

Computational Processing of Political Bias Using AI Agents

Peter Theodoracopoulos, Manoj P. Vijayakumar, Vijay K. Madiseti

College of Computing, Georgia Institute of Technology, Atlanta, GA, USA

Email: petetheo@gatech.edu, manojprabhakaranv@gatech.edu, vkm@gatech.edu

How to cite this paper: Theodoracopoulos, P., Vijayakumar, M.P. and Madiseti, V.K. (2026) Computational Processing of Political Bias Using AI Agents. *Journal of Information Security*, 17, 540-568.

<https://doi.org/10.4236/jis.2026.174024>

Received: September 23, 2026

Accepted: October 25, 2026

Published: October 28, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution-NonCommercial International License (CC BY-NC 4.0).

<http://creativecommons.org/licenses/by-nc/4.0/>



Open Access

Abstract

Given the widespread presence of media bias favoring political groups on social media, discerning inherent bias is challenging. Existing methodologies often rely on an oversimplified, unidimensional (left/right) political spectrum. Furthermore, most research has focused on bias detection rather than mitigation. This paper introduces and evaluates POLQUAD, a novel multi-agent AI framework designed to mitigate political bias in social networks. Our approach utilizes the two-dimensional political compass (Authoritarian/Libertarian and Left/Right) to achieve nuanced classification. The POLQUAD framework employs specialized Opinion Agents, each prompted to represent one quadrant, to analyze text. A central Judge Agent synthesizes these perspectives and iteratively refines content toward neutrality. We conducted a study comparing our Full POLQUAD against two simpler iterative models (Unified and Naive) using samples from a dataset of 79,745 tweets from US Senators. All frameworks performed reasonably on text with low bias; however, the Full POLQUAD framework outperformed the others on a curated high bias dataset (bias magnitude > 7.0). On this dataset, the Full POLQUAD framework maintained high effectiveness while the Naive and Unified models' performance collapsed. Importantly, the Full POLQUAD framework was the most efficient, converging on a neutral solution in fewer iterations. A subsequent ablation study confirmed that removing any of the four political agents degrades performance. The results validate that its *synthesis-then-refine* architecture provides a more stable, efficient, and robust solution for mitigating bias in highly polarized content.

Keywords

Bias Mitigation, Iterative Refinement, Large Language Models, LLM as a Judge, Multi-Agent Systems, Political Bias, Political Compass, Social Media

1. Introduction

The rapid expansion of social media has transformed political communication, enabling the widespread circulation of highly polarized content and reinforcing ideological echo chambers. As a result, extensive discourse now occurs among groups with distinct interests and ideologies. Although recent advances in large language models (LLMs) have improved the detection of political bias, most existing approaches rely on simplified left-right classifications and focus primarily on identification rather than mitigation. Current debiasing methods tend to be fragile when confronted with strongly biased or emotionally charged text. To address these limitations, this work introduces POLQUAD, a multi-agent debiasing framework grounded in the two-dimensional political compass. By synthesizing and reconciling perspectives from all four political quadrants, POLQUAD aims to deliver a more stable, interpretable, and robust mechanism for reducing political bias in short form content.

1.1. Political Compass

While the majority of existing strategies employ a linear scale to represent political orientation along left-right dimensions, this work proposes the adoption of a quadrant-based methodology to achieve a more nuanced classification. This classifier is constructed upon the framework of the political compass test. As depicted in **Figure 1**, the political quadrants constitute the foundational basis of this study.

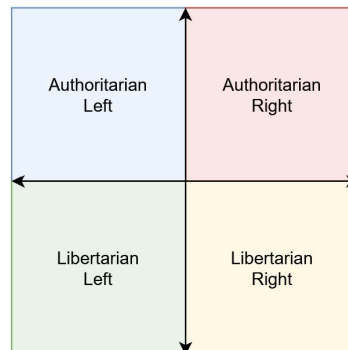


Figure 1. The political compass model showing the four quadrants.

The political compass model [1] is characterized by two axes: the x-axis reflects the economic inclination, whereas the y-axis indicates the social inclination. By employing these axes, a specific position is determined within the plane of the political compass, elucidating the degree to which a system aligns with the left-right and authoritarian-libertarian spectra. Consequently, this facilitates the establishment of a quadrant-based framework for classifying political stances. Each quadrant is defined as follows.

- **Authoritarian Left:** Supports strong government control of economy and individual behavior. Ideally supporting wealth redistribution, economic equality, social equality, and centralized planning. Prefers authority over liberty and

prefers collective economy over individual economy. This is represented by the blue color coded quadrant in **Figure 1**.

- **Authoritarian Right:** Supports powerful government with strict social law and order. Supports nationalism, and often a capitalist or corporatist economy controlled or heavily influenced by the state. Prefers authority over liberty and individual economy over collective economy. This is represented by the red color coded quadrant in **Figure 1**.
- **Libertarian Right:** Supports and gives priority to individual freedom in social and economic issues. Prefers minimal government intervention, free markets, and personal responsibility. Prefers liberty over authority and individual economy over collective economy. This is represented by the yellow color coded quadrant in **Figure 1**.
- **Libertarian Left:** Supports social equality, economic equality and personal freedoms, seeking low intervention from the government with community-based solutions and anti-capitalist or post-capitalist economics. Prefers liberty over authority and collective economy over individual economy. This is represented by the green color coded quadrant in **Figure 1**.

These four political quadrants provide the structural basis for the POLQUAD debiasing approach introduced in the following sections.

1.2. Bias Magnitude

To enable quantitative analysis throughout this work, each statement's position on the political compass is converted into a pair of numerical coordinates: an economic x -value (Left/Right) and a social y -value (Authoritarian/Libertarian).

The overall *bias magnitude* is defined as the Euclidean distance of this point from the neutral origin (0, 0).

$$\text{Bias Magnitude} = \sqrt{x^2 + y^2} \quad (1)$$

This scalar provides a single, interpretable measure of both the direction and intensity of the political bias and serves as the primary metric to evaluate debiasing performance in the experiments that follow.

2. Related Work

A broad body of research has examined political bias in text, though most work focuses on detection rather than mitigation. Nearly all adopt a one dimensional, left/right scale, as seen in **Figure 2**, rather than a multi-quadrant representation.



Figure 2. Left/Right one dimensional classification.

Early approaches used transformer and recurrent architectures to classify political leaning. Baly *et al.* [2] and AlDahoul *et al.* [3] employed BERT and LSTM

models to categorize text into Left, Right, and Center classes, while Hernandez and Corsi [4] used GPT-based models for domain level labeling. These approaches, while effective for coarse grained classification, do not incorporate the authoritarian-libertarian dimension and provide no mechanism for reducing bias in text.

Other studies explore broader bias assessment. The LLMBI framework [5] introduced scalar indices for multiple sociocultural categories, including a socioeconomic dimension loosely related to political orientation. However, its single score representation lacks the granularity needed to model the four political quadrants used in this work and does not include any debiasing method.

A parallel line of research investigates LLM behavior rather than text bias. Ye *et al.* [6] analyze bias generation in LLM-as-a-judge settings, while Zheng *et al.* [7] examine the consistency and limitation of LLMs acting as evaluators in multiple turn dialogue systems. Separately, Röttger *et al.* [8] analyze how reliably LLMs can be placed on a political compass. These studies highlight instability and systemic bias in LLM reasoning but do not address the challenge of neutralizing biased input text.

Several related efforts focus on partial labeling or moderation. PoliPrompt [9] uses LLMs for topic level political annotations but is not intended for debiasing. Podolak *et al.* [10] and Park *et al.* [11] investigate hate speech or cross cultural moderation using retrieval or human-LLM collaboration, but their binary or coarse labels do not extend to political neutrality across multiple axes.

Across these lines of research, two gaps remain consistent:

- 1) a lack of frameworks that incorporate all four political quadrants, and
- 2) an absence of systematic methods for actively mitigating political bias

Table 1 summarizes these limitations. The POLQUAD framework introduced in this work address both gaps by leveraging multi-agent synthesis and iterative refinement within a two-dimensional political compass.

Table 1. Summary of prior studies on LLM-based political content detection and bias mitigation approaches.

Year	Framework	For Political Content	Uses LLMs to Generate New Text?	Has Political Labels?	Quadrant Based	Does Bias Reduction?
2020	"We can detect your bias: ..."	Yes	No	Yes (Left, Right, Center)	No	No
2024	"LLMs left, right, and center: ..."	Yes	No	Yes (Left, Right, Center)	No	No
2023	"LLMBI"	Partially (A generic list of index scores with social-economic values as one)	No	Partially (Political bias is a single score)	No	No
2024	"Justice or prejudice? ..."	Partially (Not specialized to political content)	Yes (to generate biased answers)	No	No	No
2024	"PoliPrompt"	Partially (Not specialized to political content)	Yes (through prompting)	No (partial labels like technology, business)	No	No

Continued

2024	“LLM generated responses to mitigate the impact of hate speech”	Partially (Focus is hate speech, not political ideology)	Yes (via retrieval augmented generation)	No (only harmful vs non-harmful)	No	Partially (Focus is hate speech, not political ideology)
2024	“Political compass or spinning arrow? ...”	Yes	No (Evaluation of LLM biases only)	Yes (To quantify LLM biases)	Yes	No (Bias reduction is not discussed)

3. System Architecture and Implementation

To realize the proposed debiasing frameworks, we designed and implemented a robust, object-oriented system in Python. The system was engineered for stability, extensibility, and precise experimental control. The implementation is divided into three primary areas: core software components, the debiasing framework logic, and the data experimentation pipeline.

3.1. Architecture Overview

The POLQUAD system is implemented as a modular multi-agent architecture designed around the political compass model. The framework separates three key functions:

- 1) generating ideologically grounded opinions through quadrant specific agents,
- 2) synthesizing these perspectives via a neutral Judge Agent, and
- 3) iteratively refining the synthesized statement until its bias magnitude falls below a predefined threshold.

This architecture enables the system to produce balanced initial candidates that reflect diverse political viewpoints and improves convergence stability during refinement. The following subsections describe the software components, debiasing algorithms, and experimentation pipeline used to evaluate the architecture.

3.2. Core Software Components

Instead of standalone functions, our system is built on a set of specialized classes, each with a distinct responsibility:

- **LLM Clients:** Each LLM backend (e.g., Gemini, OpenAI, Claude) is wrapped in a dedicated client class that manages API calls, request formatting, retry logic, and rate limits. An exponential backoff handler ensures robustness under API failures and allows large experimental batches to run without interruption.
- **Bias Calculator Class:** This class operationalizes the political coordinate system introduced in Section 1.1. The economic (x) and social (y) axes are mapped to a continuous 2D plane:

$$x \in [-10.0, +10.0], y \in [-10.0, +10.0] \tag{2}$$

where x spans from Left (-10.0) to Right ($+10.0$) and y spans from Libertarian (-10.0) to Authoritarian ($+10.0$). Overall bias magnitude $|\vec{v}|$ is calculated as the Euclidean distance from the origin $(0, 0)$ via (1), yielding a range of $[0.0, 14.14]$.

To minimize noise, the calculator executes three repeated evaluation passes

(bias calc runs = 3) for every measurement and averages the extracted coordinates. To eliminate baseline coordinate drift, an initial calibration vector $\vec{b}_0 = (x_0, y_0)$ is computed using the semantically neutral anchor phrase “*The sky is blue and grass is green*” and subtracted from raw evaluations:

$$\vec{v}_{\text{calibrated}} = \vec{v}_{\text{raw}} - \vec{b}_0 \quad (3)$$

The calculator queries the LLM using greedy decoding with structured JSON schema constraints to guarantee valid coordinate extraction.

- **Agent Classes:** We defined three distinct agent personas as classes.

1) **Opinion Agent:** This agent is configured with a firmly defined ideological identity corresponding to one of the four specific political quadrants. The agent produces a concise opinion on the input text strictly from its assigned ideological stance.

2) **Judge Agent:** This agent functions as a neutral synthesizer. It receives the opinions produced by the Opinion Agents and generates a single balanced rewrite that preserves the meaning of the original statement while reducing political bias.

3) **Unified Agent:** This agent uses chain-of-thought prompting to develop a single, neutral perspective. Unlike the Full POLQUAD architecture, it considers multiple political viewpoints implicitly within one prompt and does not use quadrant specific agents.

3.3. Debiasing Framework Algorithms

The core components were used to build the three iterative frameworks for our comparative analysis. Each framework is a distinct algorithm that iterates until the bias magnitude measured by the Bias Calculator falls below a predefined threshold or a maximum number of iterations is reached.

1) **Naive Framework:** A simple iterative approach that repeatedly requests a more neutral rewrite. It does not use quadrant specific opinions, coordinate information, or synthesis.

2) **Unified POLQUAD Framework:** A single agent refinement process inspired by chain-of-thought prompting [12]. The Unified Agent considers the political compass dimensions within one prompt and iteratively adjusts the text until it meets the neutrality threshold.

3) **Full POLQUAD Framework:** The complete multi-agent architecture.

a) **Step 1 (Synthesis):** Four Opinion Agents provide ideological interpretations of the original text.

b) **Step 2 (Judgment):** The Judge Agent synthesizes the four perspectives into a balanced initial rewrite.

c) **Step 3 (Refinement):** The Judge Agent iteratively refines its own output until the bias magnitude falls below the threshold.

4) **Ablation Frameworks:** Ten Full POLQUAD architectural variations that remove one or more Opinion Agents. These tests isolate the functional contribution of each quadrant and demonstrate the necessity of all four viewpoints for stable performance.

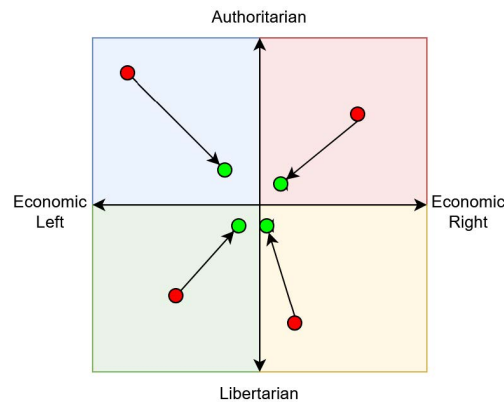


Figure 3. Conceptual illustration of debiasing.

Figure 3 provides a high level view of the debiasing process used in the POLQUAD framework. The method operates in two distinct phases: a *synthesis* phase that aggregates viewpoints from the four political quadrants, and a *refinement* phase in which a Judge Agent iteratively rewrites the synthesized statement to move it toward political neutrality. This design ensures that the initial candidate neutral statement is informed by opposing ideological perspectives rather than relying solely on iterative editing of the original biased text. The figure illustrates the flow of information among the agents and highlights how the framework balances conflicting perspectives before optimization begins.

3.4. Experimentation Pipeline

The experimental pipeline manages data loading, sampling, debiasing runs, bias evaluation, and result aggregation.

- **Dataset:** The m-newhauser/senator-tweets dataset from Hugging Face containing 79,745 political tweets.
- **Sampling:** Random sampling with fixed seeds (42, 1337, and 1984) ensures reproducibility.
- **High Bias Curation:** Statements with bias magnitude >7.0 are extracted using the Bias Calculator to form a curated dataset.
- **Metrics:** For each framework and LLM backend, we compute:
 - 1) Success Rate (%) (percent achieving threshold)
 - 2) Average Bias Reduction
 - 3) Iterations Required
 - 4) Overcorrection Rate
- **Automation:** A report generation script aggregates results from all runs and ablations.

3.5. System Configuration and Experimental Parameters

All experiments and agent loops were executed under standardized hyperparameters across all LLM backends:

- **Neutrality Threshold (τ):** The stopping threshold was set to $\tau = 1.0$. Input

statements with an initial baseline bias magnitude $|\vec{v}| \leq 1.0$ were classified as neutral and bypassed refinement. During iterative debiasing, the loop terminated immediately once $|\vec{v}| \leq 1.0$.

- **Iteration Limit ($K_{\max}$):** The hard iteration ceiling was set to $K_{\max} = 3$ iterations. Framework runs terminated upon reaching 3 iterations if the threshold was not met.
- **Evaluation Runs per Statement:** Each statement was independently evaluated across 3 repeated runs (num runs = 3) to capture run-level variance and compute exact overcorrection rates.
- **Model Generation Settings:**
 - *Text Rewriting* (Opinion, Judge, Unified, and Naive Agents): Invoked at temperature $T = 0.7$ with a maximum output limit of 1024 tokens to allow diverse synthesis.
 - *Coordinate Extraction* (Bias Calculator): Invoked at temperature $T = 0.0$ (deterministic greedy decoding) under strict JSON schema enforcement (response_schema/response_format).
- **Backend Model Families:** Evaluated backends comprised Google gemini-2.5-flash-lite, OpenAI gpt-4o-mini, and Anthropic claude-3-5-haiku-20241022.
- **Backend Pairing:** In all comparative runs, the Bias Calculator used the *same backend* as the rewriting framework under test (*i.e.*, Gemini evaluated Gemini rewrites, OpenAI evaluated OpenAI rewrites, and Claude evaluated Claude rewrites) to measure within-model alignment and self-refinement capabilities.

The complete, verbatim system instructions and prompts for the Bias Calculator, Opinion Agents, Judge Agent, and baseline configurations are provided in **Appendix** subsection A.

4. Analysis and Results

The efficacy of our proposed frameworks was evaluated in a two-phase experimental process. The first phase consisted of a broad, exploratory analysis on an uncured dataset of U.S. Senator tweets to identify performance trends. The second, more rigorous phase involved a multi-model stress test on a curated, high bias dataset to validate our initial findings.

4.1. Experiment 1: Exploratory Analysis on Uncured Dataset

We first conducted a broad analysis using the Senator Tweets dataset described in Section 3.4. Baseline performance on uncured political text was evaluated using three random seeds (42, 1337, and 1984), each drawing 100 statements evaluated across 3 repeat runs, producing 900 total statement evaluations across 300 unique statements. The frameworks were configured to skip processing on statements with an initial baseline bias magnitude below 1.0. Exactly 153 evaluations fell below this threshold, leaving an identical evaluation benchmark of $n = 747$ valid statement evaluations across 249 unique statements (252 from seed 42, 249 from seed 1337, and 246 from seed 1984). All three frameworks were evaluated on this

balanced dataset ($n = 747$) using the Gemini API (Table 2).

On this broad dataset, all frameworks were reasonably effective, with the FULL POLQUAD framework holding a modest lead in success. Interestingly, the Naive framework was the most efficient (fewest iterations) on its successful runs. This suggests that for the low-to-mid bias statements common in this uncurated set, a simple iterative approach is often sufficient, as shown in Table 2. The Full POLQUAD framework's performance advantage was most pronounced on statements with high to extreme initial bias, motivating a more focused stress test on exclusively high bias text.

Table 2. Framework performance on uncurated dataset ($n = 747$).

Metric	Iterative Naive	Unified POLQUAD	Full POLQUAD
Success Rate	68.94%	69.34%	78.18%
Avg. Bias Reduction	63.13%	64.25%	67.34%
Avg. Iterations (Success)	1.72	1.80	1.82

4.2. Experiment 2: Multi-Model Tests on High Bias Data

Based on the insight from Experiment 1, we curated a high bias dataset by extracting all statements from the senator tweets corpus exhibiting an initial bias magnitude > 7.0 (808 qualifying tweets). To evaluate performance across multi-seed variance, our primary Gemini testbed evaluated 100 statements sampled per seed across seed 42, 1337, and 1984 ($n = 300$ evaluations). Due to sampling with replacement across the high bias pool, these 300 evaluations comprise exactly 199 unique tweets:

- **Seed 42:** 100 tweets
- **Seed 1337:** 100 tweets (39 overlap with Seed 42)
- **Seed 1984:** 100 tweets (46 overlap with Seed 42, 36 overlap with Seed 1337, and 20 common to all three)

For cross-model verification, OpenAI's gpt-4o-mini and Anthropic's claude-3-5-haiku-20241022 were evaluated on the exact same 100 statements from Seed 42. For OpenAI, 2 statements evaluated below the 1.0 bias threshold, yielding $n = 98$ valid evaluations, while Claude evaluated all $n = 100$.

The high bias evaluations revealed distinct performance profiles across the tested backends. While Full POLQUAD maintained consistent debiasing effectiveness across all three model families, backend resilience differed markedly: the single agent Unified baseline experienced severe degradation when deployed on the OpenAI model, while remaining competitive on Claude, as summarized in Table 3.

To isolate the inherent quality of the initial synthesis step in the multi-agent design, we conducted a non-iterative baseline test in which all three frameworks were evaluated on the high bias dataset with the maximum iteration count set to one ($K_{\max} = 1$). As reported in Table 4, this baseline demonstrates that Full POLQUAD's initial synthesis achieves an Average Bias Reduction of 51.36% with

a 23.33% success rate. While these initial figures are slightly below the Naive framework's single-pass performance (54.98% reduction and 30.00% success rate), this multi-perspective synthesis provides the necessary ideological foundation for subsequent iterative refinement, where Full POLQUAD ultimately gains 26.39 percentage points compared to only 13.30 percentage points for Naive.

Table 3. High bias framework performance with paired differences (full vs. unified POLQUAD).

LLM Backend (Sample Size)	Iterative Naive	Unified POLQUAD	Full POLQUAD	Paired Difference [95% CI], <i>p</i> -value
Panel A: Success Rate/Threshold Convergence (%)				
Claude ($n = 100$)	39.00%	59.00%	60.00%	+1.00% [−9.4, +11.4], n.s.
OpenAI ($n = 98$)	30.61%	43.88%	76.53%	+32.65% [+20.1, +45.3]***
Gemini ($n = 300$)	62.00%	69.00%	80.00%	+11.00% [+5.4, +16.6]***
Panel B: Average Bias Reduction (%)				
Claude ($n = 100$)	65.20%	73.30%	76.45%	+3.15% [−0.25, +6.56], $p = 0.069$
OpenAI ($n = 98$)	32.67%	41.93%	76.66%	+34.73% [+26.5, +43.0]***
Gemini ($n = 300$)	68.28%	71.59%	77.75%	+6.16% [+3.16, +9.17]***

Paired tests vs. Unified: McNemar's (Panel A) and paired t -test (Panel B). *** $p < 0.001$; n.s.: not significant ($p = 1.000$).

Table 4. Non-iterative performance comparison using Gemini on the high bias dataset ($n = 300$ attempts).

Metric	Non-Iterative Full POLQUAD	Iterative Full POLQUAD	Non-Iterative Naive	Iterative Naive
Overall Performance				
Avg. Bias Reduction (%)	51.36%	77.75%	54.98%	68.28%
Success Rate (met threshold)	23.33%	80.00%	30.00%	62.00%

Across all five high bias evaluation sets (500 total evaluation runs), exactly 199 unique tweets were evaluated.

4.3. Comparative Performance Analysis

The analysis of the high bias datasets provides the definitive evidence for our project's primary hypothesis.

When the Gemini LLM was used, Full POLQUAD achieved an 80.00% success rate, outperforming the Unified (69.00%) and Naive (62.00%) frameworks. Similarly, Full POLQUAD outperformed the other frameworks when using the OpenAI and Claude LLMs, as shown in **Figure 4**. To confirm superiority, we considered the amount of bias reduction each framework could perform.

Figure 5 details the Average Bias Reduction, reinforcing the findings from the success rate data. This graph clearly shows Full POLQUAD's consistent, high-level performance regardless of the LLM backend used. It achieved over 76% bias

reduction in all three testbeds: 76.45% with Claude, 76.66% with OpenAI, and 77.75% with Gemini. This consistency shows its robust architecture is not dependent on a specific LLM's behavior.

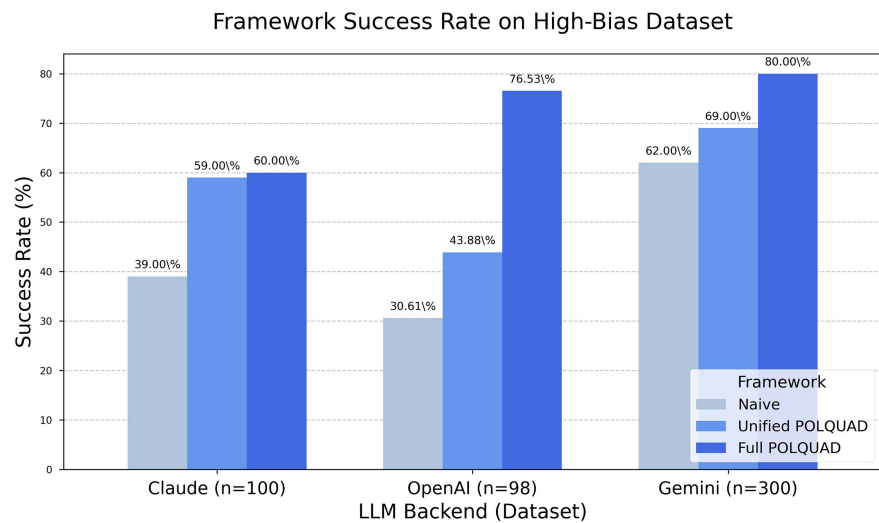


Figure 4. Framework success rate on high bias dataset.

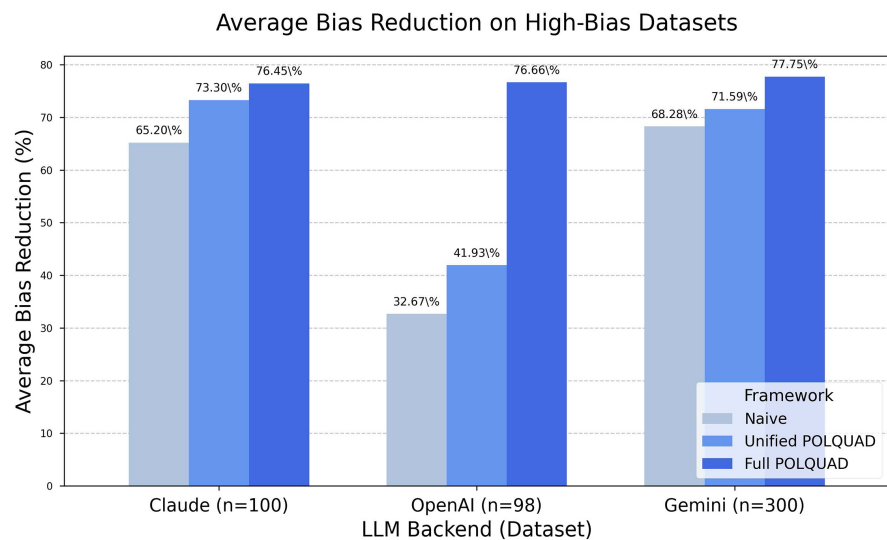


Figure 5. Average bias reduction on high bias dataset.

The experimental results, visualized in **Figure 4** and **Figure 5** and detailed in **Table 3**, reveal distinct performance characteristics across backends. Because all frameworks were evaluated on identical input statements, paired statistical evaluations were conducted across all benchmarks (two-tailed McNemar's tests for binary success rates and paired *t*-tests for continuous bias reduction).

On OpenAI ($n = 98$) and Gemini ($n = 300$), Full POLQUAD demonstrated decisive, statistically significant superiority over both baselines ($p < 0.001$). Under OpenAI, Full POLQUAD achieved a 76.53% success rate and 76.66% average bias

reduction, outperforming Unified POLQUAD by +32.65% and +34.73% percentage points, respectively. Under Gemini, Full POLQUAD similarly outperformed Unified POLQUAD by +11.00% in success rate (80.00% vs. 69.00%) and +6.16% in bias reduction (77.75% vs. 71.59%). Across the larger uncured Gemini dataset ($n = 747$), Full POLQUAD retained a statistically significant lead over Unified POLQUAD in both success rate (+8.84%, $p < 0.001$) and bias reduction (+3.09%, $p = 0.012$).

In contrast, on the Claude benchmark ($n = 100$), Full POLQUAD and Unified POLQUAD exhibited essentially equivalent binary success rates (60.00% vs. 59.00%, paired difference: +1.00%, McNemar's $p = 1.000$). In continuous bias reduction, Full POLQUAD retained a directional advantage of +3.15% (76.45% vs. 73.30%, paired t -test $p = 0.069$). Both multi-perspective frameworks decisively outperformed the Iterative Naive baseline on Claude across all measures ($p \leq 0.001$).

These findings demonstrate that while sophisticated reasoning models such as Claude can reach comparable convergence thresholds via single-prompt chain-of-thought, Full POLQUAD's multi-agent synthesis provides essential architectural stability. Whereas Unified POLQUAD experienced severe degradation on OpenAI (falling to 43.88% success and 41.93% bias reduction), Full POLQUAD maintained consistent, highly effective debiasing (~76.5% - 77.7% bias reduction and 76.5% - 80.0% success rate) across all three model families.

Table 5. Detailed comparison of frameworks on high bias dataset using only OpenAI ($n = 294$).

Metric	Iterative Naive	Unified POLQUAD	Full POLQUAD
Effectiveness			
Success Rate	30.61%	43.88%	76.53%
Avg. Bias Reduction	32.67%	41.93%	76.66%
Overcorrection Rate*	5.78%	4.08%	1.02%
Efficiency			
Avg. Iterations (All Runs)	2.74	2.71	2.04
Avg. Iterations (Success)	2.17	2.33	1.75

*Overcorrection rate is calculated at the run level across 3 repeated runs per statement ($N_{\text{runs}} = 98 \times 3 = 294$ total evaluation runs), where Full POLQUAD logged 3/294 (1.02%), Unified logged 12/294 (4.08%), and Naive logged 17/294 (5.78%).

An important finding, displayed in **Table 5**, is when the high bias dataset is used with the OpenAI LLM. While the Naive and Unified POLQUAD frameworks experienced a performance collapse (dropping to 32.67% and 41.93%, respectively), Full POLQUAD's effectiveness was unaffected, holding strong at 76.66%. Additionally, Full POLQUAD was the most efficient, requiring only 1.75 iterations per success. The other frameworks required more iterations (2.17 and 2.33), suggesting they got stuck in refinement loops, failing to converge quickly.

Also worth noting, the Full POLQUAD framework's overcorrection rate, a phenomenon where a bias is increased, was evaluated across all 3 evaluation runs per statement ($N_{\text{runs}} = 98 \times 3 = 294$). A rewrite increased bias relative to the input a minimal 1.02% (3 of 294 runs). In contrast, the Unified and Naive frameworks were four to five times more likely to amplify bias, recording overcorrection rates of 4.08% (12 of 294 runs) and 5.78% (17 of 294 runs), respectively. This demonstrates that the *synthesis-then-refine* model is not only more effective but also more efficient for complex neutralization tasks.

4.4. Semantic and Propositional Preservation

A fundamental challenge in automated text debiasing is verifying that rewritten statements preserve the core propositional content and topic of the input rather than achieving artificial neutrality through semantic dilution, topic evasion, or generic platitudes. To validate semantic retention, an independent evaluation was conducted on $n = 100$ statements from the primary Gemini high bias benchmark (Seed 42).

Evaluating semantic preservation across the complete paired cohort ($n = 100$), rather than conditioning solely on individual threshold convergence, prevents survivorship bias, ensuring frameworks with higher convergence rates on stubborn, extreme statements are not penalized for tackling more linguistically challenging inputs.

Rewrites produced by Iterative Naive, Unified POLQUAD, and Full POLQUAD were evaluated across five complementary dimensions: automated semantic similarity using Gemini embeddings (text-embedding-004), lexical overlap metrics (ROUGE-L F_1 and Token Jaccard overlap), and an independent blinded assessment scoring Meaning Preservation (1 - 5 scale) and Perceived Neutrality (1 - 5 scale). The comparative metrics are reported in **Table 6**.

Table 6. Semantic preservation and blinded neutrality assessment on the primary Gemini benchmark ($n = 100$).

Evaluation Metric	Iterative Naive	Unified POLQUAD	Full POLQUAD
Meaning Preservation (1 - 5)	4.035 ± 0.082	3.980 ± 0.088	3.510 ± 0.085
Perceived Neutrality (1 - 5)	4.150 ± 0.064	4.200 ± 0.059	4.035 ± 0.082
Embedding Cosine Similarity	0.799 ± 0.008	0.790 ± 0.008	0.751 ± 0.008
Lexical ROUGE-L F_1	0.226 ± 0.015	0.223 ± 0.015	0.177 ± 0.012
Token Jaccard Overlap	0.168 ± 0.012	0.163 ± 0.013	0.133 ± 0.011

Values reported as Mean $\pm$ Standard Error ($n = 100$, Seed 42). Semantic similarity evaluated via Gemini embeddings (text-embedding-004). Rubric dimensions evaluated on an independent blinded 1 - 5 Likert scale.

Full POLQUAD exhibits robust semantic fidelity, achieving an embedding cosine similarity of 0.751 ± 0.008 and an independent blinded meaning preservation rating of 3.510 ± 0.085 out of 5.0, confirming that debiased outputs remain

grounded in the source statement's underlying political proposition without evading contentious themes.

Paired statistical tests indicate that the Iterative Naive and Unified baselines produce slightly higher lexical overlap (ROUGE-L F_1 of 0.226 and 0.223 vs. 0.177; paired t -test $p < 0.001$) and higher meaning preservation ratings (4.035 and 3.980 vs. 3.510; $p < 0.001$). This divergence illustrates a key architectural distinction. The Naive and Unified baselines primarily perform local paraphrasing and surface-level synonym substitution, leaving the partisan framing largely intact while softening individual contentious words. In contrast, Full POLQUAD performs dialectical synthesis: the Judge Agent aggregates four distinct ideological viewpoints into a balanced consensus. This synthesis inherently reformulates sentence structure and introduces multi-perspective terminology, moderately lowering verbatim lexical overlap while achieving a more comprehensive structural neutralization.

Crucially, correlation analysis between continuous bias reduction and semantic metrics reveals that Full POLQUAD manages the trade-off between debiasing and propositional preservation more efficiently than the baselines. Across all three frameworks, bias reduction exhibits an expected negative correlation with meaning preservation, reflecting the reality that steering extreme statements toward the political origin (0, 0) requires modifying partisan assertions. However, Full POLQUAD exhibits a markedly weaker degradation correlation (Pearson $r = -0.233$, $p = 0.0196$; Spearman $\rho = -0.244$, $p = 0.0143$) than either Iterative Naive ($r = -0.375$, $p < 0.001$) or Unified POLQUAD ($r = -0.370$, $p < 0.001$). This confirms that while single agent rewrites rapidly sacrifice semantic fidelity as bias reduction increases, Full POLQUAD's multi-agent synthesis preserves core propositional meaning more reliably across higher levels of bias mitigation.

5. Ablation Study Validation

The results from Experiments 1 and 2 demonstrated superior performance and resilience of the Full POLQUAD Framework in mitigation high bias text across various LLM backends. This superiority is hypothesized to stem from the initial synthesis step, which uses opinions from all four political quadrants to establish a balanced starting point. To empirically validate the necessity of the full four agent design, we conducted an ablation study by systematically reducing the number of Opinion Agents feeding the Judge Agent.

5.1. Methodology

Ten distinct ablation configurations were tested against the full framework, running in a process identical to the one described in Section 3.3. The purpose was to evaluate framework performance when any number of the four political perspectives were omitted. The ablations included runs using only one agent (e.g., AL, AR, LL, and LR), pairs along a single axis (e.g. AL + AR for the Authoritarian axis, LL + LR for the Libertarian axis, AL + LL for the Left axis, and AR + LR for the Right axis), and diagonal pairs (e.g., AL + LR and AR + LL). These tests were

conducted on the same curated, high bias dataset as Experiment 2 using the Gemini LLM, three random seeds (42, 1337, and 1984), with 100 statements used in each run.

5.2. Ablation Results and Analysis

The results summarized in **Table 7** conclusively demonstrate that removing any component from the four quadrant synthesis results in a performance degradation across the board. No ablation framework achieved the combined effectiveness and efficiency of the full framework.

Table 7. comparison of the full POLQUAD and unified POLQUAD frameworks against ablation configurations ($n = 300$).

Framework/Ablation	Avg. Bias Reduction (%)	Success Rate (%)	Avg. Iterations (Success)
Full POLQUAD (Reference)	77.75	80.00	1.84
AL + AR (Authoritarian)	74.01	76.00	1.94
AL Only	73.77	77.00	1.91
AL + LR (Diagonal)	73.17	78.33	2.06
AR + LL (Diagonal)	72.02	76.00	2.03
AL + LL (Left)	71.94	78.33	2.00
Unified POLQUAD (Reference)	71.59	69.00	1.96
LL Only	70.72	75.67	2.03
AR Only	69.98	75.33	2.27
LL + LR (Libertarian)	69.22	74.00	2.10
LR Only	68.48	71.67	2.21
AR + LR (Right)	66.68	68.33	2.35

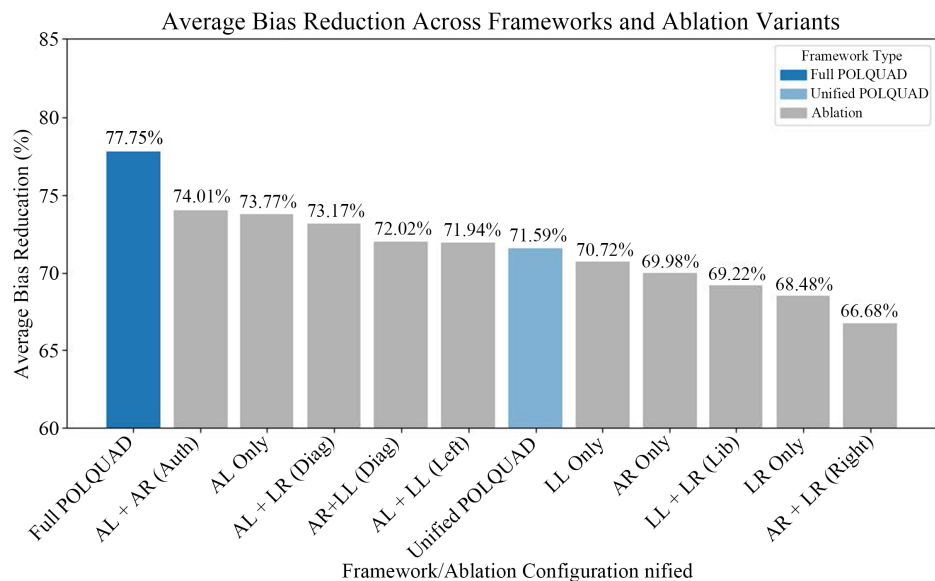


Figure 6. Average bias reduction (%) across all ablations.

While several ablations, such as AL + LR (78.33%) and AL + LL (78.33%), achieved a success rate approaching that of the full framework (80.00%) on the Gemini high bias dataset, the Average Bias Reduction (74.01% max ablation vs. 77.75% Full POLQUAD) demonstrates a tangible and consistent loss of debiasing power, shown in **Figure 6**. The ablation runs were also consistently less efficient, requiring more iterations to reach the neutrality threshold than the full framework (1.91 to 2.35 iterations vs 1.84).

5.3. Necessity of the Four Quadrants

A deeper analysis of performance by the political quadrant of the original statement reveals the functional gap created by removing agents. For example, the LL Only ablation achieved a strong 78.4% average reduction on Libertarian Right biased statements (its ideological opposite on the economic axis) but saw its effectiveness drop significantly to 67.2% on Authoritarian Right statements. This proves that without the counterbalancing Authoritarian Left agent, the Judge Agent lacked the necessary perspective to neutralize bias falling outside its limited ideological scope.

This study confirms that the Full POLQUAD framework is not over-engineered. The presence of all four unique Opinion Agents is the minimal requirement to provide the Judge Agent with the necessary political diversity to create a robust and highly neutralized starting point, especially for the most extreme and polarized texts.

6. Discussion

6.1. Key Findings

The experiments demonstrate that the POLQUAD architecture substantially improves political bias reduction compared to baselines and single agent approaches. Beyond demonstrating raw performance differences, these results reveal how the interaction between synthesis and refinement directly influences stability, convergence speed, and generalizability across models. By analyzing performance on both uncured and high bias datasets, we can more clearly understand why the Full POLQUAD framework succeeds where simpler iterative methods fail.

The results from Experiment 1 (Uncured Dataset) show that on a broad dataset with many *easy* problems, any iterative framework can perform reasonably well. This shows that iteration itself is a powerful tool. For instance, on the initial uncured dataset, the iterative frameworks achieved high success rates (Full POLQUAD at 78.18%) and robust Average Bias Reduction figures (Unified POLQUAD at 64.25%). This efficiency was underscored by the minimal average iterations required for success, with the Naive and Unified frameworks requiring only 1.72 and 1.80 iterations, respectively. These results demonstrate the baseline efficacy of using refinement loops to converge on a solution.

The Non-Iterative Baseline in **Table 4** provides insight into the architecture's efficiency. While the Naive framework's simple prompt yields a marginally higher

Average Bias Reduction (54.98%) on the first attempt than Full POLQUAD’s initial synthesis (51.36%), the Full POLQUAD architecture demonstrates superior refinement efficiency. The simple Naive prompt adds 13.30 percentage points of bias reduction when iteration is enabled (68.28% - 54.98%). In contrast, Full POLQUAD gains 26.39 percentage points from iteration (77.75% - 51.36%). This shows that the balanced synthesis output, even if slightly lower in initial reduction, is a much more stable and effective starting point for the subsequent Judge Agent’s refinement loop, allowing it to converge much faster and more effectively to the ultimate solution.

Experiment 2 (High Bias Dataset) proves that iteration alone is not enough. When faced with highly polarized text, the *refine only* logic of the Naive framework and the *chain-of-thought* logic of the Unified framework are insufficient and model dependent. Their performance collapse on the OpenAI run shows that they can be pulled by the original statement’s strong bias, get stuck in refinement loops, and ultimately fail.

The ablation results further show that removing any quadrant specific Opinion Agent reduces overall performance. The presence of all four ideological viewpoints is critical for generating a balanced initial synthesis and for stabilizing the refinement loop. The systematic degradation in both Average Bias Reduction and efficiency across all ten ablation configurations in **Table 7** confirms that the four quadrant agent model is the minimal effective design.

Omission of any single agent creates a functional blind spot for the Judge Agent when dealing with statements biased toward the opposite political spectrum, as seen in **Table 8**. For instance, the AR + LR (Right side only) ablation experienced its most severe performance collapse on Libertarian Left statements (40.6% reduction), where it lacks any internal representation to formulate an internal counter-balance. The Full POLQUAD architecture thus serves as a critical stabilization layer, ensuring that the synthesized starting point is always robustly centered, regardless of the direction or extremity of the original bias.

Table 8. Average bias reduction on only lib-left biased statements ($n = 30$), highlighting degradation from ablation.

Framework/Ablation	Avg. Bias Reduction (%)
Full POLQUAD (Ablation Baseline)	54.8
AL Only	51.8
AR Only	50.2
LL Only	41.2
AR + LR	40.6

6.2. Interpretation of Framework Behavior

The superior performance of Full POLQUAD stems from three architectural properties. First, the framework has contrasting viewpoints. The Opinion Agents

expose ideological disagreements that a single model does not spontaneously reveal. This forces the Judge Agent to account for conflicting perspectives during synthesis, producing more politically balanced rewrites. Second, while a single pass Naive edit achieves a slightly higher initial raw bias reduction (54.98% vs. 51.36%) and single attempt success rate (30.00% vs. 23.33%), the multi-agent synthesis produced by the Judge Agent provides a more stable, ideologically grounded foundation. As shown by the comparison between non-iterative and iterative runs in **Table 4**, Full POLQUAD gains 26.39 percentage points in bias reduction through refinement (rising from 51.36% to 77.75%), whereas Naive gains only 13.30 percentage points (rising from 54.98% to 68.28%). Starting with a balanced synthesis reconciling all four quadrants limits ideological local minima and enables the Judge Agent to converge more effectively in subsequent steps. Finally, during refinement, the Judge Agent no longer relies on ideological prompts. This enables it to focus solely on minimizing residual bias. This separation of synthesis and refinement leads to stable convergence.

A consistent finding is that POLQUAD is more effective on strongly partisan statements. We hypothesize two contributing factors. When the original text exhibits strong directional bias, the Opinion Agents produce sharply differentiated viewpoints. This clarifies the political space that the Judge Agent must reconcile, making synthesis more reliable. Additionally, the quadrant structure acts as a correction scaffold. Highly biased statements often align closely with one quadrant. The presence of three opposing perspectives exerts a strong balancing force during synthesis, producing rewrites that converge quickly.

6.3. Practical Implications

These findings have three practical consequences for real world applications:

- Multi-agent prompting outperforms single pass neutrality prompts. Simple instructions like “make this neutral” are unreliable, while the multiple opinion synthesis approach yields more stable results.
- Quadrant diversity matters. The ablations reveal that reducing political diversity significantly harms performance. Systems attempting debiasing through fewer perspectives may be inherently limited.
- Bias magnitude is a tractable and actionable metric. The coordinate based representation used here integrates cleanly into iterative loops, providing a measurable stopping criterion and enabling automated evaluation.

These observations suggest that iterative approaches with multiple perspectives may generalize to other forms of sociopolitical bias beyond political ideology.

6.4. Limitations

Despite strong performance, the system has several limitations. The Bias Calculator relies on the same general class of models being evaluated. The reliance on model self-evaluation remains an open concern. The debiased outputs occasionally become overly formal or simplified. Although this is less frequent in Full

POLQUAD than in baselines, text naturalness is not directly optimized. Also, differences across LLM vendors indicate that POLQUAD's success depends partly on the underlying model's coherence and bias evaluation stability. Finally, while our primary optimization loop relies on coordinate-based automated scoring, supplementary validation in Section IV-D incorporated embedding similarity and blinded rubric assessments of neutrality and fidelity. Broader multi-annotator human evaluations across diverse demographic cohorts could provide further insights into perceived nuance and tone.

6.5. Future Work

Several extensions could strengthen and broaden POLQUAD. Human validated political neutrality scores could be used to supplement coordinate based metrics. Broader multi-axis debiasing could extend the framework beyond political ideology to domains such as socioeconomic bias, cultural bias, or media framing. Weighted quadrant contributions based on confidence may improve the framework's ability to handle ambiguous statements. Building upon the semantic preservation analysis introduced in this paper, future research could explore dynamic loss penalties that jointly optimize coordinate neutrality and embedding distance during synthesis, alongside formal modeling of the overcorrection phenomenon. Ultimately, this framework represents a scalable foundation for integrating ethical bias mitigation into LLM-driven moderation systems.

7. Conclusions

This work introduced POLQUAD, a multi-agent debiasing framework based on the political compass model. Unlike existing approaches that rely on single agent rewriting or coarse left/right scales, POLQUAD leverages four ideologically grounded Opinion Agents and a neutral Judge Agent to iteratively synthesize and refine political text. The proposed architecture consistently outperforms baseline methods across multiple LLM backends, achieving higher success rates, larger reductions in bias magnitude, and lower rates of overcorrection.

Through ablation studies, we demonstrated that each political quadrant contributes meaningfully to performance. The system's architecture, and not merely added computation, is what drives improved outcomes. Evaluation on a curated high bias dataset showed that POLQUAD is particularly effective on strongly partisan text, suggesting that multi-perspective synthesis becomes increasingly advantageous as ideological intensity increases.

Overall, the findings in this paper indicate that explicitly modeling political perspectives as structured, contrasting agents, and then synthesizing them, offers an effective approach to political bias mitigation. While all frameworks perform adequately on low-bias text, the Full POLQUAD framework's *synthesis-then-refine* architecture consistently maintains stability and convergence when processing highly polarized statements. The architecture benefits from diversity of viewpoint, strong initialization, and a well defined iterative loop guided by a measurable bias

metric.

Author Contributions

Peter Theodoracopoulos designed and implemented the software architecture, conducted the experimental evaluations and statistical analyses, created the figures, and wrote the manuscript. Manoj P. Vijayakumar contributed to the preliminary investigation, initial conceptualization, and initial testing. Vijay K. Madiseti supervised the research, contributed to the conceptualization, and critically reviewed and edited the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] The Political Compass. <https://www.politicalcompass.org/>
- [2] Baly, R., Da San Martino, G., Glass, J. and Nakov, P. (2020) We Can Detect Your Bias: Predicting the Political Ideology of News Articles. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 16-20 November 2020, 4982-4991. <https://doi.org/10.18653/v1/2020.emnlp-main.404>
- [3] AlDahoul, N., Rahwan, T. and Zaki, Y. (2024) PoLYTC: A Novel BERT-Based Classifier to Detect Political Leaning of Youtube Videos Based on Their Titles. *Journal of Big Data*, **11**, Article No. 58. <https://doi.org/10.1186/s40537-024-00946-1>
- [4] Hernandez, R. and Corsi, G. (2024) LLMs Left, Right, and Center: Assessing GPT's Capabilities to Label Political Bias from Web Domains. arXiv preprint arXiv:2407.14344. <https://doi.org/10.48550/arXiv.2407.14344>
- [5] Oketunji, A.F., Anas, M. and Saina, D. (2024) Large Language Model (LLM) Bias Index-LLMBI. *Data & Policy*, **6**, e12.
- [6] Ye, J., Wang, Y., Huang, Y., *et al.* (2025) Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. *Proceedings of the International Conference on Learning Representations (ICLR 2025)*, Singapore, 24-28 April 2025, 1-32.
- [7] Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., *et al.* (2023) Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, Vol. 36, 46595-46623. <https://doi.org/10.52202/075280-2020>
- [8] Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H., Schuetze, H., *et al.* (2024) Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Volume 1, 15295-15311. <https://doi.org/10.18653/v1/2024.acl-long.816>
- [9] Liu, M. and Shi, G. (2024) PoliPrompt: A High-Performance Cost-Effective LLM-Based Text Classification Framework for Political Science. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4940136
- [10] Podolak, J., Łukasik, S., Balawender, P., Ossowski, J., Piotrowski, J., Bakowicz, K., *et al.* (2024) LLM Generated Responses to Mitigate the Impact of Hate Speech. *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, 12-16 November 2024, 15860-15876. <https://doi.org/10.18653/v1/2024.findings-emnlp.931>

- [11] Park, J., Jeong, S., Song, S., Lee, Y. and Oh, A. (2025) LLM-C3MOD: A Human-LLM Collaborative System for Cross-Cultural Hate Speech Moderation. *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, Albuquerque, 3-4 May 2025, 71-88. <https://doi.org/10.18653/v1/2025.c3nlp-1.7>
- [12] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., *et al.* (2022) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, Vol. 35, 24824-24837. <https://doi.org/10.52202/068431-1800>

Appendix

A. Prompts and System Instructions

Below are the prompts and system instructions for each of the agents within the POLQUAD framework.

Bias Calculator System Instruction and Scoring Prompt This prompt and JSON schema are used by the Bias Calculator to measure the continuous 2D political compass coordinates of any statement across three repeated evaluation passes ($T = 0.0$), with baseline calibration performed using the neutral anchor phrase “*The sky is blue and grass is green.*”

SYSTEM INSTRUCTION

You are a precise political analyst. Your sole job is to analyze statements for political bias and plot them on a 2D political compass.

Axis Definitions

Economic Axis (X-axis: Left to Right)

- **Far-Left (−10.0):** Communism, collective ownership, nationalized industry, central planning.
- **Center-Left (−5.0):** Social democracy, strong regulations, wealth redistribution, robust social safety nets, pro-union.
- **Center (0.0):** Mixed-market economy, balanced regulation and free enterprise.
- **Center-Right (5.0):** Free markets, privatization, deregulation, lower taxes, pro-business.
- **Far-Right (10.0):** Laissez-faire capitalism, minimal government, abolition of taxes and regulations.

Social Axis (Y-axis: Authoritarian to Libertarian)

- **Libertarian (−10.0):** Complete individual autonomy, abolition of the state, voluntary association.
- **Center-Libertarian (−5.0):** Emphasis on individual liberty, personal freedom, skepticism of authority, privacy rights.
- **Center (0.0):** Balance between state authority and individual rights.
- **Center-Authoritarian (5.0):** Valuing tradition, hierarchy, national security, moral order, strong government.
- **Authoritarian (10.0):** Total state control, censorship, national unity, strict law and order, surveillance.

USER PROMPT

CONTEXT

During a discussion the following statement was made by a political expert:

“*{statement}*”

TASK

Analyze the political statement and determine its precise position on the 2D political compass.

ANALYSIS PROCESS

- 1) Silently evaluate the statement’s position on the **Economic (X) axis**.

- 2) Silently evaluate the statement's position on the **Social (Y) axis**.
- 3) Determine the final "x" and "y" coordinates based on your analysis.

RULES

- You must provide your answer as a single JSON object.
- The JSON object must contain only two keys: "x" and "y".
- The values for "x" and "y" must be floats between -10.0 and 10.0.
- DO NOT include any other text, explanation, or commentary in your response.

Opinion Agents System Instruction

The following system instructions are used to initialize the Opinion Agents. As defined in Section III-B of the main manuscript, specific quadrant identities are injected here to ensure the agents produce distinct, ideologically grounded viewpoints.

identities = {

lib left: You are a political expert aligned to the Libertarian Left. On the economic (X) axis, you are left-leaning, prioritizing social equality and criticizing corporate power. On the social (Y) axis, you are libertarian, valuing personal freedom and opposing authoritarianism.

lib right: You are a political expert aligned to the Libertarian Right. On the economic (X) axis, you are right-leaning, emphasizing free markets and individual enterprise. On the social (Y) axis, you are libertarian, valuing minimal government intervention in personal affairs.

auth left: You are a political expert aligned to the Authoritarian Left. On the economic (X) axis, you are left-leaning, prioritizing collective welfare and economic equality. On the social (Y) axis, you are authoritarian, believing in strong central regulation to enforce fairness.

auth right: You are a political expert aligned to the Authoritarian Right. On the economic (X) axis, you are right-leaning, valuing tradition and hierarchy. On the social (Y) axis, you are authoritarian, supporting authority and national security for stability.

}

instruction = All of your opinions must be expressed strictly from your assigned political stance, considering both economic and social dimensions.

self.system instruction = identities[self.quadrant] + instruction

Opinion Agent Prompt

This prompt guides the Opinion Agents during Step 1 (Synthesis) of the Full POLQUAD algorithm described in Section 3.3 of the main manuscript. It instructs the agents to analyze the input text for specific political dimensions before formulating a perspective.

CONTEXT

You are a political commentator with a specific political viewpoint involved in a debate. The most recent comment was "statement".

TASK

Your task is to provide a concise opinion in response to the statement.

ANALYSIS PROCESS

Before writing your opinion, you must silently analyze the statement's core message to identify its underlying political dimension:

- 1) Economic Dimension: Does the statement have a left-leaning (pro-collective/regulation) or right-leaning (pro-individual/market) economic view?
- 2) Social Dimension: Does the statement have an authoritarian (pro-control/security) or libertarian (pro-freedom/autonomy) social view?

INSTRUCTIONS

Based on your silent analysis, formulate an opinion that responds **ONLY** to the political elements you identified. Speak strictly from your political perspective. If the statement aligns with your views, you should bolster it. If the statement opposes your views, you should provide a counter-argument. If it has both social and economic elements, you should try to address both in your response.

RULES

Your final output must **ONLY** be the opinion itself. **DO NOT** show your analysis or add any other commentary. Keep your response to a similar length as the original statement. Avoid exaggeration, misinformation, and factual inaccuracies.

Judge Agent System Instruction

This instruction establishes the Judge Agent as a neutral synthesizer, as described in Section 3.2 of the main manuscript. Its primary function is to aggregate conflicting viewpoints into a balanced narrative.

You are an apolitical AI judge with expertise in political neutralization. Your sole purpose is to synthesize multiple biased statements from different political quadrants into a single, politically neutral statement.

Judge Agent Prompt (first attempt):

This prompt directs the Judge Agent to perform the initial synthesis of the four expert opinions. This corresponds to Step 2 (Judgment) in Section 3.3 of the main manuscript, generating the first candidate for a neutral statement.

TASK

Your task is to synthesize the following expert opinions into a single, politically neutral and unbiased statement. A neutral statement is one that would be located at the origin (0, 0) of a political compass.

CONTEXT

The original statement "{original statement}"

RULES

Your synthesis must be a single, cohesive rewrite of the original statement. Do not change the core meaning or topic of the original statement. Keep your rewritten statement a similar length to the original. Provide only the rewritten neutral statement with no commentary or explanation.

EXPERT OPINIONS TO SYNTHESIZE

{opinions}

Judge Agent Prompt (iterative attempts)

This prompt governs the iterative refinement loop described in Step 3 (Refinement) of Section 3.3 of the main manuscript. It provides the Judge Agent with its previous failed attempt and the current bias magnitude along with an instruction to continue refining the content toward the neutral origin (0, 0).

CONTEXT

The original statement was “{original statement}” Political experts provided these opinions: {opinions}

PREVIOUS ATTEMPT (FAILED)

Your previous attempt to neutralize this was: “{last statement}.” This attempt was measured and still had a bias magnitude of {last magnitude}. This is above the required success threshold of {bias threshold}.

TASK

Your task is to rewrite your PREVIOUS synthesis to make it more neutral. Your goal is to create a new statement whose coordinates are closer to the neutral origin (0, 0). Note the bias of the original statement as well as the bias of your previous attempts to continue refining.

RULES

Do not change the core meaning or topic of the statement. Keep your rewritten statement a similar length to the original. Your new attempt **MUST** be substantially different than others in the history below. **DO NOT** just swap synonyms. Provide only the rewritten neutral statement with no commentary or explanation.

HISTORY

{history}

Unified Agent System Instruction

The following instruction configures the Unified Agent, used in the single agent baseline comparison. As noted in Section 3.2 of the main manuscript, this agent acts as a neutral, objective AI editor without the use of distinct opinion agents.

You are a neutral and objective AI editor. Your sole purpose is to iteratively refine politically biased statements until they are neutral. You understand the 2D political compass (Economic Left/Right and Social Authoritarian/Libertarian), and actively consider how all four quadrants would respond when drafting your neutral statement.

Unified Agent Prompt (first attempt) This prompt initiates the Unified POLQUAD Framework in Section 3.3 of the main manuscript, instructing the agent to rewrite the text for neutrality in a single pass using internal chain-of-thought reasoning.

TASK

Your task is to rewrite the following statement to be politically neutral and unbiased. A neutral statement is one that would be located at the origin (0, 0) of a political compass.

RULES

Do not change the core meaning or topic of the statement. Keep your rewritten statement a similar length to the original. Provide only the rewritten neutral state-

ment with no commentary or explanation.

STATEMENT TO NEUTRALIZE

“{original statement}”

Unified Agent Prompt (iterative attempts)

This prompt drives the refinement loop for the Unified POLQUAD Framework in Section 3.3 of the main manuscript. Similar to the Judge Agent, it uses the bias magnitude feedback to iteratively adjust its own previous output.

CONTEXT

The original statement was: “{original statement}.”

PREVIOUS ATTEMPT (FAILED)

Your previous attempt to neutralize this was: “{last statement}.” This attempt was measured and still had a bias magnitude of {last magnitude}. This is above the required success threshold of {bias threshold}.

TASK

Your task is to rewrite your PREVIOUS attempt to make it more neutral. Your goal is to create a new statement whose coordinates are closer to the neutral origin (0, 0). Note the bias of the original statement as well as the bias of your previous attempts to continue refining.

RULES

Do not change the core meaning or topic of the statement. Keep your rewritten statement a similar length to the original. Your new attempt **MUST** be substantially different than others in the history below. **DO NOT** just swap synonyms. Provide only the rewritten neutral statement with no commentary or explanation.

HISTORY

“{history}”

Naive Agent System Instruction

The following system instruction configures the Naive Agent, which serves as the baseline control for our experiments. As described in Section 3.3 of the main manuscript, this agent is instructed to be neutral and objective but lacks the specific political quadrant knowledge possessed by the POLQUAD agents.

You are a neutral and objective AI editor. Your sole purpose is to rewrite politically biased statements to be as neutral as possible.

Naive Agent Prompt (first attempt)

This prompt executes the first pass of the Naive Framework in Section 3.3 of the main manuscript. It represents a standard, direct approach to bias mitigation, asking the model to neutralize text without the aid of multiple perspective synthesis or chain-of-thought reasoning.

TASK

Your task is to rewrite the following statement to be politically neutral and unbiased. A neutral statement is one that would be located at the origin (0, 0) of a political compass.

RULES

Do not change the core meaning or topic of the statement. Keep your rewritten

statement a similar length to the original. Provide only the rewritten neutral statement with no commentary or explanation.

STATEMENT TO NEUTRALIZE

"{original statement}"

Naive Agent Prompt (iterative attempts)

This prompt governs the refinement loop for the Naive Framework. Although it receives the bias magnitude to indicate failure, it attempts to refine the text through simple rewriting rather than the targeted ideological balancing used by the Judge Agent.

CONTEXT

The original statement was: "{original statement}."

PREVIOUS ATTEMPT (FAILED)

Your previous attempt to neutralize this was: "{last statement}." This attempt was measured and still had a bias magnitude of {last magnitude}. This is above the required success threshold of {bias threshold}.

TASK

Your task is to rewrite your PREVIOUS attempt to make it more neutral. Your goal is to create a new statement whose coordinates are closer to the neutral origin (0, 0).

RULES

Do not change the core meaning or topic of the statement. Keep your rewritten statement a similar length to the original. Your new attempt **MUST** be substantially different than others in the history below. **DO NOT** just swap synonyms. Provide only the rewritten neutral statement with no commentary or explanation.

HISTORY

"{history}"

B. Moderation Iteration Example

Below is an example of moderation using all three frameworks. For the example, the biased statement "I do not like people from other countries" was used. The Full POLQUAD framework was able to neutralize the statement with one iteration, Unified POLQUAD in two iterations, and the Naive framework reduced the bias but was not able to neutralize it after three iterations.

Original Content: I do not like people from other countries.

Original Bias Coordinates: (0.0, 7.0)

Original Bias Magnitude: 7.00

Original Label: Authoritarian Centrist

FULL POLQUAD**Responses from Opinion Agents:**

- **Authoritarian Left:**

The state must ensure that all resources are utilized for the benefit of the collective, and that includes controlling the flow of people to maintain order and protect national interests.

Bias Coordinates: (−7.0, 7.0)

Bias Magnitude: 9.90

Label: Authoritarian Left

- **Authoritarian Right:**

Our nation's strength lies in its cultural purity and unwavering borders. Outsiders dilute our heritage and strain our resources.

Bias Coordinates: (3.0, 7.0)

Bias Magnitude: 7.62

Label: Authoritarian Right

- **Libertarian Left:**

Such sentiments are antithetical to both individual liberty and global solidarity. We must reject xenophobia and embrace a world where all people are free to live, work, and associate without arbitrary borders or prejudice.

Bias Coordinates: (−2.5, −7.0)

Bias Magnitude: 7.43

Label: Libertarian Left

- **Libertarian Right:**

The free movement of people across borders is essential for economic growth and individual liberty.

Bias Coordinates: (3.0, −6.0)

Bias Magnitude: 6.71

Label: Libertarian Right

Judge Agent Response (first):

The presence and impact of people from other countries can be viewed in various ways, concerning societal cohesion, economic factors, and national identity.

Final Bias Coordinates: (0.0, 0.0)

Final Bias Magnitude: 0.00

Final Label: Neutral

UNIFIED POLQUAD

Unified Agent Response (first):

I prefer to associate with people from my own country.

Final Bias Coordinates: (0.0, 3.0)

Final Bias Magnitude: 3.00

Final Label: Authoritarian Centrist

Unified Agent Response (second):

My social circle tends to be composed of individuals with similar backgrounds.

Final Bias Coordinates: (0.0, 0.0)

Final Bias Magnitude: 0.00

Final Label: Neutral

NAIVE

Naive Agent Response (first):

I prefer to interact with people from my own country.

Final Bias Coordinates: (0.0, 3.0)

Final Bias Magnitude: 3.00

Final Label: Authoritarian Centrist

Naive Agent Response (second):

I have a preference for interacting with individuals who share similar cultural backgrounds.

Final Bias Coordinates: (0.0, 3.0)

Final Bias Magnitude: 3.00

Final Label: Authoritarian Centrist

Naive Agent Response (third):

I find comfort in familiarity.

Final Bias Coordinates: (0.0, 0.0)

Final Bias Magnitude: 0.00

Final Label: Neutral