076	0.086	0.095	0.094	0.108	0.087	0.083	0.083	0.087	0.083	0.089	0.105	0.099

Table 19: Model-level subgroup JSD for Main topic under Simulated context (lagged). Cells report subgroup point estimates.

Model	Party			Age						Gender		Education		Race			
	Democrat	Republican	Independent	18-24	25-34	35-44	45-54	55-64	65+	Female	Male	College	No College	Caucasian	African-American	Asian	Hispanic
Llama-3.2-1B	0.103	0.092	0.096	0.064	0.072	0.075	0.108	0.119	0.132	0.103	0.083	0.090	0.099	0.095	0.100	0.104	0.098
Llama-3.2-3B	0.113	0.105	0.126	0.062	0.086	0.094	0.121	0.138	0.145	0.120	0.095	0.103	0.115	0.111	0.109	0.118	0.113
Qwen3-1.7B	0.084	0.073	0.085	0.044	0.063	0.059	0.084	0.102	0.094	0.082	0.068	0.071	0.080	0.075	0.079	0.084	0.085
Qwen3-4B	0.116	0.092	0.120	0.073	0.089	0.085	0.114	0.131	0.123	0.112	0.092	0.094	0.112	0.102	0.119	0.114	0.117
Qwen3-8B	0.079	0.063	0.098	0.053	0.061	0.072	0.081	0.088	0.100	0.079	0.066	0.062	0.085	0.072	0.087	0.080	0.085
Qwen3-32B	0.058	0.053	0.085	0.054	0.053	0.058	0.070	0.068	0.071	0.060	0.055	0.051	0.065	0.056	0.069	0.067	0.055
Qwen2.5-1.5B	0.117	0.104	0.108	0.070	0.079	0.083	0.121	0.140	0.140	0.113	0.102	0.104	0.111	0.110	0.110	0.109	0.119
Qwen2.5-7B	0.089	0.075	0.099	0.054	0.068	0.073	0.090	0.105	0.090	0.087	0.075	0.070	0.094	0.081	0.087	0.086	0.104
Phi-3.5-mini	0.130	0.092	0.122	0.059	0.085	0.093	0.123	0.141	0.144	0.118	0.100	0.105	0.115	0.110	0.117	0.112	0.110
Gemma-3-4B-it	0.126	0.108	0.133	0.087	0.101	0.102	0.126	0.139	0.173	0.132	0.100	0.116	0.120	0.120	0.120	0.130	0.113
Mistral-Small-24B	0.059	0.061	0.088	0.047	0.054	0.064	0.068	0.074	0.089	0.065	0.058	0.055	0.069	0.062	0.062	0.069	0.068
GPT-OSS-20B	0.090	0.081	0.112	0.073	0.082	0.094	0.090	0.087	0.097	0.087	0.077	0.079	0.087	0.081	0.090	0.096	0.082

Table 20: Model-level subgroup JSD for Main topic under Self-reported context. Cells report subgroup point estimates.

Model	Party			Age						Gender		Education		Race			
	Democrat	Republican	Independent	18-24	25-34	35-44	45-54	55-64	65+	Female	Male	College	No College	Caucasian	African-American	Asian	Hispanic
Llama-3.2-1B	0.097	0.090	0.094	0.062	0.072	0.076	0.101	0.113	0.134	0.099	0.081	0.085	0.097	0.091	0.105	0.091	0.095
Llama-3.2-3B	0.109	0.100	0.120	0.075	0.078	0.084	0.122	0.129	0.144	0.120	0.085	0.096	0.114	0.106	0.111	0.103	0.112
Qwen3-1.7B	0.079	0.076	0.074	0.049	0.057	0.059	0.087	0.094	0.101	0.080	0.067	0.067	0.080	0.074	0.086	0.071	0.079
Qwen3-4B	0.110	0.092	0.108	0.065	0.075	0.089	0.117	0.121	0.121	0.108	0.088	0.088	0.110	0.097	0.117	0.095	0.116
Qwen3-8B	0.077	0.062	0.101	0.050	0.056	0.061	0.087	0.089	0.104	0.078	0.064	0.065	0.079	0.070	0.083	0.079	0.084
Qwen3-32B	0.057	0.058	0.094	0.047	0.058	0.066	0.072	0.068	0.080	0.064	0.055	0.054	0.067	0.061	0.062	0.068	0.060
Qwen2.5-1.5B	0.115	0.102	0.096	0.080	0.080	0.082	0.114	0.135	0.138	0.111	0.096	0.100	0.109	0.105	0.111	0.111	0.107
Qwen2.5-7B	0.083	0.073	0.100	0.055	0.060	0.068	0.102	0.098	0.086	0.086	0.069	0.071	0.085	0.077	0.088	0.085	0.099
Phi-3.5-mini	0.126	0.091	0.114	0.069	0.079	0.092	0.123	0.133	0.155	0.117	0.096	0.100	0.116	0.106	0.125	0.110	0.119
Gemma-3-4B-it	0.125	0.107	0.117	0.083	0.103	0.098	0.127	0.133	0.173	0.127	0.099	0.113	0.118	0.115	0.127	0.134	0.098
Mistral-Small-24B	0.061	0.068	0.095	0.046	0.057	0.067	0.080	0.083	0.088	0.070	0.064	0.062	0.072	0.067	0.073	0.073	0.071
GPT-OSS-20B	0.087	0.080	0.113	0.080	0.078	0.087	0.088	0.086	0.106	0.086	0.074	0.079	0.084	0.079	0.085	0.107	0.086

Table 21: Model-level subgroup JSD for Main topic under Self-reported context (lagged). Cells report subgroup point estimates.

Condition	Runs	Other / all	No answer / Other	Not-Other / Other	Tail / Other	Tail / all	Tail topics
Real	1	17.1%	65.0%	0.8%	34.2%	5.9%	386
No context	12	2.5%	2.6%	6.5%	90.9%	2.3%	48.9
Uniform context	12	3.7%	0.2%	1.7%	98.1%	3.7%	134.3
Uniform context (lagged)	12	3.9%	0.2%	1.5%	98.2%	3.8%	140
Simulated context	12	8.6%	1.3%	1.0%	97.6%	8.5%	533.9
Simulated context (lagged)	12	8.4%	1.3%	1.1%	97.6%	8.2%	511
Self-reported context	12	8.4%	0.8%	6.6%	92.7%	7.9%	480.8
Self-reported context (lagged)	12	8.2%	0.3%	8.7%	91.0%	7.6%	477.6

Table 22: Composition of responses classified as *Other*. *No answer* is separated from substantive labels, and *Not-Other* counts open-ended topics that exactly match one of the fixed main-topic classifier labels and are therefore excluded from the substantive tail. Synthetic rows report means across model-condition runs within each context strategy; the real row reports the observed survey values.

Condition	Runs	Other share	No-answer count	No-answer / Other	No-answer / all
Real	1	17.1%	2100	65.0%	11.1%
No context	12	2.5%	5.1	2.6%	<0.1%
Uniform context	12	3.7%	0.6	0.2%	<0.1%
Uniform context (lagged)	12	3.9%	0.7	0.2%	<0.1%
Simulated context	12	8.6%	13.4	1.3%	<0.1%
Simulated context (lagged)	12	8.4%	13.6	1.3%	<0.1%
Self-reported context	12	8.4%	7.7	0.8%	<0.1%
Self-reported context (lagged)	12	8.2%	2.8	0.3%	<0.1%

Table 23: Prevalence of *No answer* labels among responses originally classified as *Other*. Synthetic rows report means across model-condition runs within each context strategy; the real row reports the observed survey count.

Context	Other share	Valid cov.	Topics	Entropy	Top-5	Singletons
No context	0.025	1.000	53.7	3.20	0.720	0.087
Uniform context	0.037	1.000	138.7	5.22	0.457	0.112
Uniform context (lagged)	0.039	1.000	145.0	5.22	0.470	0.109
Simulated context	0.086	1.000	541.0	6.96	0.357	0.244
Simulated context (lagged)	0.084	1.000	517.9	6.89	0.363	0.241
Self-reported context	0.084	1.000	489.8	6.87	0.343	0.224
Self-reported context (lagged)	0.082	1.000	486.8	6.84	0.349	0.229

Table 24: Pooled long-tail metrics for responses classified as *Other*. Values are averaged across model-condition runs within each context strategy.

Context	#1	#2	#3	#4	#5
Real	No answer (0.65)	Politics (0.09)	Arts and Entertainment (0.01)	Media Bias (0.01)	Foreign Policy (0.01)
No context	Healthcare (0.42)	Politics (0.16)	Election (0.03)	Climate Change (0.03)	healthcare reform (0.03)
Uniform context	Politics (0.22)	Free Speech (0.05)	Foreign interference (0.03)	Social Media (0.03)	Online Safety (0.02)
Uniform context (lagged)	Politics (0.18)	Sports (0.06)	Foreign interference (0.04)	Social Media (0.04)	Kamala Harris (0.03)
Simulated context	Politics (0.13)	Sports (0.12)	Arts and Entertainment (0.05)	Celebrity (0.02)	Media Bias (0.02)
Simulated context (lagged)	Politics (0.14)	Sports (0.12)	Arts and Entertainment (0.05)	Celebrity (0.02)	Media Bias (0.02)
Self-reported context	Politics (0.12)	Sports (0.10)	Arts and Entertainment (0.06)	Celebrity (0.03)	Entertainment (0.01)
Self-reported context (lagged)	Politics (0.12)	Sports (0.11)	Arts and Entertainment (0.05)	Celebrity (0.02)	Trump's Legal Troubles (0.02)

Table 25: Most frequent open-ended topics among responses classified as *Other*. Shares are within each pooled context strategy.

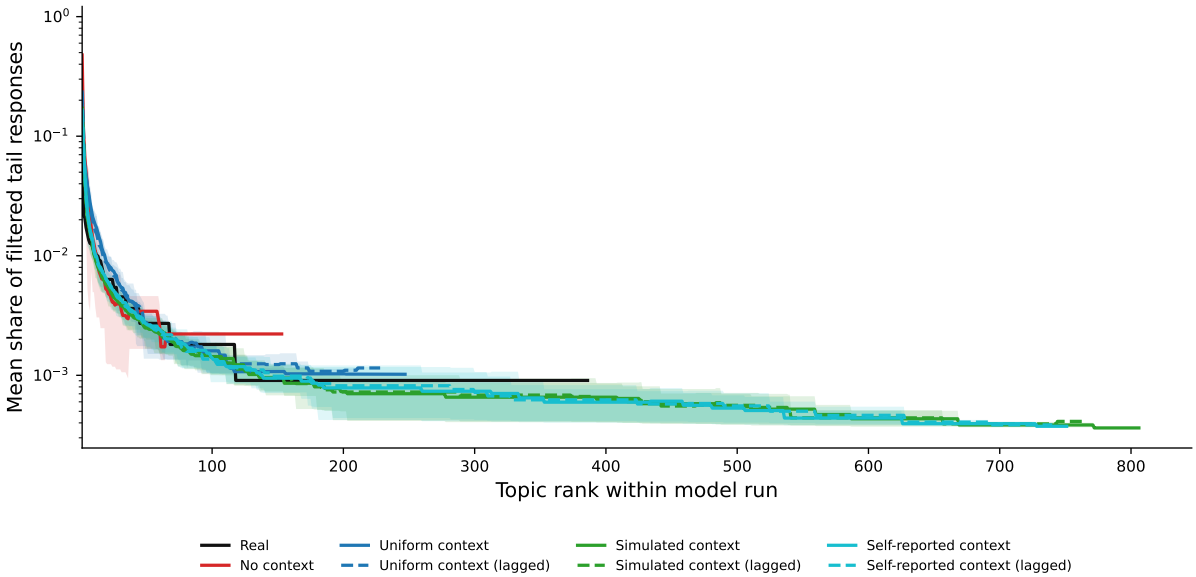

Figure 9: Rank-frequency curves for filtered open-ended tail topics, computed within each model run and then averaged by context strategy. Shaded bands show the 10th–90th percentile across model runs; the real curve is a single observed distribution. The filtered tail excludes *No answer* and exact matches to fixed main-topic labels.

Task	5	10	20	35	Standard
Main topic	0.073 (0.008)	0.074 (0.004)	0.079 (0.006)	0.082 (0.009)	0.074 (0.010)
Sentiment	0.045 (0.031)	0.059 (0.034)	0.075 (0.039)	0.080 (0.037)	0.082 (0.040)

Table 26: Pooled self-reported lagged length cap ablation JSD. Cells report mean across model runs; parentheses report the across-model standard deviation.

Task	Standard	casual	mobile	terse
Main topic	0.074 (0.010)	0.079 (0.012)	0.075 (0.009)	0.069 (0.008)
Sentiment	0.082 (0.040)	0.076 (0.037)	0.077 (0.044)	0.049 (0.040)

Table 27: Pooled self-reported lagged persona style ablation JSD. Cells report mean across model runs; parentheses report the across-model standard deviation.

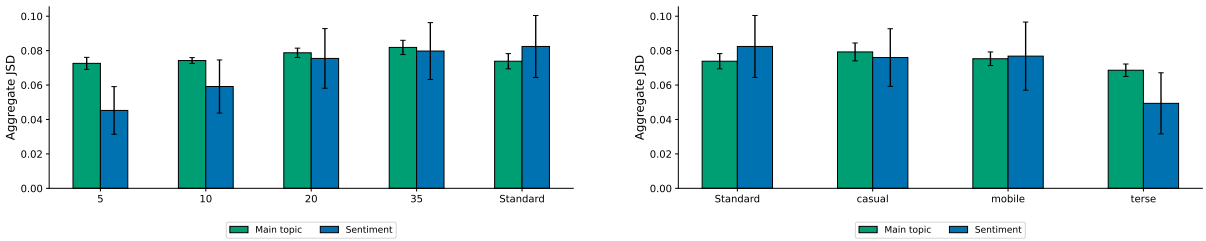

Figure 10: Mean JSD under response-length limits and prompt-style ablations.

A.19 AI Assistant Use

The authors used AI assistants during research and writing for code development, debugging, LaTeX editing, figure and table formatting, and proofreading. AI assistants were not used during the design or deployment of the survey instrument that human-subject participants received.