How Good Are LLM Agents? Benchmarking Opinion Dynamics Against Indonesian Electoral Discourse on Social Media

Musa Izzanardi Wijanarko and Derry Tanti Wijaya
Monash University Indonesia
musa.izzanardi@gmail.com

Abstract

Political polarization on social media intensifies during elections, when echo chambers harden partisan divisions and amplify online harms at the moment collective decisions are being made. Classical opinion dynamics models such as Friedkin-Johnsen provide a framework for studying belief change but abstract away the textual content that drives opinion shifts on social platforms. LLM-powered agent-based models reintroduce language to simulation, yet their fidelity to real political discourse—as opposed to synthetic personas or factual-consensus domains—remains untested against classical baselines. We construct a benchmark from the 2024 Indonesian Presidential Election: ~1,000 agents, 70,336 interactions, and five topics spanning four weeks of Indonesian-language discourse—to our knowledge the first opinion-dynamics benchmark grounded in non-English political data. We propose a two-axis evaluation framework pairing 1-Wasserstein distance with 3-class buffered-sign F1, reported on aggregate and per-topic tracks. Under this framework, we evaluate three open-weight LLMs (Large Language Models) alongside Friedkin-Johnsen and per-agent Ridge regression, analyzing how model family, interaction topology, and sentiment polarity shape simulation fidelity.[1]

1 Introduction

Rising political polarization over the past two decades has weakened public trust in democratic institutions and gridlocked policymaking across major democracies (Iyengar et al., 2019; Törnberg et al., 2021). A growing body of empirical and computational work attributes much of this acceleration to social media: by lowering the cost of selective exposure and rewarding emotionally charged content, online platforms facilitate the formation of echo chambers in which opinions are reinforced rather than challenged (Sasahara et al., 2021). The consequences are most acute during electoral periods, when echo chambers amplify misinformation about candidates and voting procedures and harden partisan divisions just as collective decisions are being made (Wang et al., 2025), Figure 1 depicts online interaction between two people. Forecasting opinion trajectories on real platforms during real elections is therefore not merely a methodological exercise: accurate simulators are a prerequisite for evaluating platform interventions and counter-polarization strategies before deployment (Chuang et al., 2024; Wang et al., 2025), and for identifying the structural conditions under which polarized discourse tips into a self-reinforcing regime (Törnberg et al., 2021).

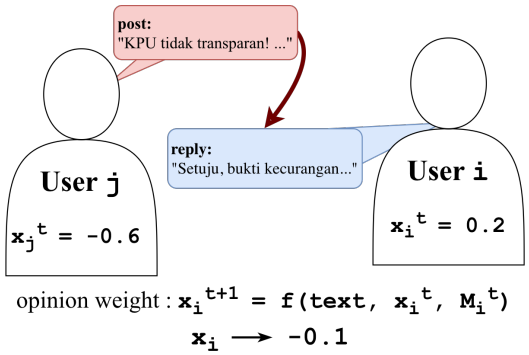


Figure 1: Illustrative example of text-conditioned opinion update. User j posts content; User i react to it, and an opinion classifier f updates x_i based on the textual content and the agent's current state. The Indonesian reads "KPU is not transparent!" (post) and "Agreed, evidence of fraud..." (reply).

Classical opinion dynamics models—DeGroot's averaging update (DeGroot, 1974), the Friedkin-Johnsen extension with stubborn agents (Friedkin and Johnsen, 1990), and bounded-confidence variants (Deffuant et al., 2000; Hegselmann and Krause, 2002)—gave the field a tractable mathematical language for belief change. Their interac-

[1] https://github.com/izzako/abm_political_polarization

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1354–1369
November 6-10, 2026 ©2026 Association for Computational Linguistics

tion effects, however, are encoded entirely as scalar weights and deterministic update rules, with the textual content actually being exchanged on social media abstracted away (Flache et al., 2017). The emergence of LLM-powered agents (Park et al., 2023) has reintroduced language and reasoning to agent-based modeling, and an active line of work now applies this paradigm to opinion dynamics: simulating attitude shifts under confirmation bias (Chuang et al., 2024), modeling partisan political discourse (Argyle et al., 2023), scaling to large social-network simulations (Gao et al., 2025), and reproducing echo-chamber formation (Wang et al., 2025; Gu et al., 2025). Yet across this body of work, simulations are run on synthetic personas or hand-constructed scenarios, with outputs scored through indirect behavioral metrics such as emergent polarization or consensus patterns. Where existing studies do engage real-world issues, they have largely focused on topics with clear factual ground truth (ex: climate change or scientific consensus) rather than the contested, identity-charged discourse that drives electoral polarization. As a result, no study has benchmarked LLM-driven opinion dynamics against real-world political opinion trajectories, nor placed it head-to-head against classical baselines such as DeGroot or Friedkin-Johnsen. This paper aims to address that gap.

A second limitation runs through the existing LLM-ABM literature: opinion is represented as a single scalar per agent, and the multi-topic structure of real political discourse—how a user's stance on one issue conditions their stance on another—is absent from the modeling frame. Real political users, however, hold topic-bundled positions: an agent may endorse one candidate, oppose another, and remain neutral on a procedural matter like a quick-count audit, with the correlation pattern across topics itself a signature of underlying partisan identity (Iyengar et al., 2019). Zuo et al. (2025) make this critique formally and propose a multi-topic LLM simulator, but evaluate only on synthetic networks; whether the cross-topic opinion structure of real users can be recovered by an LLM-based simulator remains untested. We address this directly by structuring our benchmark along five topics derived from Latent Dirichlet Allocation (LDA) methods (§4.1), and evaluating every model under two complementary tracks: an *aggregate* track scored on the pooled multi-topic opinion distribution, and a *per-topic* track scored on each topic's subgraph independently. Reporting both guards against models that produce a plausible mean stance while collapsing within-agent topic heterogeneity (aggregate only) or that miss cross-topic conditioning (per-topic only).

Concretely, this paper makes three contributions:

- **A real-world benchmark** for LLM-driven opinion dynamics on the 2024 Indonesian Presidential Election: 1,000 active Twitter users, 70,336 retweet and reply edges, tagged across five topics. We construct the network from interaction signals rather than follow relationships, since interaction networks carry partisan structure comparable to follow networks (Conover et al., 2011; Barberá, 2015) and remain accessible under contemporary platform restrictions that have largely shut down academic access to follow data (Murtfeldt et al., 2025; Blakey, 2024). To our knowledge this is the first opinion-dynamics benchmark grounded in non-English political discourse.

- **A two-axis evaluation framework** pairing a macroscopic distributional metric (1-Wasserstein) with a microscopic per-agent metric (3-class F1), each reported under aggregated data and per-topic tracks.

- **A comparative evaluation** of three open-weight LLMs against Friedkin-Johnsen and per-agent Ridge baselines under a shared protocol. We find that no single model dominates both axes: Ridge classifies individual agents more accurately, while LLMs better recover the overall shape of opinion distributions on smaller topics.

2 Recent Works

Classical opinion dynamics models—DeGroot (DeGroot, 1974), Friedkin-Johnsen (Friedkin and Johnsen, 1990), and bounded-confidence variants (Deffuant et al., 2000; Hegselmann and Krause, 2002)—are analytically tractable but rely on deterministic scalar update rules that abstract away linguistic and contextual dimensions of real discourse (Flache et al., 2017; Chuang et al., 2024). The advent of LLM-powered agents (Park et al., 2023) has introduced language and reasoning to social simulation, and this capability has been directly applied to opinion dynamics in a growing body of work.

Chuang et al. (2024) demonstrate that LLM agents qualitatively replicate attitude shifts and opinion fragmentation under confirmation bias; Gao et al. (2025) scale LLM-agent simulation to large social networks with emergent information diffusion; and Wang et al. (2025) and Gu et al. (2025) study echo-chamber formation. Further work explores LLM agents for political discourse (Argyle et al., 2023) and ideological alignment (Hwang et al., 2023).

These efforts share a common evaluation shape: simulations run under synthetic conditions, scored through emergent phenomena rather than observed opinion data. Where real issues are engaged, the focus is on factual-consensus domains such as climate change (Chuang et al., 2024) or structural clustering rather than per-agent trajectories (Gu et al., 2025), and none adopt classical baselines as reference points. Whether LLM agents actually produce more faithful simulations of real political opinion dynamics than analytically simple classical models remains unanswered.

3 Method

3.1 Problem Definition

We formalize opinion dynamics as a forecasting task over social media users. Each user is treated as an agent whose stance on a topic, which expressed through posting, replying, and retweeting, that evolves through interaction with other agents. Given a user's observed activity history up to a cutoff time, the objective is to predict their future opinion. Let $A = \{1, 2, \ldots, N\}$ be a set of N number of agents. Each agent $i \in A$ has an opinion weight $x_i^t \in [-1, 1] \subseteq \mathbb{R}$ at timestep t. These opinion weights evolve through expressed activity; we defined this space as $\tau = \{\text{retweet}, \text{reply}, \text{original post}\}$, and the values -1 and 1 represent the strongest negative and strongest positive opinions, respectively.

If we have a sequence of user activity up to time T defined by

$$\mathcal{H}_i = \{(a_i^{(k)},\ x_i^{(k)},\ t^{(k)})\}_{k=0}^{I}, t(I) \leq T$$

where user i performs an activity $a_i \in \tau$ with opinion x_i at time t. We aim to predict user's opinion x_i^θ at a random time step $\theta > T$.

3.2 Classical Models

The DeGroot averaging model (DeGroot, 1974) is one of the foundational opinion dynamics models, in which each agent updates its opinion as a weighted average of its neighbours' opinions via a row-stochastic interaction matrix $\mathbf{W} = (w_{ij})_{N \times N}$.

Friedkin-Johnsen (Friedkin and Johnsen, 1990) extends the DeGroot model by introducing a susceptibility value. It assumes that agents preserve their initial opinion by a certain degree, that is,

$$x_i^{t+1} = s_i \sum_{j \in A} w_{ij} x_j^t + (1 - s_i) x_i^0$$

where w_{ij} is the weight of interaction of i and j, $s_i \in [0, 1]$ is the agent i susceptibility to external influence, meaning $1 - s_i$ represents stubbornness, the extent to which it holds to its own initial opinion x_i at $t = 0$. Fully susceptible agent ($s_i = 1$) reduces to the DeGroot update, while a fully stubborn agent ($s_i = 0$) retains its initial opinion indefinitely.

3.3 Regression Model

As an additional baseline, we fit a per-agent Ridge regression that learns how each agent's next opinion depends on its current stance, interaction partners, and partner opinions, and is applied autoregressively at inference.

At each timestep t, let $b_{ij}^t \in \{0, 1\}$ denote whether agent i interacted with agent j at timestep t through any activity $a \in \tau$, and let $\mathbf{b}_i^t = (b_{i1}^t, \ldots, b_{iN}^t)^T$. For each agent i, we construct a feature vector:

$$\mathbf{z}_i^t = \left[\ x_i^t,\ (x_i^t)^2,\ \bar{x}_{\mathcal{N}_i}^t,\ d_i^t,\ \mathbf{b}_i^t,\ \mathbf{b}_i^t \odot \mathbf{x}^t\ \right]$$

where $d_i^t = \sum_j b_{ij}^t$ is the number of interactions, $\bar{x}_{\mathcal{N}_i}^t$ is the mean opinion of interacted neighbors, and $\odot$ is the Hadamard product.

Each agent i is assigned a Ridge regression model with parameters $\boldsymbol{\theta}_i$ trained on designated period as training set:

$$\boldsymbol{\theta}_i^* = \text{argmin}_{\boldsymbol{\theta}_i} \sum_{t=1}^{T_{\text{train}}} \left(x_i^t - \boldsymbol{\theta}_i^T \tilde{\mathbf{z}}_i^{t-1} \right)^2 + \lambda \|\boldsymbol{\theta}_i\|_2^2$$

where $\tilde{\mathbf{z}}_i^t$ is the standardized feature vector. At inference, predictions are rolled forward autoregressively:

$$\hat{x}_i^{t+1} = \text{clip}\Big(\boldsymbol{\theta}_i^{*T} \tilde{\mathbf{z}}_i^t,\ -1,\ 1 \Big)$$

where each step's output $\hat{\mathbf{x}}^t$ is fed back as input for the next. Construction details and hyperparameters will be described in Section 4.

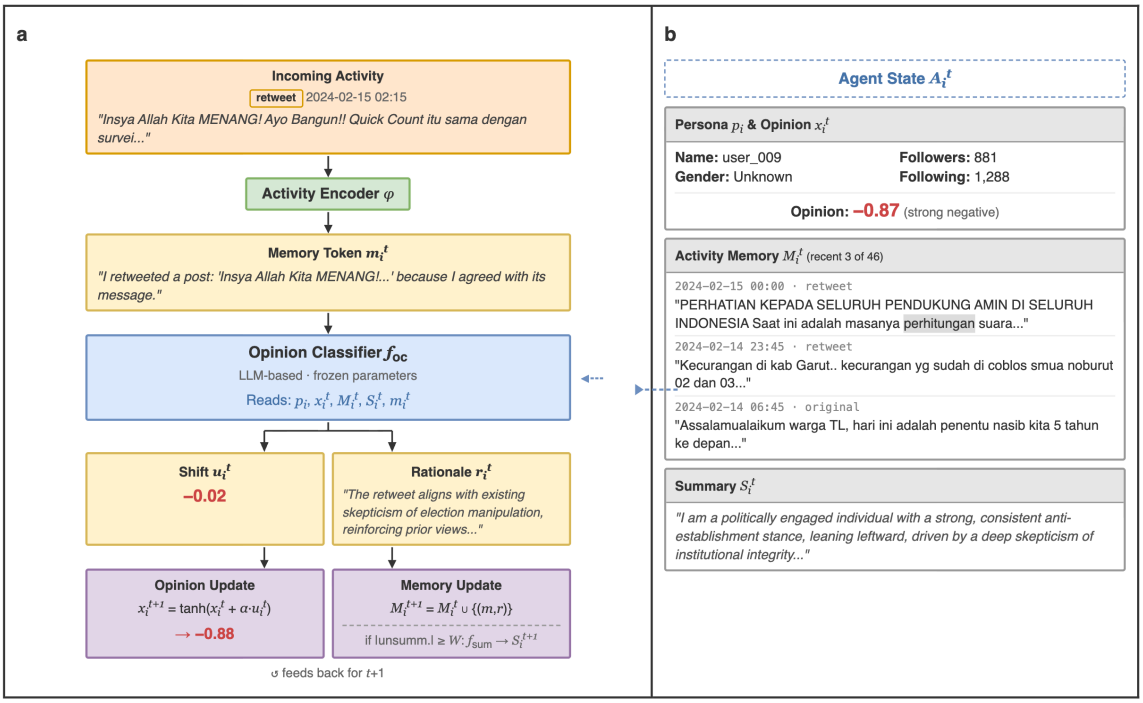


Figure 2: Overview of the LLM-enhanced agent-based model. **(a)** Single-timestep pipeline: an activity is encoded into a memory token and passed with the full agent state to the opinion classifier f_{oc}, which returns a shift u_i^t and rationale r_i^t. The shift updates the opinion via a bounded $\tanh$ rule; the rationale is appended to memory. When unsummarized entries exceed window W, the summarizer f_{sum} refreshes the compressed summary. **(b)** Agent state $\mathcal{A}_i^t$: persona p_i, opinion x_i^t, activity memory $\mathcal{M}_i^t$, and summary $\mathcal{S}_i^t$.

3.4 LLM-Enhanced Agent-Based Model

We propose an agent-based framework (Figure 2) in which each opinion update is mediated by an LLM that reasons over the agent's persona, accumulated memory, and the most recent activity. The per-step opinion shift is thus a learned, context-dependent function over natural-language records rather than fixed weights or hand-crafted features.

Each agent $i \in A$ is represented at timestep t by the tuple

$$\mathcal{A}_i^t = \left(p_i,\ x_i^t,\ \mathcal{M}_i^t,\ \mathcal{S}_i^t\right)$$

where p_i is a static persona profile encoding the agent's account-level attributes (e.g. gender, follower and following counts), $x_i^t \in [-1, 1]$ is the opinion weight defined in Section 3.1, $\mathcal{M}_i^t = \left(m_i^{(1)}, \ldots, m_i^{(L_i^t)}\right)$ is the agent's chronologically ordered activity memory of length L_i^t, and $\mathcal{S}_i^t$ is a natural-language summary of older memory entries that compresses long-range context.

Activity Encoding. At every timestep t, the data stream yields activities $\mathcal{E}^t = \{(i_k, a_k, c_k)\}_k$, where $i_k \in A$ is the acting agent, $a_k \in \tau$ is the activity type, and c_k is the associated textual content (for replies and retweets, c_k includes the parent tweet). A deterministic encoder $\phi : \tau \times \mathcal{T} \to \mathcal{T}$ maps each activity into a structured natural-language memory token $m_{i_k}^t = \phi(a_k, c_k)$.

Opinion Classifier. Let f_{oc} denote an LLM-based opinion classifier with frozen parameters. Given the most recent activity token m_i^t, the agent's persona p_i, current opinion x_i^t, the classifier predicts a scalar opinion shift jointly with a natural-language rationale:

$$(r_i^t,\ u_i^t) \ = \ f_{\text{oc}}\left(p_i,\ x_i^t,\ \mathcal{M}_i^t,\ \mathcal{S}_i^t,\ m_i^t\right)$$

where $u_i^t \in [-1, 1]$ is the relative shift induced by the activity (-1 for a strong negative shift, 0 for no change, $+1$ for a strong positive shift), and $r_i^t \in \mathcal{T}$ is the rationale, which is appended to memory and made available to subsequent classifier calls. The classifier thus conditions on the full agent state to produce both a directed opinion shift and an accompanying rationale. The rationale is retained in the agent's memory, allowing the next classification steps to access the model's prior reasoning.

Opinion Update. The opinion weight is updated by a bounded, learning-rate-controlled rule that

aggregates the LLM-predicted shift,

$$x_i^{t+1} \;=\; \tanh\big(\, x_i^t \;+\; \alpha\, u_i^t \,\big)$$

with $\alpha > 0$ a global learning rate, and the hyperbolic tangent enforces $x_i^{t+1} \in (-1, 1)$. Agents not involved in any activity at t retain their previous opinion, $x_i^{t+1} = x_i^t$.

Memory Update and Summarization. After classification, the activity token and rationale are appended to memory, $\mathcal{M}_i^{t+1} = \mathcal{M}_i^t \cup \{(m_i^t, r_i^t)\}$. To bound prompt length, a separate summarizer f_{sum} refreshes the compressed summary whenever unsummarized entries reach a fixed window W:

$$\mathcal{S}_i^{t+1} \;=\; f_{\text{sum}}\big(\mathcal{M}_i^{t+1}[L_i^{t+1} - W + 1 : L_i^{t+1}]\big)$$

Initialization. Prior to simulation, each agent is seeded with a warm-up window of historical activity $\mathcal{H}_i^{(0)} \subseteq \mathcal{H}_i$ observed before the simulation start time T_0. The initial opinion x_i^0 is set to the empirical mean of opinions in $\mathcal{H}_i^{(0)}$, and the initial summary $\mathcal{S}_i^0 = f_{\text{sum}}(\mathcal{H}_i^{(0)})$ bootstraps the agent's persona-grounded memory.

Simulation. The complete procedure is summarized in Appendix A. Iterating the loop over $t = 1, \ldots, T$ yields a per-agent opinion trajectory $\{x_i^t\}_{t=0}^T$ that constitutes the model's prediction for the inference task defined in Section 3.1. Unlike the classical and regression baselines, inter-agent influence enters implicitly through textual content and the LLM's language priors rather than through fixed weights or hand-crafted features.

4 Experiments Setup

4.1 Benchmark Dataset Preparation

We curated a dataset of Twitter (now X) discourse from the 2024 Indonesian Presidential Election in Indonesian, designed to capture the interaction patterns and sentiment trajectories required for agent-based simulation.

Data Collection. First, we acquired a 5% random sample of election-related discourse via the Brandwatch platform, covering the full month of February 2024 in order to bracket the 14 February general-election date with approximately two weeks of pre- and post-election discourse. We further restricted the sample to posts with ≥ 100 impressions, retaining only content with non-trivial audience exposure and suppressing spam and effectively unseen documents. The resulting dataset comprised $N = 497{,}364$ documents.

Agent Population and Interaction Topology. We filtered for active, opinionated users in two steps. First, using a fine-tuned Indonesian polarization classifier (Susanto et al., 2025), we kept authors whose polarizing output exceeded 50% of at least 30 activities, dropping quiet or politically lukewarm accounts. From this pool we took the top 1,000 by normalized Shannon entropy across the four activity types (Post, Reply, Quote, Retweet). A 100%-sample crawl on these accounts yielded $n = 98{,}365$ raw interactions; after removing orphans (deleted or private parent tweets), the final dataset has 70,336 edges (interactions) over 948 nodes (users), or graph density 0.08.

This filtering targets the engaged partisan core, where echo-chamber dynamics are most observable and where agents have enough activity to identify stance change. The benchmark therefore characterizes active partisan users rather than the platform population as a whole.

Opinion Space and Sentiment Attribution. We applied LDA (Blei et al., 2003) to identify five thematic axes (K=5, C_v=0.34): (1) *General Election*, (2) *Anies Baswedan*, (3) *Prabowo Subianto*, (4) *Ganjar Pranowo*, and (5) *Quick Count Analysis*. Each interaction was then labelled by a fine-tuned Indonesian RoBERTa sentiment classifier (Wilson Wongso, 2023) (F1_M=92.42%). The resulting sentiment–topic pairings define the *opinion weights* in which agent states evolve. Their per-topic distribution is reported in Table 1, and their daily evolution across the collection window in Appendix H.

We selected K=5 by scanning K=2–8 and taking the peak C_v. LDA requires the number of topics to be fixed in advance, which is a limitation relative to newer neural alternatives; we retained it for interpretability, reproducibility, and comparability with prior opinion-dynamics work. The absolute coherence (0.34) is modest but expected for short, informal election tweets, so we rely primarily on face validity: the five topics map onto recognizable real-world entities (the three candidates, the general-election frame, and the quick-count audit). Appendix F reports the ten most salient terms per topic.

Ground-Truth Validation. We treat expressed sentiment as a directional proxy for stance, following standard opinion mining practice (Pang and Lee, 2008; Liu, 2012) and recent opinion-dynamics work that uses Twitter sentiment as opinion trajectories (Zareer and Selmic, 2024). To validate the

labels, a native Indonesian speaker independently annotated a stratified sample of 250 interactions (133 negative, 66 neutral, 51 positive) on the same $\{-1, 0, +1\}$ scale. Agreement is 70.8% (weighted-F1 0.709, macro-F1 0.679; per-class F1: negative 0.80, neutral 0.54, positive 0.70). The 29% disagreement is systematic rather than random: the classifier under-detects implicit negative sentiment, absorbing sarcasm and allegations of corruption into the neutral class (annotation protocol and recurring disagreement patterns in Appendix G).

Two properties make our conclusions robust to this residual noise. *Ranking invariance*: all five models are scored against the identical label set under identical metrics, so label noise shifts every score together and preserves their relative ordering, which is the object of every claim we make. We accordingly treat rank stability, not score stability, as the criterion of robustness throughout. *Conservative direction*: since the dominant error under-detects negativity, the true signal is more negative than our labels, so our finding that instruction-tuned LLMs skew positive on negative-dominant topics is a lower bound rather than an artifact.

Topic	Negative	Neutral	Positive	Total
General Election	31,346	14,974	10,363	56,683
Anies Baswedan	211	367	271	849
Prabowo Subianto	6,246	1,828	3,117	11,191
Ganjar Pranowo	78	461	732	1,271
Quick Count	169	133	40	342
Total	**38,050**	**17,763**	**14,523**	**70,336**

Table 1: Per-topic sentiment distribution across the 70,336 curated interactions.

4.2 Data Description

The data spans from 1-28 February 2024, anchored around voting day of 14th February. Our dataset consists of 35,374 retweets, 6,188 replies, and 28,774 original posts.

The five topics are highly imbalanced in both volume and sentiment (Table 1). Aggregate stances shift across the window and cross-agent dispersion widens around 14 February; at the agent level, a small fraction of users account for most activity while the majority contribute sparingly. Volume and participation peak on election day and decay thereafter, while mean opinion drifts negative on every topic (Appendix H).

We adopt 14 February as a natural temporal pivot, treating the pre-election period (1-14 February) as a warm-up window for agent initialization and 15-28 February as the prediction window. The differing start dates of the per-topic ground-truth signals and the strong topic imbalance jointly motivate two evaluation tracks: a per-topic track operating on each topic subgraph, and an aggregate track operating on all topics combined.

Post-Election Political Scene. Voting closed on 14 February, but the official results were not certified until 20 March, placing the entire prediction window between the two. Unofficial quick-count projections gave Prabowo–Gibran a decisive lead within hours of the polls closing, the trailing camps disputed both the projections and the tabulation, and attention moved from advocating candidates to contesting the conduct of the count. Quick Count is the only topic whose volume rises after the vote—133 interactions before against 209 after, where every other topic contracts—and it turns from -0.19 to -0.50 as it grows. General Election moves from -0.24 to -0.50 over the same span as fraud vocabulary displaces campaign vocabulary (*kecurangan*, *kpu*; Appendix F). This is what widens cross-agent dispersion around 14 February, and it leaves models scored across a regime change rather than a continuation of the warm-up window—one reason every model degrades as the horizon extends (Section 5).

4.3 Model Setup and Implementation Details

We instantiate three models under a shared evaluation protocol (Section 4.4): Friedkin-Johnsen as a classical reference, a per-agent Ridge regression as a learned but linear baseline, and our proposed LLM-driven ABM. Full hyperparameter values are listed in Appendix D.

4.3.1 Friedkin-Johnsen

We follow the notation in Section 3.2, where $s_i \in [0, 1]$ is agent i's susceptibility and $\mathbf{\Lambda} = \mathrm{diag}(\mathbf{s})$. Both the interaction matrix $\mathbf{W}$ and the susceptibility vector $\mathbf{s}$ are derived from the warm-up window $\mathcal{H}_i^{\text{init}} = \{k \in \mathcal{H}_i \mid t^{(k)} \leq \text{Feb } 1 + t_{\text{init}}\}$: edge weights aggregate sentiment-conditioned reply counts and retweet counts under a small set of scalar hyperparameters, and s_i is derived from the share of agent i's activity directed at itself. The initial opinion x_i^0 is taken as the mean over agent i's warm-up tweets and retweet-inherited sentiments under the mapping $\{\text{positive: } +1, \text{ neutral: } 0, \text{ negative: } -1\}$. The

simulation rolls forward on a 2.4-hour grid from Feb $1 + t_{\text{init}}$ to Feb 28; the LLM-ABM runs on a finer grid, and all predictions are averaged to the daily level before scoring (Section 4.4). All scalar hyperparameters are jointly grid-searched on the training fold; full construction details are given in Appendix B, with hyperparameter values in Appendix D.

4.3.2 Regression Baseline

The feature vector $\mathbf{z}_i^t$, Ridge objective, and auto-regressive update rule were defined in Section 3.3. The Ridge target is the agent's *daily mean expressed sentiment* under the mapping $\{\text{positive}: +1, \text{neutral}: 0, \text{negative}: -1\}$, so the regression's task is closer to next-day expression prediction than to opinion-trajectory tracking. Features are standardized per agent via a `StandardScaler` fitted on the training fold (Feb 1–14) and reused at inference. During roll-forward over Feb 15–28, interaction indicators $\mathbf{b}_i^t$ are taken from observed data while opinions $\hat{\mathbf{x}}^t$ are fed back from the model's own predictions. Output is forward-held to the 2.4-hour grid for comparability. Hyperparameters are listed in Appendix D.

4.3.3 LLM-Powered Agent Based Model

Model Choices and Serving. We evaluate three open-weight models under identical prompts and protocol: `google/gemma-3-12b-it` (Gemma Team et al., 2025), `Qwen/Qwen3-8B-FP8` (Yang et al., 2025), and `nvidia/Llama-3.1-Nemotron-Nano-8B-v1` (Bercovich et al., 2025). All three sit in the ∼8–12B parameter class, are instruction-tuned with explicit reasoning behaviour, and have multilingual capability including Indonesian. Inference is served via vLLM (Kwon et al., 2023) and accessed through LangChain's[2] `ChatOpenAI` client. All experiments were run on a single NVIDIA A40 (48 GB).

Decoding and Structured Output. Sampling uses temperature, top-p, and top-k. Both f_{oc} and f_{sum} enforce Pydantic-validated structured output so malformed generations are rejected at the schema layer. Prompt templates are listed in Appendix C.

Simulation Grid and Agent State. The simulation runs on a 15-minute grid (the dataset's native timestamp granularity) from Feb $1 + t_{\text{init}}$ to Feb 28. At each step the simulator gathers all active agents, encodes their activities via φ (Section 3.4), and dispatches them to f_{oc} in batches.[3] The persona p_i and seed summary S_i^0 are produced once during warm-up by a single summarizer call on each agent's warm-up history $\mathcal{H}_i$. The memory window W and opinion step α follow Section 3.4.

[2] https://github.com/langchain-ai/langchain

4.4 Evaluation Metrics

Model predictions and ground-truth signal $\bar{x}_i^t$ are aggregated to the daily level before scoring. We evaluate along two complementary axes that can diverge sharply within the same system (Lanzetti et al., 2023). **Macroscopic:** the 1-Wasserstein distance W_1 between predicted and ground-truth opinion distributions on each day, capturing both location and dispersion (Panaretos and Zemel, 2019; Lanzetti et al., 2023). **Microscopic:** per-agent 3-class F1, where opinion weights are binned into *negative* $[-1, -0.25)$, *neutral* $[-0.25, 0.25]$, and *positive* $(0.25, 1]$. Given heavy class imbalance, we treat macro F1 as the primary score, with weighted F1 reported for completeness. The ± 0.25 boundaries were selected by scanning the range 0.1–0.5. Both metrics are averaged (MAE) over the prediction window; all scores are reported per topic and on the full dataset.

5 Results

Table 2 reports macro F1, weighted F1, and Wasserstein MAE across the five per-topic tracks and the aggregate track. Ridge obtains the highest macro F1 on four of six tracks (General, Prabowo, Ganjar, Aggregate) and ties with Friedkin-Johnsen on Quick Count. Qwen3-8B achieves the highest aggregate weighted F1 (0.71); Nemotron-Nano-8B leads on Anies ($\text{F1}_{\text{M}} = 0.84$). On Wasserstein distance, the ranking reverses: LLM-based models achieve the lowest W_1 on four of five per-topic tracks—Gemma-3-12B and Qwen3-8B tie on Prabowo (0.05), Qwen3-8B leads on Ganjar (0.06) and Quick Count (0.16), and Gemma-3-12B ties with Ridge on Anies (0.07). Ridge retains the lowest W_1 on General (0.07) and the aggregate track (0.12).

LLMs produce opinion distributions closer to the ground truth on minority topics but misclassify individual agents more often than Ridge. The gap is largest on General Election, which accounts

[3] Batch size 50, max concurrency 30. On batch failure, affected agents receive $u_i^t = 0$ so no spurious update is recorded. Per-agent state is persisted to JSON between steps.

Model	General			Anies			Prabowo			Ganjar			Quick Ct.			Agg.		
	$F1_M$	$F1_W$	W_1	$F1_M$	$F1_W$	W_1	$F1_M$	$F1_W$	W_1	$F1_M$	$F1_W$	W_1	$F1_M$	$F1_W$	W_1	$F1_M$	$F1_W$	W_1
Friedkin-Johnsen	0.51	0.66	0.14	0.80	0.80	0.10	0.62	0.69	0.13	0.82	0.83	0.12	**0.74**	**0.74**	0.19	0.39	0.57	0.21
Ridge	**0.59**	**0.72**	**0.07**	0.83	0.83	0.07	**0.66**	**0.73**	0.06	**0.87**	**0.88**	0.07	**0.74**	**0.74**	0.20	**0.49**	0.66	**0.12**
Gemma-3-12B	0.44	0.63	0.17	0.82	0.82	0.07	0.62	0.69	**0.05**	0.86	**0.88**	0.07	0.68	0.71	0.19	0.43	0.64	0.17
Qwen3-8B	0.47	0.68	0.24	0.83	0.83	**0.06**	0.61	0.69	**0.05**	0.86	**0.88**	**0.06**	0.69	0.73	**0.16**	0.47	**0.71**	0.19
Nemotron-Nano-8B	0.34	0.47	0.38	**0.84**	**0.84**	0.07	0.58	0.65	0.12	0.86	0.87	0.08	0.69	0.71	0.24	0.37	0.52	0.35

Table 2: Main results across per-topic and aggregate evaluation tracks. $F1_M$: macro F1 (↑); $F1_W$: weighted F1 (↑); W_1: Wasserstein Distance (↓). Best value per column in **bold**.

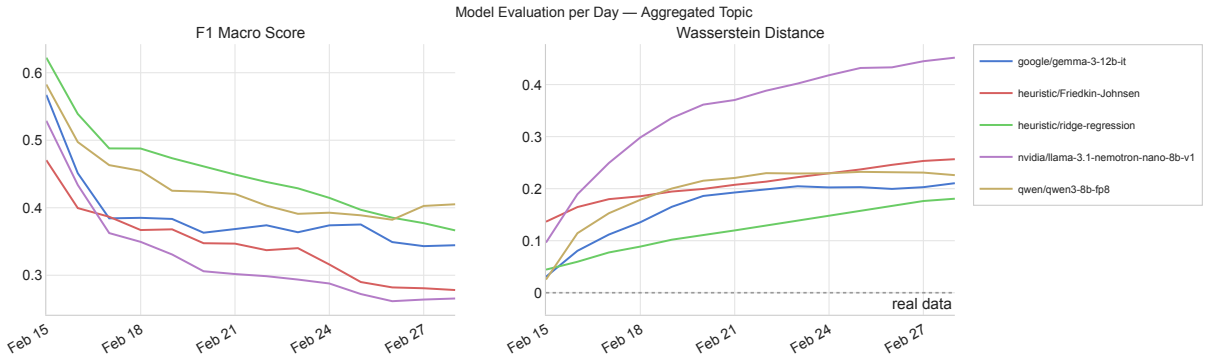


Figure 3: Daily macro F1 (left, ↑) and Wasserstein distance (right, ↓) on the aggregate track over the 14-day prediction window (Feb 15–28). Dashed line marks the ground-truth reference ($W_1 = 0$).

for over 80% of interactions and skews heavily negative (55% negative sentiment; Table 1). On positive-leaning topics such as Ganjar (58% positive), LLMs nearly match Ridge on F1 (0.86 vs. 0.87) while matching or beating it on W_1. Whether this asymmetry reflects sentiment bias in instruction-tuned models or simply data volume is examined in Section 6.

Figure 3 plots both metrics per day on the aggregate track. All models degrade over the prediction window: macro F1 drops from 0.50–0.63 on Feb 15 to 0.30–0.42 by Feb 27, and Wasserstein distance rises monotonically for every model. Ridge degrades the most gracefully—it starts with the highest F1 and the gap over other models widens as the horizon extends. On Wasserstein, all models start at comparable levels ($W_1 \approx 0.03$–0.10 on Feb 15), but their trajectories diverge: Ridge and Gemma-3-12B remain below 0.20 throughout, Qwen3-8B drifts to roughly 0.30, while Nemotron-Nano-8B and Friedkin-Johnsen exceed 0.45 by the end of the window.

Nemotron-Nano-8B diverges sharply on the aggregate track as the simulation progresses—its autoregressive errors compound faster than those of the other LLMs. Friedkin-Johnsen exhibits similar drift, consistent with a deterministic model that cannot self-correct. Ridge's per-agent parameterization anchors predictions to individual behavioural patterns, slowing error accumulation. Gemma-3-12B and Qwen3-8B occupy a middle ground—their language-conditioned memory provides partial error correction but does not fully prevent compounding.

The Frobenius distance between each model's predicted cross-topic correlation matrix and the ground truth is lowest for Ridge and Nemotron-Nano-8B (0.40 each) and highest for Gemma-3-12B (0.74), indicating that per-agent learned parameters recover cross-topic opinion structure more reliably than language-conditioned updates alone (full analysis in Appendix E).

6 Discussion

Per-agent precision vs. distributional fidelity. The consistent split between the two evaluation metrics—Ridge leading on per-agent F1, LLMs leading on Wasserstein for minority topics—reflects a structural difference in how the model types represent agents. Ridge fits a separate coefficient vector $\boldsymbol{\theta}_i$ for each agent, directly encoding individual behavioural patterns from the training window. This per-agent parameterization makes it effective at classifying *who* holds *which* stance, but it cannot generalize beyond the feature combinations observed during training. LLMs, by contrast,

share a single set of frozen parameters across all agents and condition opinion updates on natural-language memory rather than per-agent features. The result is a model that captures population-level distributional patterns—the relative proportions of negative, neutral, and positive agents—without reliably pinning the correct class to the correct individual. This trade-off is sharpest on the General Election topic, where Ridge has abundant per-agent training signal, and narrowest on minority topics like Ganjar and Anies, where sparse data limits what per-agent fitting can learn and the LLM's language priors compensate.

Sentiment polarity and instruction tuning. The F1 gap between LLMs and Ridge covaries with topic sentiment composition: it is narrowest on positive-leaning Ganjar and widest on negative-dominant General Election (Table 1). Data volume confounds—General Election has $67\times$ more interactions than Ganjar—but the direction of the asymmetry is consistent with a known property of instruction-tuned models: RLHF training rewards agreeable, constructive outputs, producing a positivity bias that works against faithful simulation of sustained hostile discourse. All three LLMs are instruction-tuned, and the pattern holds across all three despite differing base architectures. Disentangling sentiment bias from data volume would require controlled downsampling experiments, which we leave to future work, but the correlation warrants caution when deploying LLM-based simulators on negative-dominant political discourse.

Qualitative error analysis. Two agents from the General Election topic illustrate where language-conditioned updates help and where they misfire. Agent A is a high-activity user (365 interactions) whose summarized memory describes a "consistent critic of the KPU… anti-establishment stance." Ground-truth opinion is persistently negative ($\bar{x} = -0.45$); Qwen3-8B tracks this correctly ($\bar{x} = -0.79$, 100% class accuracy), while Ridge drifts toward neutral ($\bar{x} = -0.19$, 0% accuracy)—its linear features average out the mixed-sentiment interactions this agent is exposed to, losing the ideological consistency that the LLM's memory preserves. Agent B shows the opposite failure. The summarised memory identifies a "staunch pro-Prabowo supporter… dismissing fraud claims," and the LLM shifts this agent's opinion toward positive ($\bar{x} = +0.50$), apparently conflating partisan *support* for a candidate with *positive* sentiment toward the General Election topic. The ground truth is negative ($\bar{x} = -0.64$); Ridge recovers this faithfully ($\bar{x} = -0.62$). The failure exposes a reasoning shortcut in the LLM: political identity cues in the persona override the actual opinion signal, a form of sentiment–identity conflation that scalar baselines are immune to by construction.

7 Conclusion

We constructed a benchmark from the 2024 Indonesian Presidential Election—1,000 agents, 70,336 interactions, five topics—and evaluated LLM-driven opinion dynamics against this real political discourse. Under a two-axis framework pairing 1-Wasserstein distance with 3-class buffered-sign F1, no single model family dominates both axes: Ridge classifies individual stances most accurately, while LLMs better recover population-level distributions on minority topics. The gap covaries with sentiment polarity—LLMs struggle on negative-dominant discourse, consistent with positivity bias in instruction-tuned models—and Friedkin-Johnsen underperforms all other models, confirming that fixed scalar weights are insufficient for this task.

Limitations

Construct validity. The ground-truth opinion weights come from a fine-tuned sentiment classifier ($F1_M$=92.42%) rather than exhaustive human annotation, so classifier errors propagate into both the training signal and the evaluation target. Our validation study (§4.1) bounds this: agreement with human judgement is 70.8% on a 250-item sample, and the dominant error direction is conservative with respect to our conclusions. That study used a single annotator, so it measures agreement rather than inter-annotator reliability. Our prompt compounds this, reading partisan identity as sentiment toward the topic (Agent B, §6); instructing the classifier to prioritize the current activity over persona priors, or windowing the persona summary to a recent horizon, would likely mitigate it, but we test neither.

Internal validity. Every score comes from a single run per model, with no error bars. Three properties bound the risk: temperature is 0.3 (Appendix D); the qualitative pattern holds across three independently pretrained families, which generation noise alone would not reproduce; and scores are MAE-aggregated over the 14-day window.

None substitutes for independent-seed replication, which the cost forecloses: each step issues one classifier call per active agent, and the summarizer fires only when unsummarized memory exceeds W, so a run costs $\approx O(|E|)$ calls—70,000+ per model including warm-up. On one A40 (48 GB) the aggregate track takes roughly 20 hours for Qwen3-8B and half that for the other two; per-topic tracks scale with their interaction counts, so General Election approaches the aggregate cost. The tracks also run independently, providing no mechanism for cross-topic conditioning within a single simulation.

External validity. The benchmark covers a single election in a single language; generalizability to other political contexts remains open. All three LLMs fall in the 8–12B range, and larger or proprietary models (e.g., GPT-4o, Claude) may narrow the gap with Ridge, but were excluded at the cost above. And we observe only on-platform activity: rallies, televised debates, and the official announcements inside the prediction window—the KPU tabulations, the 20 March certification—enter the data only as reaction, so shifts they drive are attributed to whatever interactions coincide with them.

Societal impact. The benchmark is dual-use: a simulator accurate enough to test counter-polarization interventions is accurate enough to test which messages polarize best. Current fidelity is far short of either (0.49 aggregate macro F1). Results obtained here describe active partisan accounts, not Indonesian public opinion.

Ethical Considerations

To adhere to the ACL Ethics Guidelines, we de-identified all account handles and applied semantic-preserving text perturbations to the raw tweets. This ensures that while the linguistic and social signals remain intact for simulation benchmarking, individual user privacy is protected against reverse-search Deanonymization. The anonymized dataset will be publicly available for research purposes upon publication.

Acknowledgments

Generative AI tools were used solely for writing assistance, generating low-novelty text, and helping identify relevant literatures, with all outputs carefully verified by the authors.

References

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.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, and 1 others. 2025. Llama-nemotron: Efficient reasoning models. *Preprint*, arXiv:2505.00949.

Elizabeth Blakey. 2024. The day data transparency died: How Twitter/X cut off access for social research. *Contexts*, 23(2):30–35.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, Mexico City, Mexico. Association for Computational Linguistics.

Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 89–96.

Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 03(01n04):87–98.

Morris H. DeGroot. 1974. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121.

Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. 2017. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4):2.

Noah E. Friedkin and Eugene C. Johnsen. 1990. Social influence and opinions. *The Journal of Mathematical Sociology*, 15(3-4):193–206.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2025. S^3: Social-network simulation system with large language model-empowered agents. *Preprint*, arXiv:2307.14984.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Chenhao Gu, Ling Luo, Zainab Razia Zaidi, and Shanika Karunasekera. 2025. Large language model driven agents for simulating echo chamber formation. *Preprint*, arXiv:2502.18138.

Rainer Hegselmann and Ulrich Krause. 2002. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3):1–33.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. 2019. The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22(Volume 22, 2019):129–146.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Nicolas Lanzetti, Florian Dörfler, and Nicolò Pagan. 2023. The impact of recommendation systems on opinion dynamics: Microscopic versus macroscopic effects. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 4824–4829.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Ryan Murtfeldt, Sejin Paik, Naomi Alterman, Ihsan Kahveci, and Jevin D. West. 2025. Rip twitter api: A eulogy to its vast research contributions. *Preprint*, arXiv:2404.07340.

Victor M. Panaretos and Yoav Zemel. 2019. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(Volume 6, 2019):405–431.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1–2):1–135.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2021. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, 4(1):381–402.

Lucky Susanto, Musa Izzanardi Wijanarko, Prasetia Anugrah Pratama, Zilu Tang, Fariz Akyas, Traci Hong, Ika Karlina Idris, Alham Fikri Aji, and Derry Tanti Wijaya. 2025. A multi-labeled dataset for Indonesian discourse: Examining toxicity, polarization, and demographics information. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18863–18890, Vienna, Austria. Association for Computational Linguistics.

Petter Törnberg, Claes Andersson, Kristian Lindgren, and Sven Banisch. 2021. Modeling the emergence of affective polarization in the social media society. *PLOS ONE*, 16(10):1–17.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. 2025. Decoding echo chambers: LLM-powered simulations revealing polarization in social networks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3913–3923, Abu Dhabi, UAE. Association for Computational Linguistics.

Wilson Wongso. 2023. indonesian-roberta-base-sentiment-classifier (revision e402e46).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, and 1 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Mohamed N. Zareer and Rastko R. Selmic. 2024. Modeling interactions in social media networks using an asynchronous and synchronous opinion dynamics. *Social Network Analysis and Mining*, 14(1):235.

Dingyi Zuo, Hongjie Zhang, Jie Ou, Chaosheng Feng, and Shuwan Liu. 2025. MTOS: A LLM-driven multi-topic opinion simulation framework for exploring echo chamber dynamics. *Preprint*, arXiv:2510.12423.

A Proposed LLM-Based ODM Algorithm

Algorithm 1 details the full simulation loop, corresponding to the components defined in Section 3.4.

B Construction of the Friedkin-Johnsen Matrices

This appendix gives the explicit construction of $\mathbf{W}$ and $\mathbf{s}$ used in Section 4.3.1.

Algorithm 1: LLM-Driven Agent-Based Simulation of Opinion Dynamics

Input: N agent personas $\{p_i\}_{i=1}^{N}$; warm-up histories $\{\mathcal{H}_i^{(0)}\}_{i=1}^{N}$; activity stream $\{\mathcal{E}^t\}_{t=1}^{T}$; opinion classifier f_{oc}; summarizer f_{sum}; learning rate α; summary window W.

Output: Opinion trajectories $\{x_i^t\}_{t=0}^{T}$ for each agent $i \in A$.

// Agent initialization

for $i = 1$ **to** N **do**

  $x_i^0 \leftarrow \frac{1}{|\mathcal{H}_i^{(0)}|} \sum_{(a,x,t)\in\mathcal{H}_i^{(0)}} x$

  $\mathcal{M}_i^0 \leftarrow \{\phi(a,c) : (a,c) \in \mathcal{H}_i^{(0)}\}$

  $\mathcal{S}_i^0 \leftarrow f_{\text{sum}}(\mathcal{M}_i^0)$

end

// Simulation

for $t = 1$ **to** T **do**

  foreach $(i, a_i^t, c_i^t) \in \mathcal{E}^t$ **do**

    // encode activity

    $m_i^t \leftarrow \phi(a_i^t,\, c_i^t)$

    $(r_i^t,\, u_i^t) \leftarrow f_{\text{oc}}(p_i,\, x_i^t,\, \mathcal{C}_i^t,\, m_i^t)$

    // classify shift

    $x_i^{t+1} \leftarrow \tanh(x_i^t + \alpha\, u_i^t)$

    // update opinion

    $\mathcal{M}_i^{t+1} \leftarrow \mathcal{M}_i^t \cup \{(m_i^t,\, r_i^t)\}$

    // append to memory

    if $L_i^{t+1} \bmod W = 0$ **then**

      $\ell \leftarrow L_i^{t+1}$

      $\mathcal{S}_i^{t+1} \leftarrow f_{\text{sum}}\big(\mathcal{M}_i^{t+1}[\ell - W + 1 : \ell]\big)$

    end

  end

  foreach $i \in A$ *not acting at* t **do**

    $x_i^{t+1} \leftarrow x_i^t$,

    $\mathcal{M}_i^{t+1} \leftarrow \mathcal{M}_i^t$,

    $\mathcal{S}_i^{t+1} \leftarrow \mathcal{S}_i^t$

  end

end

return $\{x_i^t\}_{t=0}^{T}$ for each $i \in A$

Interaction Matrix. $\mathbf{W}$ is constructed from the warm-up window $\mathcal{H}_i^{\text{init}} = \{k \in \mathcal{H}_i \mid t^{(k)} < \text{Feb } 1 + t_{\text{init}}\}$. For each ordered pair (i, j) with $i \neq j$, let $\mathcal{H}_{ij,s}^{\text{rep}} \subseteq \mathcal{H}_i^{\text{init}}$ denote the subset of replies from i to j with sentiment $s \in \{+, 0, -\}$, and let $\mathcal{H}_{ij}^{\text{rt}} \subseteq \mathcal{H}_i^{\text{init}}$ denote the retweets by i of j's posts. The unnormalized edge weight aggregates the two interaction channels:

$$\tilde{w}_{ij} = \left(\sum_{s\in\{+,0,-\}} \beta_s^{\text{rep}} \left|\mathcal{H}_{ij,s}^{\text{rep}}\right| \right) + \beta^{\text{rt}} \left|\mathcal{H}_{ij}^{\text{rt}}\right|$$

with the four scalars $\beta_+^{\text{rep}}, \beta_0^{\text{rep}}, \beta_-^{\text{rep}}$, and β^{rt} as hyperparameters. Observed self-interactions are discarded so the diagonal is not contaminated by reply or retweet counts; instead we set $\tilde{w}_{ii} = 1$ as a self-trust prior. $\mathbf{W}$ is then row-normalized,

$$w_{ij} \;=\; \tilde{w}_{ij} \Big/ \textstyle\sum_{k\in A} \tilde{w}_{ik},$$

so that $\sum_{j\in A} w_{ij} = 1$, satisfying the stochasticity requirement of the DeGroot or FJ update.

Susceptibility Vector s. Susceptibility is computed over the same warm-up window. Aggregating replies regardless of sentiment as $|\mathcal{H}_{ij}^{\text{rep}}| = \sum_{s\in\{+,0,-\}} |\mathcal{H}_{ij,s}^{\text{rep}}|$, we define the self-directed and total weighted activity of agent i as

$$\begin{aligned}
\sigma_i^{\text{self}} &= \beta_{\text{rep}}^{\text{self}} \left|\mathcal{H}_{ii}^{\text{rep}}\right| \;+\; \beta_{\text{rt}}^{\text{self}} \left|\mathcal{H}_{ii}^{\text{rt}}\right|, \\
\sigma_i^{\text{total}} &= \beta_{\text{rep}}^{\text{self}} \sum_{j\in A} \left|\mathcal{H}_{ij}^{\text{rep}}\right| \;+\; \beta_{\text{rt}}^{\text{self}} \sum_{j\in A} \left|\mathcal{H}_{ij}^{\text{rt}}\right|,
\end{aligned}$$

where the sums in σ_i^{total} run over all $j \in A$ including $j = i$, so $\sigma_i^{\text{self}} \leq \sigma_i^{\text{total}}$ by construction. The susceptibility is then

$$\rho_i \;=\; \frac{\sigma_i^{\text{self}}}{\sigma_i^{\text{total}}}, \qquad s_i \;=\; 1 - \text{clip}\big(\rho_i,\, 0.1,\, 0.9\big).$$

The clip prevents degenerate fully-rigid ($s_i \to 0$) or fully-malleable ($s_i \to 1$) agents. Agents with no recorded activity in the warm-up window are assigned a baseline $s_i = 0.6$, i.e. a stubbornness floor of 0.4.

C LLM Prompts

This appendix lists the two prompt templates used by the LLM-powered ABM (Section 4.3.3). Both are rendered verbatim; `{...}` placeholders are filled in at runtime with per-agent state. The memory summarizer is called once per agent during

warm-up to produce the persona p_i and seed summary S_i^0 from the warm-up history $\mathcal{H}_i$. The opinion classifier f_{oc} is called at every simulation step for every active agent and returns the (r_i^t, u_i^t) pair defined in Section 3.4.

C.1 Memory Summarization Prompt

```
You are an analyst who summarizes a person's
    ↪ recent activities in order to infer their
    ↪  traits.
You are given a list of the agent's few most
    ↪ recent memories.
Each memory contains a timestamp and a short
    ↪ description of an activity.

Your task:
- Read all the memories carefully.
- Identify patterns in personality, political
    ↪ views, and behavioural tendencies.
- Then produce a single coherent paragraph
    ↪ describing how this agent's personality,
    ↪ political leaning, and behaviour are
    ↪ constructed based on these activities.

Do NOT restate each memory individually.
Instead, infer trends, tendencies, and
    ↪ motivations behind the activities.

Input memories:
{memories}

Output format (one paragraph only):
A summary written from a first-person
    ↪ perspective ("I ..."), expressing:
- core personality traits
- political leaning (if detectable)
- behavioural patterns
- what these traits seem to be based on in the
    ↪ memories
- Strictly output the summary paragraph only.
```

C.2 Opinion Classifier Prompt

```
You are simulating a Twitter user's opinion
    ↪ about the following topic: {topic}.

## User Profile
- Gender: {gender}
- Followers: {followers_count} | Following: {
    ↪ following_count}
- Opinion before this activity: {
    ↪ current_opinion_weight} (continuous, -1.0
    ↪  to 1.0)

## Personality (long-term)
{summarized_memory}

## Recent Memory (short-term, chronological)
{memories}

## Activity
- Type: {activity_type} (original_post | retweet
    ↪  | reply)
- Content: "{activity_content}"

---

## Your Task
Based on the user's personality, memory, social
    ↪ context, and the nature of this activity,
determine:

1. **expressed_opinion** -- What sentiment does
    ↪ this activity express toward the topic?
   Scale: -2 (strongly against), -1 (against), 0
    ↪  (neutral), 1 (for), 2 (strongly for)

2. **opinion_delta** -- How much does this
    ↪ activity shift the user's internal
    ↪ opinion?
   Scale: -1.0 to 1.0

   Consider:
   - A retweet signals weaker conviction than an
    ↪  original post
   - A reply's shift depends on whether it's
    ↪ defensive or exploratory
   - A user whose memory shows consistent views
    ↪ shifts less than one with conflicted
    ↪ history
   - Prior opinion distance from the activity's
    ↪ sentiment amplifies or dampens the delta

Output ONLY valid JSON, no markdown, no extra
    ↪ fields:

{
  "reasoning": "<one sentence, first-person,
    ↪ less than 80 words>",
  "expressed_opinion": <integer -2 to 2>,
  "opinion_delta": <float -1.0 to 1.0>
}
```

D Hyperparameter Values

This appendix lists the hyperparameter values used to produce the results in Section 5. Tables 3–5 report one set per model. Values prefixed with β in Table 3 are jointly grid-searched on the training fold; the remaining values are fixed (defaults or configuration constants).

Symbol	Description	Value
β_{+}^{rep}	Edge weight, positive reply	0.1
β_{0}^{rep}	Edge weight, neutral reply	0.0
β_{-}^{rep}	Edge weight, negative reply	0.0
β^{rt}	Edge weight, retweet	0.1
$\beta_{\text{rep}}^{\text{self}}$	Self-share weight, reply	$0.1^{\dagger}$
$\beta_{\text{rt}}^{\text{self}}$	Self-share weight, retweet	0.1
$\tilde{w}_{ii}$	Diagonal self-trust prior	1.0
$[\rho^{\min}, \rho^{\max}]$	Clip range for ρ_i	$[0.1, 0.9]$
s_i^{base}	Susceptibility baseline (no activity)	0.6
t_{init}	Warm-up window (days)	14
Δt	Simulation step	2.4 h

Table 3: Hyperparameters for the Friedkin-Johnsen baseline. Values are shared across all experiments except $^{\dagger}\beta_{\text{rep}}^{\text{self}} = 0.9$ on the aggregate track.

Symbol	Description	Value
λ	Ridge regularization strength	1.0
—	Standardization	`StandardScaler`
t_{init}	Warm-up window (days)	14
T_{train}	Training buckets (Feb 1–14)	14
—	Test horizon (Feb 15–28)	14 days
Δt_{model}	Training/inference grid	1 day
Δt_{out}	Output grid (forward-held)	2.4 h

Table 4: Hyperparameters for the Ridge regression baseline.

Symbol	Description	Value
—	Decoding temperature	0.3
p	Nucleus sampling threshold (top-p)	0.8
k	Top-k cutoff	20
α	Opinion-update step (tanh)	0.5
$\lvert W \rvert$	Memory window (recent entries)	5
—	Classifier batch size	50
—	Warm-up summarizer batch	50
T_{init}	Warm-up window (days)	14
Δt	Simulation step	15 min

Table 5: Hyperparameters for the LLM-powered ABM. We evaluate three open-weight models under identical settings.

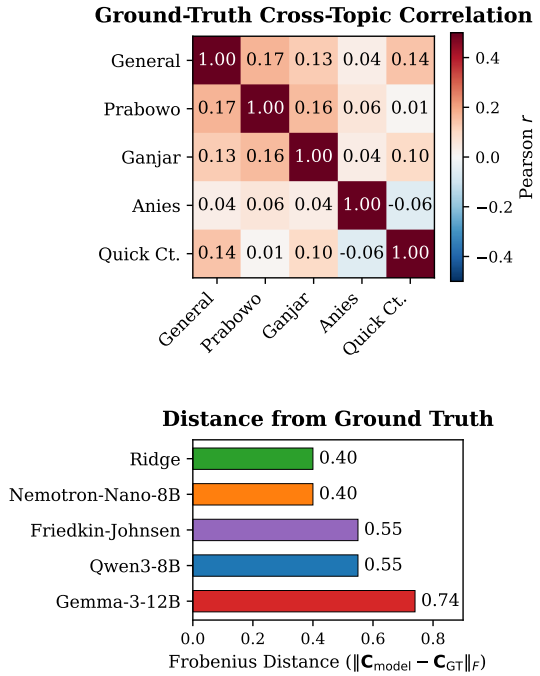


Figure 4: Top: ground-truth cross-topic Pearson correlation matrix over 914 agents. Bottom: Frobenius distance between each model's predicted correlation matrix and the ground truth (↓).

E Cross-Topic Correlation Analysis

To assess whether models recover the multi-topic opinion structure of real users, we compute per-agent mean opinion across the prediction window for each topic, build a 5×5 Pearson correlation matrix over agents active in at least two topics, and compare each model's matrix against the ground truth via Frobenius norm $\|\mathbf{C}_{\text{model}} - \mathbf{C}_{\text{GT}}\|_F$. Figure 4 reports the ground-truth correlation structure (left) and each model's Frobenius distance (right). Ridge and Nemotron-Nano-8B recover cross-topic structure most faithfully (0.40), while Gemma-3-12B diverges most (0.74).

F Topic Model Terms

Table 6 reports the ten most salient terms for each of the five LDA topics, supporting the face-validity argument of Section 4.1.

Terms are ranked by log-odds ratio against the rest of the corpus under an informative Dirichlet prior (Monroe et al., 2008) rather than by raw frequency, which returns the same generic election vocabulary (*pemilu*, *presiden*, *pilpres*) for every topic. Counting runs over stems so that morphological variants collapse together, but the term displayed is the most frequent surface form within each stem group: Indonesian stemming strips the political sense of a word, reducing *perubahan* ("change", the campaign slogan of the Anies–Muhaimin ticket) to *ubah*, and *pemilu* itself to the non-word *milu*. The unit of analysis is the deduplicated source text rather than the 70,336 interactions, so a single viral post cannot dominate a topic through its retweets.

The terms recover campaign specifics that the topic names only gesture at. Topic 2 carries the *kampanye akbar* at Jakarta International Stadium (*jis*) alongside *perubahan* and the nicknames *abah* and *cak imin*; Topic 3 carries Prabowo's free-meals programme (*makan*, *gratis*) and his running mate *gibran*; Topic 4 carries the *Hajatan Rakyat* rallies, the *salam metal* hand sign, and the Gelora Bung Karno stadium (*gbk*). Topic 1 is procedural rather than candidate-centred, dominated by *suara*, *kecurangan*, and *kpu*.

Quick Count is the weakest of the five. Its 203 documents yield *count*, *quick*, *persen*, and *real*, the vocabulary of tallying returns, but also *iklan*, *download*, and *klik*, which come from promotional posts riding on quick-count traffic. The topic is coherent enough to name, but it is the smallest and the noisiest track in the benchmark.

	General Election (N=28,151)	Anies Baswedan (N=406)	Prabowo Subianto (N=5,379)	Ganjar Pranowo (N=608)	Quick Count (N=203)
1	pemilu (election)	akbar (grand)	prabowo	ganjar	count
2	suara (vote)	jis (stadium)	jokowi	mahfud	quick
3	kecurangan (fraud)	anies	ahok	pranowo	persen (percent)
4	kpu (electoral comm.)	amin	pendukung (supporter)	warga (residents)	baca (read)
5	pilpres (pres. election)	kampanye (campaign)	gratis (free)	hajatan (public feast)	percaya (believe)
6	partai (party)	baswedan	makan (meal)	hadir (attend)	real
7	pilih (vote)	abah (nickname)	kasih (give)	total (total)	iklan (advert.)
8	presiden (president)	imin	gibran	ribu (thousand)	download
9	hasil (result)	perubahan (change)	orang (people)	salam (salute)	klik (click)
10	indonesia	cak (title)	serangan (attack)	disambut (welcomed)	survey

Table 6: Ten most salient terms per LDA topic, ranked by log-odds ratio against the rest of the corpus under an informative Dirichlet prior (Monroe et al., 2008). English glosses in parentheses; proper nouns and terms already in English are left as they appear. N is the number of deduplicated source texts assigned to each topic.

G Annotation Protocol and Disagreement Analysis

This appendix details the human validation study summarized in Section 4.1.

Protocol. A native Indonesian speaker annotated a stratified sample of 250 interactions drawn across the five topics: 133 labelled negative by the classifier, 66 neutral, and 51 positive. The annotator worked from the interaction text alone, without access to the classifier output, and assigned each item a label on the same $\{-1, 0, +1\}$ scale used for the ground-truth opinion weights. The annotator is an author, so the study measures agreement between a human reading and the classifier rather than inter-annotator reliability.

Recurring disagreement patterns. Table 7 lists the three patterns that recur most often among the disagreements. All three are cases of the classifier defaulting to neutral on text a human reads as directional. Two concern implicit negativity: mockery carried by sarcasm, and accusations of vote-rigging or corruption phrased as statements of fact. The third runs the opposite way, with enthusiastic candidate endorsements falling into neutral rather than positive.

Sarcasm was the hardest of the three to adjudicate, since the surface form is often approving and only the surrounding context inverts it. Agreement over the full sample is 70.8%.

Implication for the benchmark. The two negative-direction patterns outnumber the positive-direction one, so the net effect of the classifier's neutral default is to under-state negativity in the corpus. This grounds the *conservative direction* argument of Section 4.1: correcting these labels would push the ground-truth signal further negative, widening the gap against LLMs that already skew positive rather than closing it.

Pattern	Classifier → Human	Cases
Sarcasm / mockery	neutral → negative	11
Corruption or fraud claims	neutral → negative	16
Enthusiastic endorsement	neutral → positive	16

Table 7: The three most frequent disagreement patterns between the RoBERTa sentiment labels and the human annotation on the 250-item validation sample.

H Ground-Truth Signal Over Time

Figure 5 reports the daily evolution of the ground-truth signal introduced in Section 4.2, separated by topic. Daily volume spans three orders of magnitude across the five topics, and Ganjar Pranowo and Quick Count fall to zero recorded activity on several days, which is what motivates the per-topic evaluation track. Both interaction volume and the number of active users peak on 14 February and fall away over the prediction window, so models are scored on a progressively sparser signal. Mean opinion shifts downward on every topic after the election, crossing into net-negative territory on all but Ganjar Pranowo. Quick Count is the only topic to gain volume after the vote, and it turns sharply negative as it does, consistent with post-election discourse concentrating on disputes over the count itself.

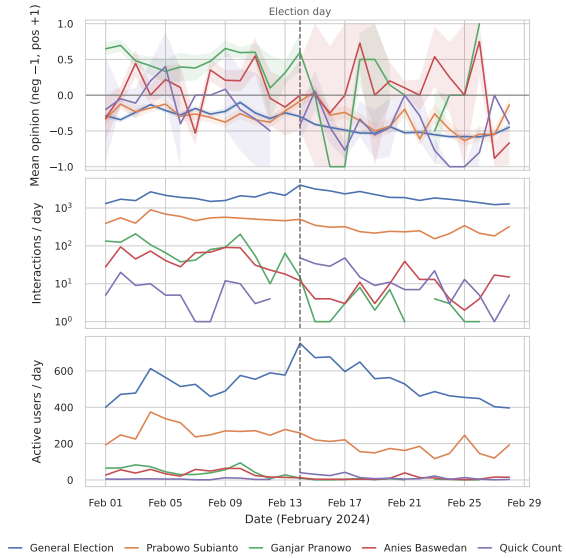


Figure 5: Daily evolution of the ground-truth signal per topic: (top) mean opinion with 95% CI, (middle) interaction volume (log scale), (bottom) unique active users. The dashed line marks the 14 February general election. Days with no recorded activity on a topic appear as gaps rather than zero-valued points.