

How Do Self-Referential LLM Loops Contract Political Diversity?

Wooyoung Kim

Department of Industrial Engineering
Yonsei University
timothy@yonsei.ac.kr

Wooju Kim

Department of Industrial Engineering
Yonsei University
wkim@yonsei.ac.kr

Abstract

As Large Language Models (LLMs) increasingly consume their own outputs through retrieval-augmented generation and synthetic data pipelines, self-referential loops are becoming prevalent in modern information ecosystems. While prior work has studied variance collapse during recursive training, the dynamic consequences for political diversity during inference remain underexplored. We investigate this by conducting iterative paraphrasing experiments across eight widely used LLMs, and formalize self-referential generation as a discrete-time Markov process to characterize its long-term behavior. Our results reveal a systematic contraction of political diversity, with several models converging toward center-dominant equilibria. Notably, this homogenization is strongly front-loaded: the majority of diversity loss occurs within the earliest iterative rounds, indicating that even a few AI-mediated transformations can disproportionately narrow the political spectrum. We further examine two interventions: Stance-Conditioned Generation (SCG) substantially reduces marginal stance-distribution drift even without ground-truth labels, leaving no statistically detectable shift for three of the four homogenizing models, whereas Self-Identified Stance Preservation (SISP) overshoots into over-polarization; even SCG improves but does not guarantee instance-level stance retention. These findings highlight a structural risk associated with self-referential generation in AI-driven information pipelines. Our code and prompts are publicly available at <https://github.com/timothy-coshin/SelfReferentialPoliticalDiversity>.

1 Introduction

The widespread proliferation of Large Language Models (LLMs) is fundamentally reshaping information ecosystems. Current estimates from [Spennemann \(2025\)](#) suggest that approximately 30% of active web content is now AI-generated. How-

ever, this transition brings a critical challenge: the erosion of information diversity.

[Shumailov et al. \(2024\)](#) established that models trained recursively on synthetic data suffer from the disappearance of tail distributions, leading to a profound loss of data diversity. Although this phenomenon was initially studied in the context of recursive training, it raises broader concerns about how similar feedback dynamics may operate during inference-time generation. [Wang et al. \(2025\)](#) show, in a simulated multi-model ecosystem whose shared retrieval memory accumulates model-generated text, that such self-referential loops are not limited to retraining phases but also arise through recursive consumption within prompts, as in Retrieval-Augmented Generation (RAG) ([Lewis et al., 2020](#)) pipelines, causing outputs to become increasingly homogenized.

This phenomenon is particularly detrimental to political discourse, where the preservation of diverse perspectives is widely regarded as essential—and political discourse reaches the public primarily through online news, a channel that is rapidly becoming synthetic. One industry analysis estimates that primarily AI-generated articles now account for roughly half of newly published online articles ([Paredes et al., 2026](#)), and in a survey of journalists across the Global South and emerging economies, over 80% reported using AI in their work ([Radcliffe, 2025](#)). In light of these trends, this study provides a systematic evaluation of the extent to which political stance distributions are eroded within LLM self-referential cycles.

To empirically investigate how political stance evolves under self-referential generation, we conduct an iterative paraphrasing experiment that tracks stance transitions across successive generations. The results reveal a strongly front-loaded pattern of political homogenization, with most of the contraction in stance diversity occurring within the earliest generation rounds.

To formally characterize this dynamic, we introduce a Markovian framework that models the self-referential information loop as a discrete-time stochastic process. This formulation enables the analysis of long-term stationary behavior across different model ecosystems and provides a principled analytical account of the observed reduction in political entropy. Because the loop procedure is held fixed across models, this formulation separates what each model contributes at a single step from what repetition of the loop does to it. In addition, we examine two complementary strategies for mitigating homogenization: Stance-Conditioned Generation (SCG), which enforces explicit external control through stance conditioning, and Self-Identified Stance Preservation (SISP), which relies on internal consistency by allowing the model to infer and maintain the stance of the input text.

This study has important societal implications for AI-mediated political communication. Unchecked self-referential generation may quickly narrow the spectrum of political viewpoints exposed to the public, posing a substantial risk to pluralism in information ecosystems.

2 Related Work

2.1 Static Political Bias in LLMs

Bang et al. (2024) and Rettenberger et al. (2025) have investigated political bias in LLMs, focusing primarily on snapshot evaluations of model outputs. These studies analyze political leanings through curated prompts or survey-style queries, documenting how models align with specific political spectra.

However, these evaluations are inherently static, failing to account for the dynamic evolution of political framing when LLMs are deployed in recursive generation and consumption cycles. Our work shifts the focus from *what a model believes* to *how political diversity collapses* over successive iterations. By characterizing stance drift as a systematic outcome of recursive generation–consumption loops, we highlight the urgent need to address the structural risks posed by the emerging LLM Web Dynamics (Wang et al., 2025).

2.2 Model Collapse and Variance Loss

Model Collapse (Shumailov et al., 2024) refers to the phenomenon where recursive training on synthetic data leads to the loss of tail distributions and statistical variance, resulting in a profound decline in data diversity. We extend this discourse by

demonstrating that such technical variance loss is reflected in the erosion of political diversity, where nuanced partisan stances are progressively stripped away. Unlike prior studies centered on the retraining phase, our work focuses on the drift occurring specifically within recursive inference loops, where each model-generated output serves as the conditional input for subsequent transformations.

3 Methodology

We establish a controlled framework for evaluating the effects of generative self-consumption through recursive language transformation. The resulting paraphrasing loop is a deliberately simplified proxy for the repeated model-mediated transformations that arise in RAG and synthetic-data pipelines; we do not implement retrieval, ranking, or mixing with human-authored content in the main loop. Within this self-referential generative regime, we assess the preservation of political diversity as a key evaluation criterion for a pluralistic information ecosystem. Specifically, we examine whether the nuanced political stances present in human-authored content remain stable or progressively converge toward a simplified, neutral state under repeated LLM-mediated generation. Additionally, by formalizing these iterative cycles as a Markovian process, we characterize the long-term stationary behavior of political diversity and analyze the systematic reduction of entropy in political discourse.

3.1 Data Collection

We adopt a focused seed-set construction strategy designed to facilitate controlled analysis of stance drift by emphasizing political framing and anchoring all instances to the same underlying topics and events. To systematically study political diversity, we construct a dataset centered on policy issues that exhibit pronounced partisan polarization in the United States. Following Pew Research Center (2024), we identify seven core topics whose perceived importance differs substantially across political affiliations: Climate Change, Economy and Jobs, Environment, Healthcare, Immigration, Race and Racism, and Terrorism. This design prioritizes stance clarity over topical breadth, and the resulting dataset is intended to function as a *diagnostic benchmark* for characterizing self-referential drift dynamics under minimal initial noise, rather than as a representative corpus for population-level stance estimation.

For each topic, we curate news articles associated with concrete real-world events that elicited diverse political responses. We rely on AllSides¹, a media-bias rating platform also used in prior research (Haak and Schaer, 2023), to select coverage from outlets classified as Left (L), Center (C), and Right (R).

In total, the dataset comprises 39 seed articles spanning 13 events across the seven topics, each carrying a stance label derived from the AllSides media-bias rating of its source outlet; further details regarding the data collection and expert curation process are provided in Appendix A. From these seeds, we analyze a total of 6,160 model-paraphrased instances drawn from complete loops across eight models and 20 recursive rounds (out of a nominal 6,240; four articles failed to complete the loop for one model, see Appendix B). This design enables controlled comparisons across political stances under a fixed factual context.

We formally represent each data instance as $d_i = (t_i, e_i, c_i, y_i)$ consisting of the following components:

- t_i : The high-level topic category, where $t_i \in \{\text{Climate Change, Economy and Jobs, } \dots, \text{Terrorism}\}$.
- e_i : The event headline and a factual summary describing the associated political context.
- c_i : The initial title and textual content of a specific news article, expressing a particular political stance toward the event.
- y_i : The reference political stance label inherited from the outlet’s AllSides rating, where $y_i \in \{L, C, R\}$.

3.2 Evaluating Political Stance Stability

To analyze the temporal stability of political framing, we introduce a recursive paraphrasing and evaluation framework. This process consists of two stages: (1) iterative text generation to simulate a self-referential loop and (2) stance estimation.

Recursive Generation. Given the initial article content c_i of d_i , we utilize a generative language model $\mathcal{G}$ to perform R successive rounds of paraphrasing. For each round $r \in \{1, \dots, R\}$, the paraphrased title and content are generated based on the output of the previous round:

$$c_i^r = \mathcal{G}(c_i^{r-1}) \quad \text{where } c_i^0 = c_i \quad (1)$$

¹<https://www.allsides.com>

This iterative approach allows us to observe how political nuances shift or dissipate as the text undergoes repeated transformations. The paraphrasing prompt explicitly instructs the model to preserve facts and political nuance (Appendix E), so the contraction reported in Section 4.1 occurs despite an explicit preservation instruction.

Stance Scoring and Classification. To quantify the political stance across successive paraphrasing rounds, we adopt the LLM-as-a-judge paradigm (Zheng et al., 2023). For each paraphrased instance at round r , we employ a dedicated judge model $\mathcal{J}$ to estimate its political orientation. The model $\mathcal{J}$ takes the generated text c_i^r and the factual event summary e_i as inputs to produce a continuous stance score $v_i^r \in [-1, 1]$:

$$v_i^r = \mathcal{J}(e_i, c_i^r) \quad (2)$$

where $v_i^r = -1$ and $v_i^r = 1$ represent the extremes of the Left and Right stances. To promote consistency in assessment, the judge model is explicitly provided with e_i as a formal reference, which defines the characteristic arguments of both the Left and Right sides for each event to ground the stance classification. To derive discrete stance labels, we map the continuous score v_i^r to a discrete stance label $\hat{y}_i^r$ using a predefined threshold τ as follows:

$$\hat{y}_i^r = \begin{cases} \text{Left} & \text{if } v_i^r < -\tau \\ \text{Center} & \text{if } -\tau \leq v_i^r \leq \tau \\ \text{Right} & \text{if } v_i^r > \tau \end{cases} \quad (3)$$

3.3 Implementation Details

Generator Selection. To examine whether our observations generalize across commonly used LLMs, we evaluate eight generators $\mathcal{G}$: four proprietary models (GPT5.1, GPT4o, Gemini2.5, Sonnet4.5) (OpenAI, 2026, 2025, 2024; Gemini Team, 2025; Anthropic, 2025) and four open-weight models (Exaone3.5, LLaMa3.1, Gemma3, Qwen3) (LG AI Research et al., 2024; Llama Team, 2024; Gemma Team, 2025; Qwen Team, 2025). Exact model versions are listed in Appendix B.

Judge Model Selection. The accuracy of the judge model $\mathcal{J}$ is a critical prerequisite for reliable stance evaluation. We benchmark four candidate LLMs by assessing zero-shot stance prediction performance against the outlet-derived reference labels y_i , and select GPT4o (OpenAI, 2024) as

	GPT5.1	Gemini2.5	Qwen3	Exaone3.5	Gemma3	LLaMa3.1	GPT4o	Sonnet4.5
Initial Dist.	0.38 / 0.33 / 0.28 (0.43/0.34/0.23 for LLaMa3.1)							
Final Dist.	0.41 / 0.36 / 0.23	0.41 / 0.36 / 0.23	0.36 / 0.38 / 0.26	0.33 / 0.46 / 0.21	0.13 / 0.72 / 0.15	0.14 / 0.74 / 0.11	0.21 / 0.62 / 0.18	0.28 / 0.67 / 0.05
<i>p</i> -value	0.7788	0.7788	0.8382	0.4657	0.0071	0.0068	0.0169	0.0109
Initial Variance(v^0)	0.3626 (0.3269 for LLaMa3.1)							
Final Variance(v^R)	0.3373	0.3254	0.3230	0.2394	0.1219	0.1387	0.1995	0.1493
VRR	6.98%	10.26%	10.92%	33.98%	66.38%	57.57%	44.99%	58.83%

Table 1: Stance distribution dynamics and score variance stabilization across models over 20 iterative rounds. All models except LLaMa3.1 share the same initial stance distribution and initial score variance, while the final distributions (slashes denote Left/Center/Right stances), Variance Reduction Ratios (VRR), and Stuart–Maxwell test results reveal model-specific differences in stance drift and stabilization behavior. For LLaMa3.1, four articles whose generation failed mid-loop are excluded from all rounds ($n = 35$; see Appendix B), so all of its entries, including the initial distribution (0.43/0.34/0.23) and variance, are computed on those 35 articles.

the primary judge. Full benchmarking results and a robustness analysis addressing potential judge-specific biases are provided in Appendix C.

Hyperparameters and LLM Prompts. We use a temperature of 0.8 for text generation to balance diversity and consistency, and a temperature of 0.0 for evaluation to minimize sampling variability. The recursive depth is fixed to $R = 20$, and the stance classification threshold is set to $\tau = 0.3$. The initial score variance $\text{Var}(v^0) = 0.3626$ is taken from a single round-0 judging pass over the original 39 articles and shared as the common baseline across models, with the exception noted for LLaMa3.1 in Table 1. Additional details and ablation analyses for these parameter settings are provided in Appendix D. Also, to ensure experimental reproducibility, all LLM prompts used for generation and stance evaluation are reported in Appendix E.

4 Experiments

4.1 Political Stance Drift Across Different LLMs

In this section, we analyze how political stances change during recursive paraphrasing using two statistical measures. First, we conducted a Stuart–Maxwell test for marginal homogeneity to compare the initial distribution of political stance labels with the final-round distribution. This test determines whether the distribution of political perspectives changed between the initial and final rounds. Second, we tracked the variance of the continuous political scores (v_i^r). To quantify the extent to which the political spectrum shrinks, we define the Variance Reduction Ratio (VRR) as follows:

$$\text{VRR} = \left(1 - \frac{\text{Var}(v^R)}{\text{Var}(v^0)}\right) \times 100(\%) \quad (4)$$

where $\text{Var}(v^0)$ and $\text{Var}(v^R)$ represent the variance of the political scores at the initial and final rounds, respectively. A decrease in variance indicates that the diversity of views is shrinking toward a single, homogenized mean.

As shown in Table 1, the models divide into two distinct groups. The first group, which we call Preserving Models, includes GPT5.1, Gemini2.5, Qwen3, and Exaone3.5. For these models we detect no statistically significant marginal shift after 20 rounds of generation ($p > 0.05$, Stuart–Maxwell). Exaone3.5 sits between the two groups: its marginal distribution is statistically unchanged, yet it loses 33.98% of its score variance, showing that categorical stability and intensity preservation can dissociate.

In contrast, the Homogenizing Models—GPT4o, Sonnet4.5, Gemma3, and LLaMa3.1—showed a statistically significant shift in their political distributions ($p < 0.05$). These models experienced a strong central drift, where Left and Right stances rapidly diminished and transitioned into the Center category. By the final round, the Center proportion rose sharply to 62–74%, a marked departure from the initial balanced state. Gemma3 exhibited the largest variance reduction (66.38%), followed by Sonnet4.5 (58.83%) and LLaMa3.1 (57.57%). This variance reduction indicates that the dispersion of political intensity scores in the original data was compressed into a narrower range; the round-by-round analyses below locate that range at the neutral center.

The temporal dynamics of political homogenization are further elucidated by the round-by-round progression illustrated in Figure 1 and Figure 2. A key finding is that Homogenizing Models tend to lose political diversity most rapidly during the early stages of iterative paraphrasing.

As shown in Figure 1, Homogenizing Models exhibit an abrupt increase in the proportion of

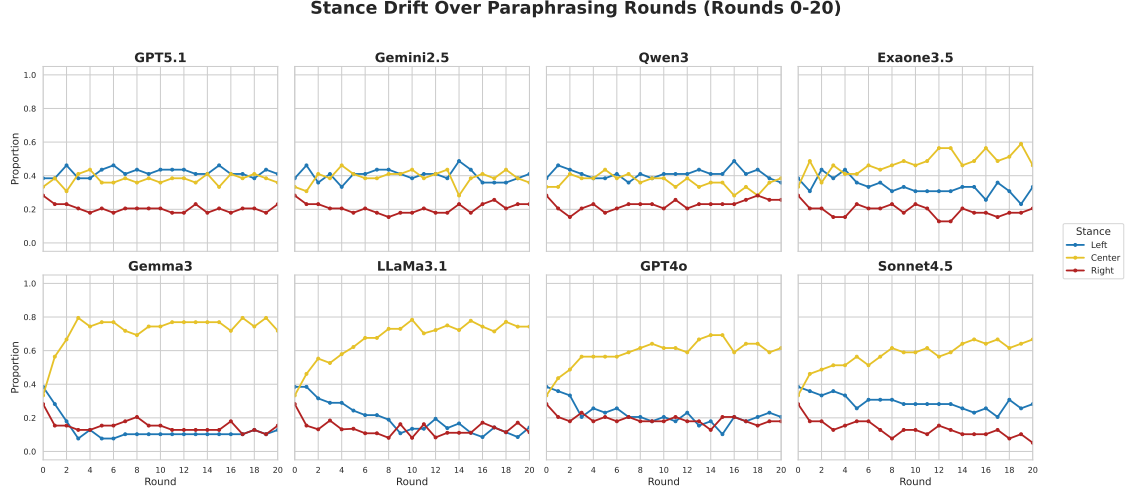


Figure 1: Evolution of political stance distribution over 20 recursive paraphrasing rounds. Homogenizing Models (bottom row) exhibit a rapid surge in the Center proportion (yellow line) during the early rounds, whereas Preserving Models (top row) largely retain their initial balance, with Exaone3.5 drifting partly toward the Center.

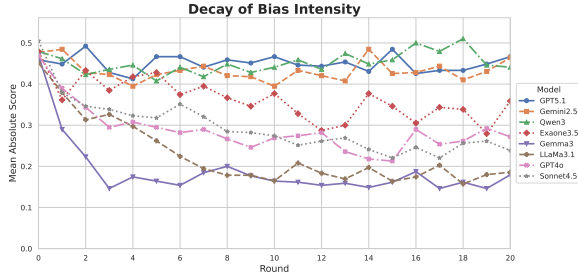


Figure 2: Decay of bias intensity across recursive rounds. The Mean Absolute Score serves as a proxy for the strength of political framing. Homogenizing Models show an immediate and sharp decline in intensity within the first few rounds, indicating a rapid reduction in partisan nuance, while Preserving Models largely retain it.

Center stances, often within the first few rounds. Among these models, more than half of the total increase in the Center proportion occurs within the initial 1–3 rounds. This observation suggests that the diversity-reducing effects of recursive generation are not merely cumulative but are disproportionately concentrated in the earliest transformation steps.

This pattern is further supported by Figure 2. To quantify the intensity of political stance, we compute the Mean Absolute Score (MAS) at each round as follows:

$$\text{MAS} = \frac{1}{n} \sum_{i=1}^n |v_i^r| \quad (5)$$

where n denotes the total number of articles. For Homogenizing Models, the MAS decreases sharply during the initial iterations. In particular, Gemma3 and LLaMa3.1 exhibit the most pronounced decline, with stance intensity rapidly collapsing before stabilizing at a persistently low level for the remainder of the 20-round cycle. In contrast, Preserving Models such as GPT5.1 and Gemini2.5 maintain a consistently high and stable MAS across all rounds, indicating a sustained ability to preserve the magnitude of the original political stance.

Two complementary analyses further characterize this contraction. First, it operates through discrete absorption into the Center state rather than gradual within-label moderation: instances that retain a Left or Right label maintain a comparatively stable stance intensity across all rounds (Appendix F). Second, we detect no association between initial stance intensity and absorption: even the most polarized articles are absorbed at comparable rates (Appendix G).

Taken together, these results suggest that for the Homogenizing Models, the most severe nuance-stripping effect is front-loaded, occurring during the earliest stages of information transformation. The transition matrices in Figure 3 explain why this contraction is also largely irreversible. While the probabilities of transitioning from non-central stances to the Center are high, the reverse transition probabilities from the Center to non-central stances

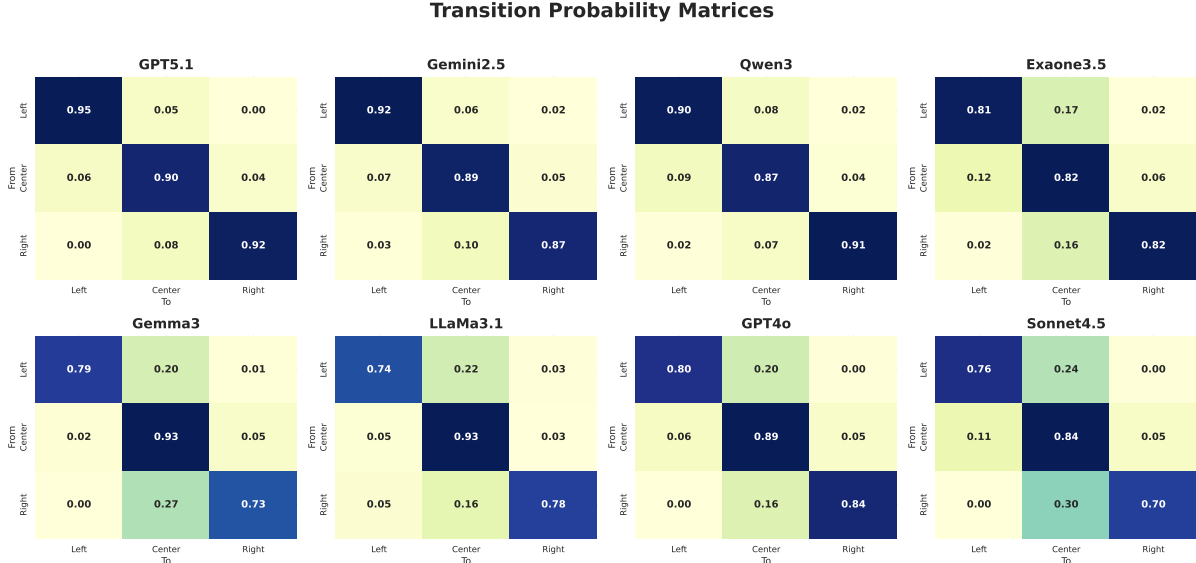


Figure 3: Transition probability matrices representing the likelihood of shifting between political stances (Left, Center, Right) in a single paraphrasing round. Preserving Models (top row) exhibit high diagonal values, indicating a strong tendency to preserve the original stance. In contrast, Homogenizing Models (bottom row) show substantial transition probabilities from polarized stances toward the Center state.

remain markedly low.

As a result, once generated content is absorbed into the Center state, the system exhibits limited capacity to recover political diversity. The speed of the initial contraction, in turn, follows from the Markov analysis in Section 4.2: a fixed transition kernel initialized far from its equilibrium moves fastest in the earliest rounds and slows as it approaches a neutral equilibrium.

The predominance of the Center state is broadly consistent with evidence that preference-alignment methods such as RLHF and constitutional training (Ouyang et al., 2022; Bai et al., 2022) can reduce semantic and viewpoint diversity in model outputs (Slocum et al., 2025); we refer to the resulting tendency to favor moderate framings at the expense of political variance as a *neutrality trap*. What explains the split between the two model groups? Several candidate factors do not, on their own, separate them: model scale does not (the 14B Qwen3 shows only 10.92% VRR while the similarly sized 12B Gemma3 homogenizes most of all), decoding does not (all models share the same temperature and prompts), and instruction tuning by itself does not (all eight generators are instruction-tuned, yet they split into two groups). Even within a single provider, GPT4o and GPT5.1 behave oppositely, suggesting recipe-level rather than family-level determinants.

To probe the role of alignment training directly, we compare base and alignment-trained checkpoints within three open-weight model families under matched few-shot prompting, so that only the weights differ within each pair (Appendix H). Alignment training amplifies self-referential homogenization in two of the three families (Llama-3.1-8B: VRR 42.9% $\rightarrow$ 55.0%; Gemma-3-12B: 22.0% $\rightarrow$ 59.3%), leaves one essentially unchanged (Qwen2.5-7B), and decreases it in no family.

Notably, the base checkpoints already exhibit substantial contraction (VRR 22–43%), indicating that the recursive loop itself induces homogenization and that alignment acts on top of it. Alignment training thus acts as a model-dependent amplifier of homogenization rather than a universal cause, with a magnitude that varies by family (see Limitations for attribution caveats). In summary, the dynamics are model-dependent: where contraction occurs it is concentrated in the earliest iterations, and the low probability of recovering non-central stances once Center absorption occurs is what makes it persistent.

4.2 Long-term Projection: Markov Chain Analysis

To estimate the long-term equilibrium of the information ecosystem beyond the observed rounds,

we model recursive paraphrasing as a discrete-time Markov process. This allows us to estimate the stationary distribution. We define the political stance as a state in a finite space $S = \{L, C, R\}$. For each model $\mathcal{G}$, we construct a transition probability matrix $P \in \mathbb{R}^{3 \times 3}$, where $p_{jk} = P(s_r = k \mid s_{r-1} = j)$ represents the probability of transitioning from stance j to k in a single round. The long-term behavior is determined by the stationary distribution π , satisfying:

$$\pi P = \pi, \quad \text{where } \sum_{i \in S} \pi_i = 1 \quad (6)$$

This π represents the long-term equilibrium of the stance distribution under the estimated transition dynamics. To quantify the diversity of this equilibrium, we calculate the stationary Political Entropy $H(\pi)$:

$$H(\pi) = - \sum_{i \in S} \pi_i \log_e \pi_i \quad (7)$$

This formulation assumes time-homogeneity, i.e., that a single kernel P estimated by pooling transitions across rounds adequately describes the dynamics. A fixed kernel initialized far from its equilibrium produces exactly the geometric, early-concentrated convergence reported in Section 4.1. We validate the assumption empirically in Appendix I through early- versus late-round kernel comparisons, full-trajectory reproduction, and a 10-round held-out extension of the GPT4o loop.²

The estimated transition matrices (Figure 3) reveal the transition dynamics underlying the observed drift. Preserving Models exhibit strong diagonal dominance, with polarized-state self-transition probabilities p_{jj} ranging from 0.81 to 0.95 (0.945 and 0.923 for GPT5.1). In contrast, Homogenizing Models show substantially higher transition probabilities toward the Center state; for instance, Gemma3 and Sonnet4.5 exhibit a 0.20 to 0.30 probability of polarized articles shifting to Center in a single iteration.

These transition dynamics lead to drastically different final equilibria (Figure 4). The stationary distributions of Homogenizing Models are heavily skewed, with the Center state probability reaching 0.79 for Gemma3 and 0.73 for LLaMa3.1. Conversely, Preserving Models converge to a balanced stationary state that closely mirrors the initial

²For GPT4o, the kernel is estimated from this extended 30-round run; the estimate from rounds 1–20 alone differs by at most 0.017 per entry and yields indistinguishable stationary quantities (Appendix I).

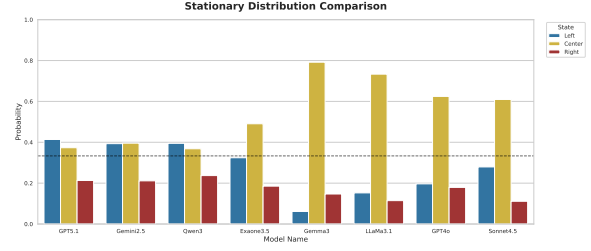


Figure 4: Comparison of predicted stationary political distributions across evaluated LLMs. The dashed line marks the uniform reference share of 1/3. Homogenizing Models (right side) converge toward a distribution heavily dominated by the Center stance, while Preserving Models (left side) sustain a pluralistic distribution that closely mirrors the initial state.

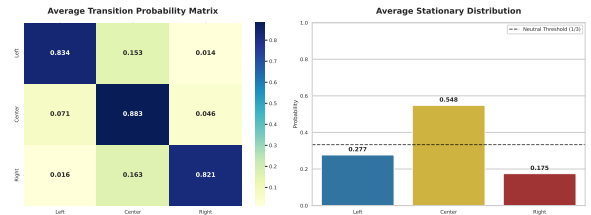


Figure 5: Aggregate dynamics of a simulated multi-model ecosystem. The average transition matrix $\bar{P}$ (left) and the average of the per-model stationary distributions (right) reveal that even a mixed environment leans heavily toward a center-dominant equilibrium (54.8%; the stationary distribution of $\bar{P}$ itself places 57.3% on Center), far exceeding the ideal neutral threshold of 1/3.

distribution, effectively maintaining a pluralistic environment.

As summarized in Table 2, this divergence results in substantial entropy reduction for specific models. Starting from an initial entropy of 1.09, Gemma3 reaches a stationary entropy of only 0.64, a 41.53% reduction in political diversity. LLaMa3.1 follows with a 28.62% reduction against its own 35-article baseline. Meanwhile, GPT5.1, Gemini2.5, and Qwen3 maintain reduction ratios below 3%, sustaining a diverse stationary state. These results indicate that for certain models, the neutral drift leads to a strongly convergent trajectory toward a homogenized stationary state where original political nuances are largely lost.

In a realistic environment, content is rarely generated by a single model but rather by an ensemble of diverse LLMs. To simulate this aggregate behavior, we assume an ecosystem where all eight evaluated models are utilized with equal frequency. By calculating the average transition probability

Model	GPT5.1	Gemini2.5	Qwen3	Exaone3.5	Gemma3	LLaMa3.1	GPT4o	Sonnet4.5
Initial Political Entropy	1.09 (1.07 for LLaMa3.1)							
Stationary Political Entropy	1.06	1.06	1.08	1.03	0.64	0.76	0.92	0.90
Entropy Reduction Ratio	2.59%	2.61%	1.35%	5.84%	41.53%	28.62%	15.48%	17.27%

Table 2: Stationary entropy analysis. Entropy reduction ratios quantify the relative loss of political diversity from the initial to the stationary state, revealing substantial long-term entropy reduction in Homogenizing Models compared to the stability of Preserving Models.

matrix $\bar{P}$,

$$\bar{P} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} P_g \quad (8)$$

we can observe the collective trajectory of the information lifecycle. As shown in Figure 5, the aggregate ecosystem still exhibits a strong tendency toward the Center state. The two natural aggregations answer slightly different questions. Averaging the per-model stationary distributions describes an ecosystem in which each content lineage is handled throughout by a single model, and predicts a Center share of 54.8%; the stationary distribution of $\bar{P}$ describes lineages whose generator is redrawn at every round, and is even more Center-concentrated (57.3%).

Even with the inclusion of Preserving Models, the homogenizing gravity of others remains dominant enough to skew the entire ecosystem toward a center-heavy distribution. Under this simplified ecosystem model, diversity-preserving models alone do not suffice to counteract the broader trend of political homogenization. We further note that the closed loop analyzed here is a worst-case limit in which no new human-authored content enters the pipeline: Appendix J extends the analysis with a restart formulation in which fresh human-authored content is injected at every step, and shows that a substantial Center bias persists even under high injection rates—27% above the balanced ideal even when 30% of the content entering the pipeline at each step is human-authored.

Figure 6 applies the same machinery to two-model mixtures. For a mixing weight w we form the kernel $wP_{\text{GPT5.1}} + (1 - w)P_{\text{GPT4o}}$, the transition operator induced by drawing the generator at random at each round, and compute its stationary distribution. The equilibrium shifts monotonically toward the Center as the weight of GPT4o grows, even when GPT5.1 remains prevalent: an equal mixture already places 0.536 on the Center state. The homogenizing effect therefore persists in mixed-model environments rather than being averaged away.

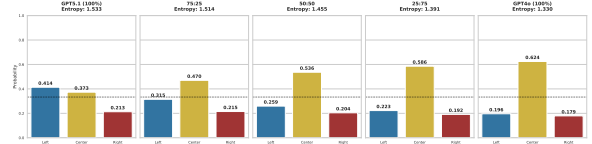


Figure 6: Simulation of model mixing dynamics between a preserving model (GPT5.1) and a homogenizing model (GPT4o) under varying invocation probabilities. As the proportion of GPT4o increases, the stationary political distribution progressively shifts toward Center dominance, indicating that even partial exposure to homogenizing models is sufficient to erode pluralistic stability maintained by preserving models. Entropy values in the panel titles are reported in bits, whereas Eq. (7) and Table 2 use nats.

4.3 Mitigating Homogenization: Self-Identified Stance Preservation and Stance-Conditioned Generation

The Center-directed drift documented in Section 4.1 substantially erodes political diversity, and is difficult to reverse under unmodified recursive generation. We hypothesize that it stems from the progressive loss of stance-specific signals in the input rather than from a loss of the models’ capacity to express partisan framing. SCG and SISP serve primarily as *diagnostic probes* of this hypothesis: if restoring the stance signal through conditioning recovers the distribution, the observed drift reflects recoverable signal loss—a default generation behavior under weak stance-preservation signals. Accordingly, we focus this evaluation on the previously identified Homogenizing Models, which exhibited the most pronounced diversity erosion.

We categorize our generation strategies into three increasingly constrained control levels to evaluate their effectiveness across the selected models. No Control (w/o Control) serves as the baseline where the generator $\mathcal{G}$ performs vanilla recursive generation defined as $c_i^r = \mathcal{G}(c_i^{r-1})$, while internal control is implemented using the SISP prompt, which adopts a chain-of-thought-style prompting strategy (Wei et al., 2022; Kojima et al., 2022) to

Model	GPT4o	Sonnet4.5	Gemma3	LLaMa3.1
Initial Dist.	0.38 / 0.33 / 0.28			
w/o Control	0.21 / 0.62 / 0.18	0.28 / 0.67 / 0.05	0.13 / 0.72 / 0.15	0.14 / 0.74 / 0.11
<i>p</i> -value	0.0169	0.0109	0.0071	0.0068
SISP	0.42 / 0.14 / 0.42	0.48 / 0.11 / 0.40	0.70 / 0.08 / 0.21	0.72 / 0.11 / 0.16
<i>p</i> -value	0.0539	0.0364	0.0182	0.3821
Predicted SCG	0.53 / 0.17 / 0.28	0.43 / 0.27 / 0.29	0.51 / 0.17 / 0.30	0.71 / 0.00 / 0.28
<i>p</i> -value	0.1307	0.5853	0.1738	0.0045
Ground Truth SCG	0.41 / 0.30 / 0.28	0.33 / 0.36 / 0.30	0.41 / 0.25 / 0.33	0.61 / 0.07 / 0.30
<i>p</i> -value	0.9394	1.0000	0.5968	0.0224

Table 3: Left/Center/Right stance distributions under different control strategies. For each model, stance distributions obtained under the w/o Control, SISP, Predicted SCG, and Ground Truth SCG conditions are compared, using the Stuart–Maxwell test, against the round-0 labels of the articles completed in that condition; the Initial Dist. row shows the full 39-article reference. Proportions are rounded or truncated and may not sum to one (Appendix B).

allow the model to infer the stance from the input text without relying on explicit stance labels.

In contrast, external control—SCG—explicitly constrains the generator with a target label $s \in \{L, C, R\}$ such that $c_i^r = \mathcal{G}(c_i^{r-1}, s)$, using either ground-truth ($s = y_i$) or predicted ($s = \hat{y}_i^{r-1}$) labels produced by the judge model $\mathcal{J}$. The statistical homogeneity between the resulting $L/C/R$ distributions and the initial state for each model is evaluated using the Stuart–Maxwell test.

The experimental results in Table 3 reveal that internal control via SISP does not restore the initial balance: it suppresses the Center share well below its initial level (0.08–0.14 vs. 0.33), trading center-collapse for over-polarization, and two of the four SISP distributions still differ significantly from the initial state. External control provides a substantially stronger corrective effect.

Notably, the effectiveness of SCG remains robust even when the generator is conditioned on the predicted stance ($\hat{y}_i^{r-1}$) without ground-truth stance labels. This finding is particularly important for practical applications, as ground-truth stances are rarely available in real-world scenarios; it suggests that a model-driven feedback loop, where the system estimates and then enforces a stance, offers a potentially practical mechanism for mitigating diversity erosion in self-referential generation pipelines. For models such as GPT4o and Sonnet4.5, this external guidance substantially reduced the Center drift, leaving distributions that are no longer statistically distinguishable from the initial state ($p = 0.1307$ and $p = 0.5853$, respectively). Predicted SCG operates at the *distribution level*: it brings the marginal stance distribution back toward its initial state, while at the instance level judge errors persist across rounds without cascad-

ing multiplicatively (Appendix K).

However, the efficacy of these strategies is model-dependent. For LLaMa3.1, no strategy restores the initial balance: both Predicted SCG ($p = 0.0045$) and Ground Truth SCG ($p = 0.0224$) remain significantly shifted, and while its SISP distribution is not statistically distinguishable from the initial state ($p = 0.3821$), that distribution is visibly polarized (0.72/0.11/0.16) and the non-significance may partly reflect the smaller sample for this model.

5 Conclusion

This study empirically demonstrates that self-referential loops in generative AI pipelines pose a systematic risk to the diversity of political framing in AI-mediated content. Our Markov formalization quantifies how the loop compounds each model’s measured per-step tendencies; for the homogenizing models, the resulting aggregate contraction is front-loaded and largely irreversible owing to the asymmetry of Center absorption. Furthermore, our analysis confirms that homogenization drives strong convergence toward a center-dominant equilibrium even within mixed-model ecosystems.

Erosion of stance diversity at scale would threaten democratic pluralism, but this outcome is not inevitable. Stance conditioning substantially reduces marginal drift for three of the four homogenizing models, although self-identified preservation overshoots rather than restoring balance. Preserving political diversity will therefore require deliberate model design, not passive reliance on generative systems.

6 Limitations

This study employed a precision-oriented design focusing on high-contrast events to isolate the structural mechanisms by which self-referential loops erode political pluralism. This strategic choice minimized initial noise, enabling the precise capture of convergence characteristics of LLMs. However, such a controlled approach inherently limits the representation of the nuanced and multidimensional political distributions found in the real world. In particular, the moderate-intensity range is nearly unpopulated among our polarized seeds (three of 96 pooled Left/Right article–model instances), so generalization to less polarized content remains an open question (Appendix G).

Future research should expand this framework to real-world datasets to explore how a broader spectrum of granular political perspectives is preserved or extinguished within complex AI-mediated information ecosystems. One concrete direction is *lineage-aware* generation: if the transformation history of content were represented structurally—e.g., as a provenance graph over documents and their AI-mediated derivations—retrieval pipelines could condition on it, prioritizing human-authored or low-depth content and thereby raising the effective rate of fresh human content entering the pipeline (Appendix J). Recent methods for conditioning LLMs on graph-structured inputs (Kim and Kim, 2026; Kim et al., 2026) offer a technical basis for exploring such lineage-conditioned pipelines.

In addition, all stance evaluations rely on an LLM judge. The judge agrees with the outlet-derived reference labels y_i on 84.62% of round-0 articles—the same set on which it was selected (Appendix C)—and is cross-checked against a second high-capability judge, and the Ground Truth SCG condition provides an internal control against judge artifacts; nevertheless, a full per-round human evaluation of the 6,160 paraphrased instances was beyond the resources of this study and remains an explicit item of future work. Relatedly, the Stuart–Maxwell tests reported throughout are exploratory and are not corrected for multiple comparisons, and non-significant results indicate only that we detect no marginal shift, not that the distributions are equivalent.

Furthermore, although we categorize models into Preserving and Homogenizing groups based on their observed stance dynamics, the limited transparency surrounding training data composition and

alignment procedures constrains deeper causal analysis. While our base-versus-aligned comparison (Appendix H) establishes alignment training as a model-dependent amplifier of homogenization, this opacity prevents us from attributing the effect to specific alignment components (e.g., safety tuning, preference data, or data composition), whether for proprietary models or for undisclosed open-weight recipes.

As a result, systematically examining the relationship between training methodologies and the stability of political signals remains an open and critical research challenge. Addressing this gap is essential for the development of next-generation language models that not only avoid excessive homogenization but also explicitly support and sustain pluralistic representation.

Ethical Considerations

This work studies political stance preservation and homogenization in recursive LLM generation. While the goal is to diagnose risks to political pluralism, stance-preserving generation can also be misused to maintain or amplify partisan persuasion, propaganda, or polarizing rhetoric. We therefore frame SCG and SISP as diagnostic and mitigation tools for controlled research settings, not as recommendations to preserve all political framings unconditionally. The dataset is constructed from public news articles and media-bias labels, and the study analyzes aggregate stance dynamics rather than individual-level political beliefs. We do not infer or profile the political preferences of private individuals.

Acknowledgments

This work was supported by research funding from a project conducted in collaboration with the Korea Technology Finance Corporation (KOTEC) (Project No. 2026-11-1514). This work was supported (in part) by the Yonsei University Research Fund (Post Doc. Researcher Supporting Program) of 2026 (project no.: 2026-12-0177).

We used ChatGPT (OpenAI) and Claude (Anthropic) to assist with English translation and writing during the preparation of this manuscript. All scientific content, experimental design, and conclusions are solely the responsibility of the authors.

References

- Anthropic. 2025. [Claude Sonnet 4.5 system card](#). Technical Report.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.
- Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. [Measuring political bias in large language models: What is said and how it is said](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11142–11159. Association for Computational Linguistics.
- Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Fabian Haak and Philipp Schaer. 2023. Qbias – a dataset on media bias in search queries and query suggestions. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 239–244.
- Wooyoung Kim, Haemin Jung, Byeongjin Kim, Suhyeon Kwon, and Wooju Kim. 2026. [Graph discrete prompt optimization for knowledge graph question answering](#). In *Proceedings of the ACM Web Conference 2026*, WWW ’26, pages 8353–8356, New York, NY, USA. Association for Computing Machinery.
- Wooyoung Kim and Wooju Kim. 2026. [Addressing information bottlenecks in graph augmented large language models via graph neural summarization](#). *Information Fusion*, 127:103784.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, 33:9459–9474.
- LG AI Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Yountae Jung, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, and 14 others. 2024. EXAONE 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:2412.04862*.
- Llama Team. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.
- OpenAI. 2025. GPT-5.1 instant and GPT-5.1 thinking system card addendum. <https://openai.com/index/gpt-5-system-card-addendum-gpt-5-1/>.
- OpenAI. 2026. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Jose Luis Paredes, Gregory Druck, Bevin Benson, and Ethan Smith. 2026. AI now writes as many online articles as humans. <https://graphite.io/five-percent/ai-now-writes-as-many-online-articles-as-humans-do>.
- Pew Research Center. 2024. Americans’ top policy priority for 2024: Strengthening the economy. <https://www.pewresearch.org/politics/2024/02/29/americans-top-policy-priority-for-2024-strengthening-the-economy/>.
- Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Damian Radcliffe. 2025. Journalism in the AI era: Opportunities and challenges in the global south and emerging economies. *Available at SSRN 5229014*.
- Luca Rettenberger, Markus Reischl, and Mark Schutera. 2025. Assessing political bias in large language models. *Journal of Computational Social Science*, 8(2):42.
- Ilya Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.
- Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. 2025. Diverse preference learning for capabilities and alignment. In *The Thirteenth International Conference on Learning Representations*.

Dirk HR Spennemann. 2025. Delving into: The quantification of AI-generated content on the internet (synthetic data). *arXiv preprint arXiv:2504.08755*.

Tianyu Wang, Akira Horiguchi, Lingyou Pang, and Carey E Priebe. 2025. LLM web dynamics: Tracing model collapse in a network of LLMs. *arXiv preprint arXiv:2506.15690*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Expert-Curated Dataset for High-Fidelity Analysis

A.1 Systematic Selection and Expert Curation

To isolate the fundamental mechanisms underlying political diversity erosion, the dataset was constructed through a multi-stage expert curation process. We initially identified four real-world events within each of seven major policy domains, for 28 candidate events in total. For each event, we collected news articles from Left-, Center-, and Right-leaning outlets as classified by AllSides, yielding a candidate pool of 84 articles.

To ensure an unambiguous signal for tracking stance transitions, this pool was subjected to a stringent filtering procedure conducted by researchers holding at least a Bachelor’s degree in the humanities or social sciences. Their task was to assess whether each event exhibited clear, robust, and consistent partisan divergence in framing and evaluative stance. Events characterized by neutral consensus, weak political salience, or ambiguous framing were systematically excluded to minimize potential confounding noise. Filtering did not leave a complete Left/Center/Right triple for every retained event: two events yielded a second Left-leaning article in place of a Right-leaning one, which is why the totals in Table 4 are 15/13/11 rather than 13/13/13. Table 4 summarizes the final dataset statistics.

	Left	Center	Right	Total
Climate Change	2	2	2	6
Economy and Jobs	3	3	3	9
Environment	3	2	1	6
Healthcare	1	1	1	3
Immigration	2	2	2	6
Race and Racism	3	2	1	6
Terrorism	1	1	1	3
Total	15	13	11	39

Table 4: Distribution of news articles across topics and political categories. Each row reports the number of articles associated with a given topic, grouped by media bias labels (Left, Center, Right) as defined by AllSides.

A.2 Strategic Rationale for Precision-Targeted Design

The dataset design follows a precision-targeted strategy intended to serve as a stringent diagnostic setting for LLMs. By focusing on events with the clearest political contrast, we adopt an approximately controlled setting in which political framing

is the dominant varying factor.

This high-resolution design reduces the likelihood that observed convergence toward centrist stances is attributable primarily to ambiguity or noise in the initial information distribution, making it more readily attributable to the recursive-generation dynamics studied here. Accordingly, the curated dataset is not intended to support population-level inference, but rather to function as a controlled testbed for characterizing the structural risks of diversity collapse in recursive AI-driven information ecosystems.

B Model Details

Table 5 lists the exact model identifiers of the eight generators used in this study. Judge model versions are detailed in Appendix C. For LLaMa3.1, four of the 39 seed articles failed to complete the 20-round loop due to persistent generation-format errors and are excluded from the main recursive-generation analyses of this model ($n = 35$); all other models completed all 39 articles in the main experiment. The control-strategy runs of Section 4.3 were executed separately and had their own format failures, so the number of articles completing all 20 rounds varies by model and condition; each Stuart–Maxwell test in Table 3 compares a condition against the round-0 labels of the articles completed in that condition.

Type	Short Name	Model Version
Proprietary	GPT5.1	gpt-5.1-2025-11-13
	GPT4o	gpt-4o-2024-11-20
	Gemini2.5	gemini-2.5-flash
	Sonnet4.5	claude-sonnet-4-5-20250929
Open-weight	Exaone3.5	EXAONE-3.5-7.8B-Instruct
	LLaMa3.1	Llama-3.1-8B-Instruct
	Gemma3	gemma-3-12b-it
	Qwen3	Qwen3-14B

Table 5: Exact model versions of the eight generators $\mathcal{G}$ used in this study.

C Judge Model Evaluation

This appendix reports the judge selection benchmark summarized in Section 3.3. We measured the zero-shot classification accuracy of four candidate models on the original articles (round 0) against the reference labels. As detailed in Table 6, the three strongest candidates perform comparably (84.62%, 82.05%, and 79.49%, i.e., differences of at most two articles at $n = 39$), while gemini-2.5-pro is clearly weaker (69.23%). We

adopt gpt-4o-2024-11-20, the most accurate of the four, as the primary judge, and verify below that the round-20 distributions are not sensitive to this choice.

Model	GPT4o	GPT5.1	Gemini2.5-Flash	Gemini2.5-Pro
Accuracy	84.62	82.05	79.49	69.23

Table 6: Performance comparison of candidate judge models. Accuracy is measured by comparing zero-shot stance predictions with the reference labels on the initial dataset. The value in bold indicates the best performance.

Because all stance evaluations in our framework rely on a single LLM acting as a judge, a natural methodological concern is that the observed stance dynamics could reflect model-specific biases of the chosen judge rather than genuine generative effects. To assess this risk, we conduct a robustness analysis by comparing the round-20 stance distributions produced by two high-capability judge models, gpt-4o-2024-11-20 and gpt-5.1-2025-11-13.

We quantify the discrepancy between the final predicted stance distributions using the Jensen–Shannon (JS) divergence to measure distributional similarity. In addition, we apply the Stuart–Maxwell test to explicitly evaluate whether the categorical stance distributions (Left/Center/Right) produced by the two judges differ significantly. As summarized in Table 7, the observed JS divergence is small, and the Stuart–Maxwell tests do not detect a significant difference between the stance distributions produced by the two judges.

Taken together, these results indicate that the observed stance dynamics are unlikely to be explained solely by the choice of judge model, although they do not establish equivalence between the two judges. This finding mitigates concerns regarding judge-induced artifacts and supports the robustness of our analyses under the LLM-as-a-judge evaluation framework.

D Hyperparameter Settings

D.1 Stance Classification Threshold

The stance classification threshold was established at $\tau = 0.3$ based on the empirical distribution of scores observed in the reference-labeled data. This threshold is justified by the clear clustering of Center samples around a score of 0.0, which positions them well within the $[-0.3, 0.3]$ range. In contrast, the Left and Right stances are strongly polarized, with their scores predominantly concentrated below

Model	GPT5.1	Gemini2.5	Qwen3	Exaone3.5	Gemma3	LLaMa3.1	GPT4o	Sonnet4.5
JS Divergence	0.0702	0.1205	0.0431	0.0980	0.0867	0.0449	0.0295	0.1004
p -value	0.4931	0.1614	0.5000	0.7034	0.5052	0.5000	0.8404	0.3385

Table 7: Results of a judge-model robustness analysis. Comparison of round-20 stance distributions obtained using gpt-4o-2024-11-20 and gpt-5.1-2025-11-13 as judge models. Jensen–Shannon divergence measures distributional similarity, while the Stuart–Maxwell test assesses statistical differences between categorical stance distributions.

−0.6 and above 0.6, respectively. As illustrated in Figure 7, the vertical dashed lines at ± 0.3 serve as a robust demarcation to minimize overlap and effectively distinguish neutral content from polarized political stances.

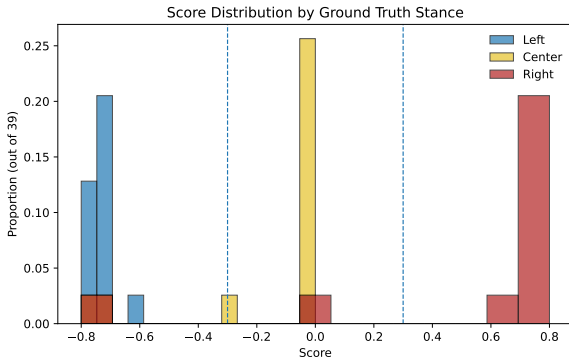


Figure 7: Score distribution by reference stance. The vertical dashed lines indicate the stance classification thresholds at $\tau = \pm 0.3$, which effectively separate Center, Left, and Right stances.

D.2 Temperature of LLM Decoding

The temperature parameter for GPT4o was tuned to balance the trade-off between text diversity and linguistic fidelity. In our preliminary experiments using GPT4o, lower temperature settings led to a verbatim reproduction phenomenon, where the model frequently regenerated the original articles word-for-word instead of performing the intended paraphrasing. Specifically, at $T = 0.0$ we observed 107 instances of such verbatim copying out of 780 generations, and even at $T = 0.4$ the issue persisted with 16 cases.

To mitigate this structural rigidity, we shifted our focus toward a relatively high-temperature regime for GPT4o, conducting analyses across $T = 0.8$, 1.0, and 1.2 following the same procedure described in Section 4.1. To assess the shift in political stances, we performed the Stuart–Maxwell test to compare the marginal distributions of the stances generated by GPT4o against the initial reference distribution.

The results indicated a statistically significant departure from the original distribution across the entire high-temperature spectrum: $T = 0.8$ ($p = 0.0169$), $T = 1.0$ ($p = 0.0065$), and $T = 1.2$ ($p = 0.0013$). The selection criterion was therefore the elimination of verbatim copying, which is a generation artifact rather than a stance phenomenon, and $T = 0.8$ is the lowest setting we tested at which it disappears; we fixed this value for all models. That the marginal shift is significant at $T = 0.8$, 1.0, and 1.2 alike indicates that the effect is not an artifact of this particular setting.

E LLM Prompt Details

Political Bias Evaluation Prompt. We use the following prompt to classify the political bias of news articles based on a reference summary.

System Prompt:

You are a political analyst. Your task is to classify the political bias of a news article into one of three categories: Left, Center, or Right. Use the provided baseline summary as a reference.

User Prompt:

[Baseline Summary]
{Summary from the AllSides News Team}
 [Target Article]
 Title: *{News Title}*
 Content: *{News Content}*

Task:

1. Analyze the target article in comparison with the baseline summary.
2. Determine the political stance of the article (Left, Center, or Right).
3. Assign a continuous bias score between -1.0 (Left) and 1.0 (Right).

Output Format (JSON):

```
{
  "reasoning": "Brief explanation
    of the classification",
  "score": float,
  "predicted_stance":
```

"Left/Center/Right"

}

Paraphrasing Prompt. To generate recursive variants of news articles while preserving political stances, we use the following paraphrasing prompt.

System Prompt:

Paraphrase the news title and content while strictly preserving facts and political nuance.

User Prompt:

Title: {News Title}

Content: {News Content}

Output Format (JSON):

```
{
  "p_title": "...",
  "p_content": "..."
}
```

Stance-Conditioned Generation (SCG) Prompt.

To generate stance-consistent paraphrases, we condition the LLM explicitly on a target political orientation and require strict preservation of both factual content and political framing.

System Prompt:

You are a professional news editor. Your goal is to paraphrase the text while strictly maintaining its original political stance, which is specified as {Target Political Stance}, and preserving factual integrity. Do not neutralize the tone; preserve the unique perspective associated with the {Target Political Stance} orientation.

User Prompt:

Target Stance to Preserve: {Target Political Stance}

Title: {News Title}

Content: {News Content}

Output Format (JSON):

```
{
  "p_title": "...",
  "p_content": "..."
}
```

Self-Identified Stance Preservation (SISP) Generation Prompt. To allow the model to self-identify the political stance of an article and preserve it during paraphrasing, we employ a stance-aware chain-of-thought prompting strategy.

System Prompt:

You are an expert news analyst and professional editor. Your task is to:

1. Analyze the input text to identify its political stance, underlying tone, and key arguments.
2. Explain this analysis in the ‘thought’ field as a chain-of-thought.
3. Paraphrase the text while strictly preserving the identified political stance and factual integrity.

Do not neutralize the tone; maintain the original perspective’s intensity and nuance.

User Prompt:

Title: {News Title}

Content: {News Content}

Output Format (JSON):

```
{
  "thought": "Detailed analysis of
    the original stance and
    how to preserve it",
  "p_title": "...",
  "p_content": "..."
}
```

F Within-Label Intensity Decomposition

The three-state representation could, in principle, mask gradual within-label moderation (e.g., far-left to moderate-left). The continuous-score metrics of the main text (VRR in Table 1, MAS in Figure 2) already capture such shifts; here we decompose the MAS by label to isolate them. At each round we track the mean intensity conditional on retaining a polarized label, $\mathbb{E}[|v_i^r| | \hat{y}_i^r \in \{L, R\}]$.

This quantity is essentially flat: from 0.73–0.75 at round 0 to 0.67–0.73 at round 20 across models, with the fraction of retained-label instances at $|v_i^r| \geq 0.6$ remaining between 0.78 and 1.0 throughout. The contraction thus proceeds not by gradual moderation within labels but by discrete absorption into the Center state—the phenomenon itself is state-transitional, which is precisely why the three-state Markov representation is appropriate.

G Dose–Response Analysis of Initial Stance Intensity

Because the seed set emphasizes high-contrast articles, one may ask whether the observed absorption merely reflects borderline articles near the classification threshold. Pooling the four Homogenizing

Models over articles judged Left or Right at round 0 ($n = 96$ article-model instances), we relate each article’s initial intensity $|v_i^0|$ to the absorption event $\hat{y}_i^{20} = C$. The Spearman rank correlation is $\rho = -0.074$ ($p = 0.475$): we detect no monotonic association between initial intensity and absorption.

By intensity band, articles with $0.6 \leq |v_i^0| < 0.75$ are absorbed at 67.3% and the most polarized articles ($|v_i^0| \geq 0.75$) at 59.1%—a small difference consistent with the null rank correlation. Absorption is therefore not a threshold artifact; if anything, high-contrast seeds are the *hard* case, carrying the most stance signal to strip away. We caution that the moderate band ($|v_i^0| < 0.6$) contains only three of the 96 pooled instances, so generalization to genuinely moderate content remains open (see the Limitations section).

H Base vs. Alignment-Trained Models

To test whether alignment training contributes causally to the observed homogenization—rather than the recursive loop or prompting alone—we compare base and alignment-trained checkpoints within three open-weight model families: Llama-3.1-8B (Llama-3.1-8B vs. Llama-3.1-8B-Instruct), Qwen2.5-7B (Qwen2.5-7B vs. Qwen2.5-7B-Instruct) (Qwen Team, 2024), and Gemma-3-12B (gemma-3-12b-pt vs. gemma-3-12b-it).

Design. The pipeline is identical to the main experiments: the same 39 seed articles, 20 paraphrasing rounds at temperature 0.8, and judging by gpt-4o-2024-11-20 at temperature 0 with $\tau = 0.3$. Because base models lack a chat template, a naive base-vs-instruct comparison would confound alignment with prompting format. We therefore run *both* members of each pair with the same few-shot completion prompt, so that within a pair only the weights differ. (We verified in preliminary runs that few-shot prompting leaves instruct-model behavior essentially unchanged relative to chat prompting.) Since round 0 is identical across arms, we judge the original 39 articles once and share this round-0 baseline across all arms ($\text{Var}(v^0) = 0.3213$; 39/39 valid), so that all VRR values share a common denominator. This baseline comes from a separate judging pass and therefore differs slightly from the 0.3626 used in the main experiments. Such run-to-run variability is small at the article level: each of the eight main-experiment runs judged round 0 sep-

arately, and across those passes—one of which supplies the shared baseline of Table 1—article-level scores correlate at pairwise Pearson $r = 0.945$ –0.999 and stance labels agree on 92–100% of articles, so the variability affects the variance baseline more than the stance assignments. Note that these VRR values are computed under few-shot prompting with this shared baseline and are therefore not directly comparable to the chat-prompted VRR values in Table 1.

Family	Base	Aligned	ΔVRR	MAS@20
Llama-3.1-8B	42.9%	55.0%	+12.1	0.249 $\rightarrow$ 0.213
Qwen2.5-7B	28.1%	27.1%	−1.0	0.326 $\rightarrow$ 0.323
Gemma-3-12B	22.0%	59.3%	+37.3	0.338 $\rightarrow$ 0.200

Table 8: VRR (higher = stronger homogenization) for base vs. alignment-trained checkpoints under matched few-shot prompting. ΔVRR is aligned minus base, in percentage points; MAS@20 is the round-20 Mean Absolute Score (base $\rightarrow$ aligned). All arms completed 39/39 articles with no judging errors.

Results. As shown in Table 8, alignment training increases homogenization markedly in two of the three families (Llama +12.1, Gemma-3 +37.3 VRR points), leaves Qwen2.5 essentially unchanged (−1.0, within single-pass judging noise), and decreases it in no family. The two amplified families correspond to the two open-weight Homogenizing Models of the main study, whose base counterparts homogenize substantially less. At the same time, the base checkpoints still contract variance by 22–43%, showing that the recursive loop alone induces homogenization and that alignment amplifies it.

Scope. These results support alignment training as a model-dependent amplifier, not a universal cause: the magnitude is family-dependent, and this comparison measures the net effect of the full alignment stage without attributing it to specific components (safety tuning, preference data, or data composition). We also note that the Qwen family member used here (Qwen2.5-7B) differs from the model in the main study (Qwen3-14B, whose base checkpoint was unavailable in a comparable form); it serves as an additional family data point rather than a direct counterpart. VRR is computed from rounds 0 and 20 with a single judging pass and $n = 39$ seeds, without confidence intervals.

I Validation of the Time-Homogeneous Markov Approximation

The Markov analysis in Section 4.2 pools transition pairs across rounds into a single kernel $\hat{P}$, assuming time-homogeneity. Front-loaded aggregate change does not contradict this assumption: with a fixed kernel, a chain initialized far from its stationary distribution moves fastest in early rounds and slows as it approaches equilibrium (geometric convergence). Indeed, the fitted homogeneous chains themselves predict that 61–81% of the total Center increase of the Homogenizing Models occurs within the first three rounds, matching the observed front-loading. We validate the assumption directly in three ways.

(a) Early- vs. late-round kernels. For each model we estimate $\hat{P}_{\text{early}}$ from transitions into rounds 1–5 and $\hat{P}_{\text{late}}$ from transitions into rounds 6–20, and test row-wise (per source state) whether the two arrival distributions are homogeneous. Because some late-round cells are sparse, we use permutation tests (10,000 resamples) rather than the χ^2 approximation, yielding 8 models $\times$ 3 rows = 24 tests. In 22 of 24 rows, the difference is not significant ($p > 0.05$). The two exceptions both belong to Gemma3 and point in the direction of $\hat{P}_{\text{early}}$ being *more* absorbing than the pooled kernel—i.e., the homogeneous approximation, if anything, underestimates early contraction.

(b) Trajectory reproduction. We apply the pooled kernel iteratively to the observed initial distribution to obtain predicted marginals $\hat{\pi}_r = \pi_0 \hat{P}^r$, and compare them with the observed distributions using the total variation distance $\text{TV} = \frac{1}{2} \|\hat{\pi}_r - \pi_r^{\text{obs}}\|_1$. Across all eight models, the mean TV over rounds is 0.031–0.059 (maximum 0.173, a single Gemma3 round)—an error on the order of one to two articles out of 39.

(c) Held-out forecasting. We extended the GPT4o loop to 30 rounds, estimated $\hat{P}_{1:20}$ from rounds 1–20 only, and forecast the stance distributions of rounds 21–30 starting from the observed round-20 distribution. The held-out mean TV is 0.038 (maximum 0.094), and the empirical kernel estimated separately from rounds 21–30 differs from $\hat{P}_{1:20}$ by at most 0.058 per entry. The rounds 1–20 estimate likewise differs from the pooled 30-round kernel used for Figure 3 by at most 0.017 per entry, with indistinguishable stationary quantities.

J Ecosystem Projection with Human Content Injection

The closed loop of the main analysis is a worst-case limit in which no new human-authored content enters the pipeline. To quantify how the equilibrium changes when it does, we replace the transition matrix P of each model with

$$P(\alpha) = (1 - \alpha) P + \alpha \mathbf{1} \pi_0^\top, \quad (9)$$

where π_0 is the stance distribution of the human-authored seeds and the second term corresponds to a fresh human-authored article entering the pipeline with probability α at every step. For each α we compute, for each model, the stationary distribution $\pi(\alpha)$ satisfying $\pi(\alpha)P(\alpha) = \pi(\alpha)$, and average across models, following the same convention as Figure 5; no new generation experiments are required. The closed loop of the main text is the $\alpha \rightarrow 0$ limit, and $\alpha = 0$ exactly reproduces the 0.548 Center equilibrium of Figure 5.

α	Ecosystem Center	Homogenizing-only	Entropy loss
0.00	0.548	0.690	9.2%
0.05	0.516	0.637	6.6%
0.10	0.490	0.594	4.9%
0.20	0.450	0.529	2.7%
0.30	0.422	0.482	1.6%
0.50	0.383	0.417	0.5%

Table 9: Equilibrium Center share under human-content injection rate α , for the eight-model ecosystem and for the four Homogenizing Models alone, with the corresponding loss of political entropy. The balanced ideal is $1/3 \approx 0.333$.

As shown in Table 9, the Center bias decays continuously with α but remains substantial even under high injection rates: at $\alpha = 0.3$ (30% fresh human content at every step), the ecosystem Center equilibrium is still 27% above the balanced ideal (45% above for the homogenizing subset). As the share of AI-generated content on the web grows, the effective α of real information ecosystems declines, moving them along this curve toward the closed-loop limit.

K Instance-Level Error Dynamics of Predicted SCG

Predicted SCG conditions each round on the judge’s previous-round label, $c_i^r = \mathcal{G}(c_i^{r-1}, \hat{y}_i^{r-1})$, so judge errors can propagate across all 20 rounds; Table 3 is therefore an end-to-end test of error accumulation. Here we characterize the error dynamics

at the instance level, using the per-round conditioning labels recorded during the experiment.

If judge errors compounded multiplicatively, an $\approx 85\%$ -accurate judge would leave agreement of only $0.85^{20} \approx 4\%$ after 20 rounds. Instead, agreement with the original reference stance stabilizes at 58–73% by round 20 (GPT4o 72%, Sonnet4.5 73%, Gemma3 64%, LLaMa3.1 58%). The round-to-round dynamics explain the plateau: once an instance’s label is wrong, it rarely recovers ($\Pr[\hat{y}_i^r = y_i \mid \hat{y}_i^{r-1} \neq y_i] = 3\text{--}5\%$ per round), but new errors also enter slowly ($\Pr[\hat{y}_i^r = y_i \mid \hat{y}_i^{r-1} = y_i] = 97\text{--}99\%$ per round). Errors persist but do not cascade; the system approaches a linear-saturating equilibrium rather than a multiplicative collapse.

Model	Predicted SCG	w/o Control	Δ
GPT4o	0.718	0.615	+0.103
Sonnet4.5	0.730	0.487	+0.243
Gemma3	0.641	0.359	+0.282
LLaMa3.1	0.579	0.429	+0.150

Table 10: Per-instance retention (fraction of instances whose round-20 label equals the reference stance) under Predicted SCG vs. uncontrolled generation.

Per-instance retention also improves substantially over uncontrolled generation (Table 10, +10 to +28 percentage points). Nevertheless, these dynamics confirm that Predicted SCG should be understood as a distribution-level mitigation (it brings the marginal stance distribution back toward its initial state; Table 3) rather than an instance-level guarantee.