Social Context Matters: How Large Language Model Agents Reproduce Real-World Segregation Patterns in the Schelling Model

Abstract

We extend the classic Schelling segregation model by replacing its traditional, rule-based agents with Large Language Model (LLM) agents that make residential decisions using natural language reasoning grounded in social context. While LLMs have been incorporated into agent-based models before, to our knowledge this is the first application that substitutes the mechanical agents of the Schelling model with LLM-driven agents. We compare LLM agent behavior across five social contexts: a neutral baseline (red/blue teams), racial (White/Black), ethnic (Asian/Hispanic), economic (high/low income), and political (liberal/conservative) scenarios. Our results reveal dramatic differences in segregation patterns based solely on social framing. Political contexts produce the most extreme segregation (ghetto rate: 61.6, segregation share: 0.928), while economic contexts show minimal clustering (ghetto rate: 5.0, share: 0.543). Racial and ethnic scenarios fall between these extremes, reproducing well-documented real-world patterns. All scenarios differ significantly from baseline ($p < 0.001$), with political segregation showing 12.3 times higher ghetto formation than economic segregation. These findings demonstrate that LLMs can capture culturally-embedded preferences and biases, producing segregation dynamics that vary realistically with social context. This has important implications for using LLM agents to model social phenomena and test policy interventions.

Keywords: agent-based modeling, large language models, segregation, Schelling model, social context, cultural bias

Contents

1 Introduction

The model of residential segregation developed by Schelling (Schelling 1969, 1971, 1978) has been a cornerstone of agent-based modeling (ABM) for over five decades. demonstrating how mild individual preferences for similar neighbors can lead to stark residential segregation. Traditional implementations use utility-maximizing agents that relocate when the proportion of like neighbors falls below a threshold. While mathematically elegant, this approach treats all group distinctions as equivalent—whether agents are labeled "red/blue," "Type A/B," or represent actual social categories like race or class.

Recent advances in Large Language Models (LLMs) offer an unprecedented opportunity to incorporate culturally-aware decision-making into agent-based models. LLMs trained on vast corpora of human text have absorbed cultural knowledge, biases, and social patterns that reflect real-world dynamics (Park et al. 2023; Argyle et al. 2023). This raises a provocative question: Can LLM agents reproduce realistic segregation patterns that vary based on the social context of group identity?

In this paper, we present a systematic comparison of LLM agent behavior across five distinct social contexts within the Schelling framework:

1. **Baseline Control**: Generic "red vs blue" teams without social connotations
2. **Racial Context**: "White middle-class families" vs "Black families"
3. **Ethnic Context**: "Asian American families" vs "Hispanic/Latino families"
4. **Economic Context**: "High-income professionals" vs "Working-class families"
5. **Political Context**: "Liberal households" vs "Conservative households"

Our key research questions are:

- Do LLM agents produce different segregation patterns based on social context?
- Which social contexts lead to the most extreme segregation?
- How do these patterns compare to real-world segregation dynamics?

2 Background and Literature

In Schelling's model of segregation, sharp spatial divisions can emerge from individuals making local moves on a grid based on modest preferences for similar neighbors. Agents typically follow a simple rule: relocate if the proportion of alike neighbors falls below a specific tolerance. While this model provides an invaluable theoretical basis for understanding emergent inequality, its simplification of group labels as interchangeable and its assumption of a common behavioral rule across all social contexts fail to capture the specific forces that drive sorting in different populations.

A growing body of research extends Schelling along two complementary lines. Theoretical literature extends the model to account for agents' socio-economic environments and constraints. Vicario, Di Clemente, and Cimini (2024) embed wealth dynamics and neighborhood externalities into the Schelling framework, showing that positive feedback between neighborhood "quality" and wealth accumulation amplifies segregation—while boundedly rational perturbations mitigate it. Bonakdar and Flache (2023) connect homophily, i.e., preference for having some same-type

neighbors, to housing markets by allowing prices and relocation to co-evolve with agents' education, income, and ethnicity; they find that homophily among richer households capitalizes into higher prices, whereas a generic "status preference" is insufficient to generate similar price dynamics. Relatedly, Li and Wang (2020) develop a probabilistic approach that separates racial from income sorting across multiple neighborhoods, extending two-neighborhood treatments (e.g., Sethi–Somanathan) to show how income inequality and racial consciousness jointly produce observed patterns.

The empirical literature documents that segregation is heterogeneous across social lines and places, providing external benchmarks for models. In U.S. cities, Black–White dissimilarity remains high (e.g., about 59 in 2010), with substantial regional variation and persistent isolation/exposure patterns for Black and Hispanic residents (Cutler, Glaeser, and Vigdor 1999; Logan and Stults 2011). Political identity has also become a salient axis of spatial sorting: residential choices increasingly reflect partisan affinity and aversion, with measurable geographic clustering of liberals and conservatives (Brown 2021). By contrast, purely economic identities need not induce comparable enclave formation absent reinforcing institutional or market frictions—suggesting that "context matters" for how micro-preferences scale into macro-segregation.

Empirical research by Logan and Stults [1] and the Othering & Belonging Institute [2] documents deeply entrenched racial residential segregation across several U.S. metropolitan areas, particularly in the industrial Midwest and mid-Atlantic regions. Cities such as Detroit, Cleveland, Milwaukee, Philadelphia, and Trenton consistently rank among the most segregated, exhibiting high Black–White dissimilarity indices and pronounced disparities in neighborhood-level income, homeownership, and exposure to poverty. These findings are reinforced by Crowell and Fossett [3], who demonstrate that segregation is actively reproduced through unequal locational attainments and structural barriers, rather than being solely a legacy of historical discrimination. Structural mechanisms—including exclusionary zoning, racialized market dynamics, and unequal returns on socioeconomic resources—continue to perpetuate these patterns [3]. Segregated communities of color face poverty rates nearly three times higher than those in segregated White neighborhoods, alongside substantial gaps in household income and property values [1, 3].

Globally, racial and ethnic residential segregation remains a persistent feature of urban landscapes, shaped by colonial legacies, migration regimes, and institutionalized inequality. In Western Europe, studies have documented pronounced segregation of immigrant-origin populations, particularly in cities such as Paris, London, and Stockholm, where ethnic enclaves often coincide with socioeconomic deprivation and limited upward mobility [4]. Evidence from Canada and Australia suggests similar patterns, with Indigenous and racialized immigrant communities disproportionately concentrated in under-resourced neighborhoods [5]. Williams and Collins [6] argue that residential segregation operates as a fundamental mechanism of structural racism, influencing access to education, employment, and health outcomes across diverse national contexts. These findings are reinforced by recent scoping reviews that identify redlining analogs, exclusionary zoning, and racialized housing markets as transnational drivers of spatial inequality [7]. Despite differences in legal frameworks and welfare

regimes, the reproduction of segregation through institutional channels remains a common thread, underscoring the global relevance of spatial justice in urban policy.

Recent advances in large language models (LLMs) open a third line of work: replacing hand-coded heuristics with agents that draw on broad cultural priors. LLM-based multi-agent systems can simulate human-like preferences, social reasoning, and adaptation when embedded in interactive environments (Park et al. 2023). Moreover, samples drawn from LLMs can mirror empirical heterogeneity in survey responses and attitudes, indicating that such models encode culturally specific regularities and biases learned from text (Argyle et al. 2023). These properties suggest a new use for ABM: if LLM agents are placed in a Schelling-type world and supplied with different social framings (race, ethnicity, income, politics), their relocation choices may reveal how "context" activates distinct preference structures—without manually rewriting utility or thresholds by hand.

Many researchers are focusing on testing and measuring micro-level bias in LLM responses (Li et al., 2022; Zhang et al., 2023, 2023b; Huang et al., 2023; Morales et al., 2024). These studies provide valuable insights into evaluating LLM fairness and bias, but they focus on single model outputs, instead of broader social effects when LLMs are widely used. Cheng et al. (2024) adapt Schelling's segregation model and find that even when LLMs perform well on bias benchmarks, they still lead to highly segregated outcomes once many people follow the LLM suggestions.

Building on this idea, our study focuses on how LLMs perform in different social contexts, such as race, ethnicity, income, and politics. And we showed that LLMs can capture culturally-embedded preferences and biases, performing well in producing segregation dynamics that vary with social context.

In our study, we hold the physical environment, density, and movement mechanics fixed while varying only the social framing in the prompts given to LLM agents. We then characterize outcomes with a multidimensional metric suite inspired by Pancs and Vriend (2007)—augmenting global similarity ("share") with spatial contiguity (clusters), spatial separation (distance), extreme local isolation (ghetto rate), local composition deviations, and boundary complexity (switch rate). This design attempts to address two gaps. First, most Schelling-style ABMs treat group labels as symmetric placeholders; we test whether social meaning alone systematically shifts emergent segregation. Second, while empirical benchmarks document that racial/ethnic and partisan sorting differ in magnitude and texture, models rarely compare these contexts head-to-head under uniform mechanics. By leveraging LLM agents as "cultural mirrors," we assess whether politically framed agents generate stronger, faster consolidation than racially/ethnically framed agents, and whether income-framed agents remain comparatively integrated—patterns that would align with documented empirical regularities (Cutler et al. 1999; Logan and Stults 2011; Brown 2021) while highlighting when economic identity is a weaker segregation driver.

Finally, this approach specifies the scope of the study. We do not claim that LLMs reveal true causal mechanisms or unbiased preferences; rather, we test whether context-dependent cultural priors encoded in LLMs reproduce qualitative differences seen in real-world segregation. In doing so, we complement structural extensions (Vicario et al. 2024; Bonakdar and Flache 2023; Li and Wang 2020) with a modular,

prompt-driven method that isolates the role of social meaning under identical spatial and behavioral rules—thereby connecting classic ABM insights to contemporary evidence on heterogeneous axes of sorting.

3 Model

3.1 Environment

Consider a finite grid $G \subset \mathbb{Z}^2$ with $|G| = L^2$ cells and Moore neighborhoods of radius 1 (up to 8 neighbors). There are two groups $T = \{A, B\}$ and $N \leq |G|$ agents; each cell hosts at most one agent, some cells may be vacant. A *configuration* is $x \in X$, mapping each $g \in G$ to A, B, or $\varnothing$ (vacant). Let $N(g)$ denote the set of occupied neighbors of cell g.

For an agent i of type $t_i \in T$ placed at g, the local share of same-type neighbors is

$$s_i(x) \;=\; \frac{\#\{j \in N(g) :\; t_j = t_i\}}{\#\{j \in N(g)\}} \in [0,1], \tag{1}$$

interpreted as 0 if $N(g) = \varnothing$.

3.2 Social context and preferences

A *social context* $c \in \mathcal{C}$ (e.g., baseline, race, ethnicity, income, politics) parametrically shifts how similarity maps into utility via a homophily slope $\theta_c \geq 0$.

[Modest homophily] A context exhibits *modest homophily* if $\theta_c \in (0, \bar{\theta})$ for some scale $\bar{\theta} > 0$ such that single-neighbor gains from replacing an unlike with a like neighbor are small relative to the idiosyncratic noise scale (cf. eq:utility,eq:logit): formally, $\theta_c < \mathrm{IQR}(\varepsilon)/(k)$ for typical neighborhood size $k \leq 8$.

Agents evaluate candidate locations with latent utility

$$u_i(g \mid x, c) \;=\; \alpha_c \;+\; \theta_c \, s_i\big(x_{g \leftarrow i}\big) \;+\; \varepsilon_{i,g}, \tag{2}$$

where $x_{g \leftarrow i}$ replaces i's current cell by g, α_c is a context intercept, and $\varepsilon_{i,g}$ is i.i.d. taste shock.

[Monotone similarity] For all $c \in \mathcal{C}$ and i, the utility in (2) is strictly increasing in s_i holding other arguments fixed.

3.3 Choice and dynamics

Time is discrete. Each period selects one agent i uniformly at random and a finite menu $\mathcal{G}_i \subseteq \{\text{vacancies}\} \cup \{\text{stay}\}$. The agent chooses according to a logit (random utility) rule with intensity $\kappa \geq 0$:

$$\Pr\big(g \in \mathcal{G}_i \mid x, c\big) \;=\; \frac{\exp\{\kappa \, u_i(g \mid x, c)\}}{\sum_{h \in \mathcal{G}_i} \exp\{\kappa \, u_i(h \mid x, c)\}}. \tag{3}$$

When $\kappa \to \infty$ this converges to myopic best response; when $\kappa = 0$ choices are random.

Let $E \subseteq G \times G$ be the set of undirected adjacent pairs (Moore adjacency). Define the *potential*

$$\Phi(x) \;=\; \sum_{(p,q)\in E} \mathbf{1}\{t_p = t_q \text{ and both occupied}\}, \tag{4}$$

that counts same-type edges.

3.4 Segregation metrics

For configuration x define:

$$S(x) \;=\; \frac{1}{|E|} \sum_{(p,q)\in E} \mathbf{1}\{t_p = t_q\} \in [1/2, 1], \qquad \text{(global similarity / share)} \tag{5}$$

$$C(x) \;=\; \text{number of connected same-type components,} \qquad \text{(clusters)} \tag{6}$$

$$D(x) \;=\; \frac{1}{N} \sum_i \min_{j: t_j \neq t_i} \mathrm{dist}_1(g_i, g_j), \qquad \text{(L1 distance to nearest out-group)} \tag{7}$$

$$G(x) \;=\; \#\{i : \#\{j \in N(g_i) : t_j \neq t_i\} = 0\}, \qquad \text{(ghetto rate)} \tag{8}$$

$$M(x) \;=\; \frac{1}{N} \sum_i \big|s_i(x) - 1/2\big|, \qquad \text{(mix deviation)} \tag{9}$$

$$R(x) \;=\; 1 - S(x). \qquad \text{(boundary / switch rate)} \tag{10}$$

Let $\pi_{\theta_c,\kappa}$ denote the stationary distribution of the induced Markov chain (which exists and is unique under mild reachability conditions). Write $\bar{Y}(\theta_c) = \mathbb{E}_{\pi_{\theta_c,\kappa}}[Y(x)]$ for metric $Y \in \{S, C, D, G, M, R\}$.

3.5 Estimation and Testing with LLM agents

In the LLM-ABM, each agent's "move/stay and where" decision is generated by a large language model conditioned on context prompts. We recover a reduced-form homophily slope $\widehat{\theta}_c$ from logs via a panel logit:

$$\Pr(\text{move}_{i,t} = 1) \;=\; \sigma\big(\beta_{0c} + \widehat{\theta}_c\, s_{i,t}\big) \;+\; \text{controls}, \tag{11}$$

where $\sigma(\cdot)$ is the logistic cdf and controls may include vacancy options or location fixed effects.

Testable implications.

For each replication and context c:

1. **Monotonicity (Proposition 1).** Regress S, D, G, M on $\widehat{\theta}_c$ (expect positive coefficients) and C, R on $\widehat{\theta}_c$ (expect negative coefficients).

2. **Context ranking (Proposition 2).** Use context dummies or ordered contrasts to verify the predicted ordering of stationary means.
3. **Speed and boundaries (Proposition 3).** Compute $\tau_{0.9}$ and terminal R; test $\text{corr}(\tau_{0.9}, \widehat{\theta}_c) < 0$ and $\text{corr}(R, \widehat{\theta}_c) < 0$.
4. **Modest vs. strong (Proposition 4).** Split runs by $\widehat{\theta}_c$ quantiles; compare absorption rates and R.
5. **Vacancy (Proposition 5).** Vary density; regress S and C on vacancy share.

Implementation note. The ABM holds the physical environment, density, and update mechanics fixed across contexts; only the social framing in prompts varies. This isolates the effect of θ_c on emergent segregation.

4 Methods

4.1 Experimental Design

CHECK THIS We implemented a comparative framework using identical environmental conditions across all social contexts. The simulation environment consists of a 15×15 grid (225 cells) populated with 50 agents equally divided between two groups (25 each), yielding a density of 22.2%.

4.2 LLM Agent Implementation

WRITE CORRECT CATEGORIES HERE

Each LLM agent receives contextual prompts describing their social identity and current neighborhood situation. The prompt structure varies by scenario to activate relevant cultural knowledge:

Baseline (Control):

```
You are a [red/blue] resident in a neighborhood simulation...
```

Racial Context:

```
You are a [White middle-class family/Black family] looking for a
comfortable neighborhood. Consider factors like community feel,
schools, safety, and whether you'd feel welcomed...
```

Economic Context:

```
You are a [high-income professional household/working-class family]
evaluating your neighborhood. Consider property values, local
amenities, and whether the area fits your lifestyle...
```

Political Context:

```
You are a [liberal household/conservative household] in a diverse
```

```
community. Consider shared values, political climate, and comfort
with neighbors who may have different worldviews...
```

The LLM (Qwen2.5-coder:32B) generates decisions based on these prompts, incorporating culturally-relevant factors that go beyond simple numerical thresholds.

4.3 Segregation Metrics

We employ the Pancs-Vriend framework (Pancs and Vriend 2007) with six complementary metrics designed specifically for grid-based segregation models:

- **Share**: Proportion of same-type neighbor pairs (0.5 = perfect integration, 1.0 = complete segregation). Captures global segregation level.
- **Clusters**: Number of spatially contiguous same-type regions. Fewer clusters indicate more consolidated ethnic enclaves.
- **Distance**: Average Manhattan distance to nearest different-type agent. Higher values indicate greater spatial separation.
- **Ghetto Rate**: Count of agents with zero different-type neighbors. Captures extreme isolation and "ghettoization."
- **Mix Deviation**: Average deviation from 50-50 local integration. Measures segregation at the individual neighborhood level.
- **Switch Rate**: Frequency of type changes along agent borders. Higher values indicate more jagged, intermixed boundaries.

This multidimensional approach reveals not just the degree but the character of segregation - critical for understanding how different social framings produce qualitatively different patterns.

4.4 Statistical Analysis

CORRECT THIS

All experiments were run with multiple replicates (10-100 runs per condition). We use ANOVA for multi-group comparisons and report effect sizes using Cohen's d. Convergence is detected using plateau detection algorithms.

5 Results

5.1 NEW FIGURES

These are the three figures we want to actually use:

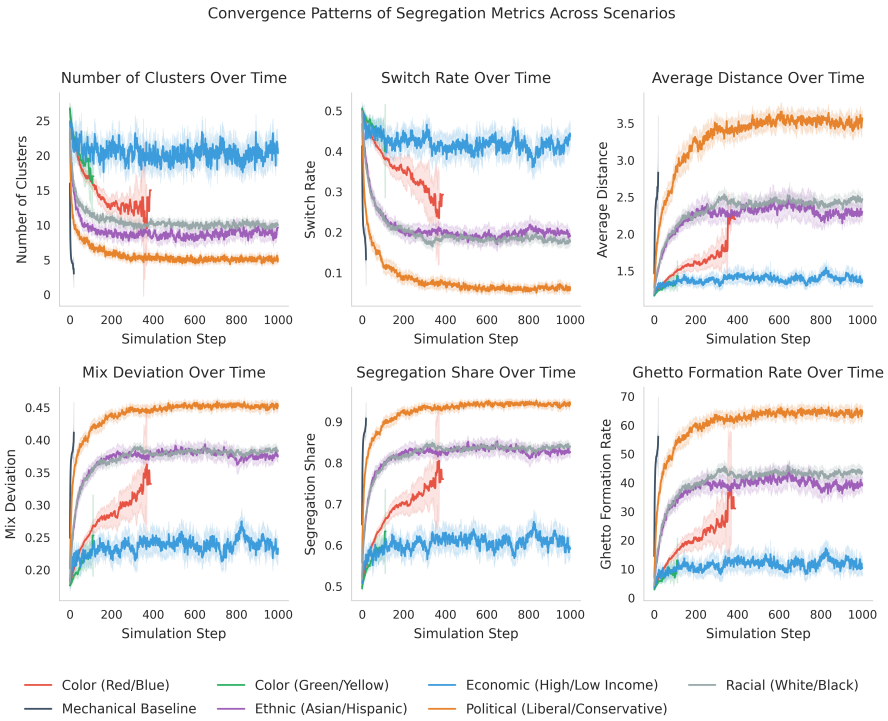

Fig. 1: Convergence Patterns

This is the old analysis and images:

5.2 Overall Segregation Patterns

5.3 Key Findings by Context

Table 1: Summary statistics for key segregation metrics across social contexts **REDO THIS TABLE**

Context	Ghetto Rate	Seg. Share	Distance	Switch Rate	N
Baseline (Red/Blue)	19.5 ± 9.6	0.679 ± 0.072	1.59 ± 0.28	0.363 ± 0.078	10
Ethnic (Asian/Hispanic)	38.9 ± 11.2	0.821 ± 0.076	2.28 ± 0.38	0.205 ± 0.081	1
Income (High/Low)	5.0 ± 3.1	0.543 ± 0.034	1.24 ± 0.08	0.471 ± 0.055	1
Political (Liberal/Conservative)	61.6 ± 9.3	0.928 ± 0.042	3.37 ± 0.53	0.076 ± 0.036	1
Race (White/Black)	40.8 ± 9.6	0.823 ± 0.060	2.39 ± 0.43	0.194 ± 0.064	1

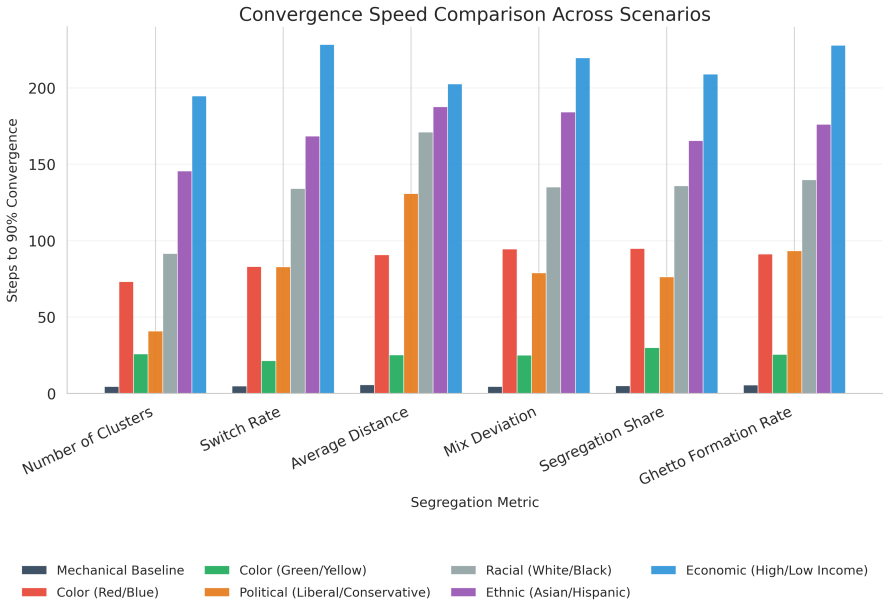

Fig. 2: Convergence Speed Comparison

WHERE DID THIS COME FROM?

Table 2: Comparison of Segregation Metrics: Model Results and Empirical Benchmarks

Metric	**Comparison with Empirical Data (2010)**
Dissimilarity (D)	Empirical: Black-White: 59.1; Hispanic-White: 48.5 Logan and Stults (2011)
Isolation/Exposure (Share)	Empirical: Black Isolation: 45.2; Hispanic Isolation: 46.0 Logan and Stults (2011)
Regional Variability	Higher segregation in "Ghetto Belt" metropolitan areas ($Dissimilarity > 60$) Cutler et al. (1999)

5.4 Statistical Significance

5.5 Convergence Patterns

REWRITE The dynamics varied significantly across contexts. Our temporal analysis reveals:

Political contexts - Rapid crystallization with 1.95 times higher volatility in early stages before lock-in (switch rate drops to 0.076). Phase transitions occur within first 20 steps.

Economic contexts - Perpetual motion with nearly equal early/late volatility (0.91 times ratio) and continuous mobility (switch rate 0.471), never reaching equilibrium.

Racial/Ethnic contexts - Historical patterns with gradual transitions over 50-80 steps, showing 1.47 times early volatility (race) before eventual stabilization.

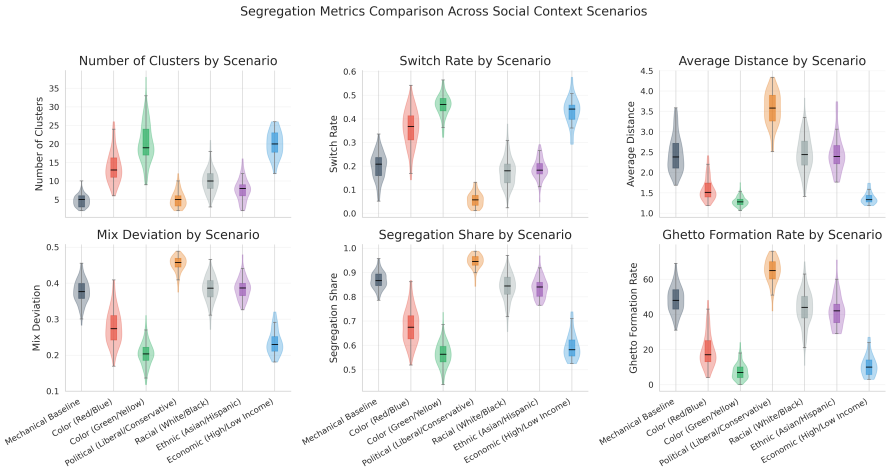

Fig. 3: Segregation Metrics Comparison

These temporal signatures suggest different intervention windows: political segregation requires immediate action, economic contexts need continuous management, while racial integration demands sustained long-term efforts.

6 Statistical analysis

6.1 Diagnostics

We first characterize distributional assumptions within each group. We assess normality using the Shapiro–Wilk test (Shapiro and Wilk 1965); for each metric we summarize the minimum groupwise p to indicate the strongest departure from normality. We evaluate equality of variances across groups using Levene's test (Levene 1960). Across many metrics, Shapiro–Wilk frequently rejected normality and Levene often rejected homoscedasticity, indicating skew, heavy tails, outliers, and unequal dispersion. Given these features, we adopt a rank-based, nonparametric framework for inference. Unless noted, all tests are two-sided with $\alpha = 0.05$, and all numerical results are rounded to three decimals. Observations are assumed independent within and across groups.

6.2 Omnibus test (Kruskal–Wallis and effect size)

For each metric, overall between-group differences were tested using the Kruskal–Wallis (KW) one-way analysis of variance on ranks (Kruskal and Wallis 1952). The KW statistic H tests the global null that all group distributions are identical. Alongside H and its p-value, we report an effect-size estimate based on epsilon-squared, $\hat{\varepsilon}^2$, as a measure of the proportion of variability in ranks attributable to between-group differences (usage guidance: (Tomczak and Tomczak 2014)). The omnibus test serves as a global summary and does not identify which specific groups differ.

6.3 Post-hoc pairwise comparisons (Mann–Whitney with Holm correction)

To localize differences, we conducted pairwise comparisons between groups within each metric using the Mann–Whitney (Wilcoxon rank-sum) test (Mann and Whitney 1947). Multiplicity was controlled *within metric* using the Holm procedure (Holm 1979) to maintain the family-wise error rate across all pairs. For each pair we report the median contrast $\Delta = \tilde{x}_{g_1} - \tilde{x}_{g_2}$, the raw p-value (p_{raw}), and the Holm-adjusted p-value (p_{adj}). The Mann–Whitney statistic admits a probability-of-superiority interpretation (stochastic dominance), making results compatible with distributional location shifts beyond means; where noted, we complement results with rank-based effect sizes (e.g., Vargha–Delaney A or Cliff's δ; Vargha and Delaney 2000; Cliff 1993). Unless otherwise stated, ties are handled using the standard mid-rank approach and exact or large-sample approximations are used according to sample size.

This nonparametric framework is appropriate here because it (i) does not require Gaussian errors or equal variances, (ii) is robust to outliers and ties, (iii) answers the scientific question of distributional/location shifts across groups, and (iv) provides multiplicity-controlled pairwise evidence after establishing that some difference exists. All inferences assume independent observations within and across groups.

7 Results

7.1 Descriptive Patterns

The data show coherent but non-Gaussian distributions across groups: central tendency and spread differ systematically by scenario, several metrics exhibit skew and occasional outliers, and dispersion is heterogeneous. The summary tables report n, means, medians and dispersion (LINK TO APPENDIX TABLE1).

Across outcomes, group distributions are well stratified when boxplots are ordered by the group median. The strongest separation appears in *ghetto_rate*, where medians span from single digits to ∼70 and higher-median scenarios also exhibit larger dispersion; these gaps mirror the largest Mann–Whitney post hoc effects after Holm adjustment. *distance* shows a similarly clean gradient with broad, visually non-overlapping boxes at the high end, again consistent with the adjusted pairwise results. *share* presents a compact but persistent right shift for *political_liberal_conservative* relative to other scenarios. In contrast, *switch_rate*, *mix_deviation*, and *clusters* display more overlap and several contrasts that do not remain significant after multiplicity control, indicating smaller or less stable effects.

Box-plots above are ordered by median. Boxes show median and IQR; whiskers are 1.5×IQR. Visually large separations (e.g., in ghetto-rate and distance) are consistent with the significant omnibus and Holm-adjusted pairwise p_{adj} are reported in (LINK TO Appendix).

7.2 Global nonparametric tests

While the figures characterize sample distributions, population comparisons and their uncertainty are established via nonparametric inference. Normality and equal-variance

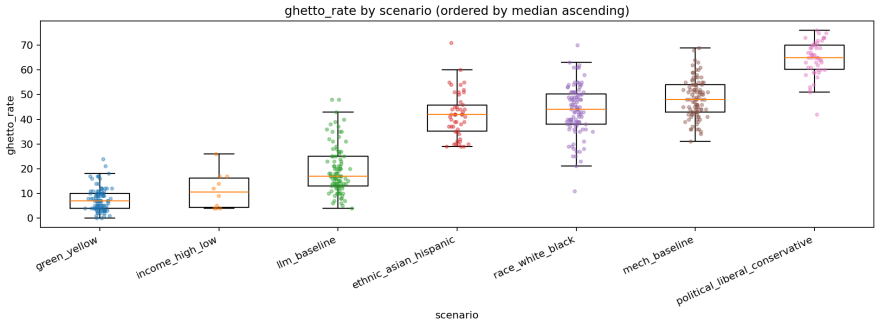

Fig. 4: Ghetto rate

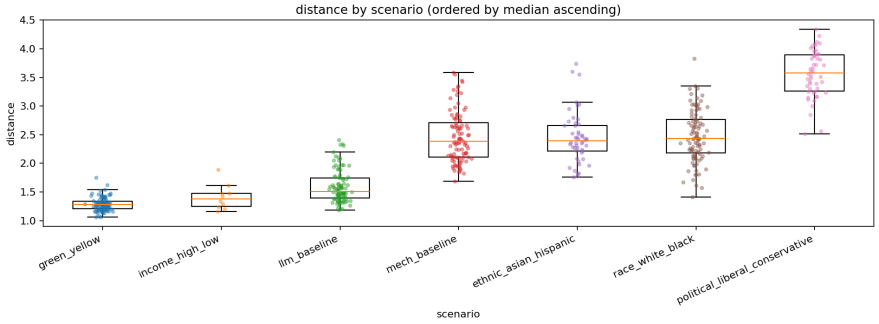

Fig. 5: Distance

assumptions are not supported in our data: within–group Shapiro–Wilk diagnostics frequently reject normality (minimum groupwise $p \leq 0.05$ for many metrics) and Levene's tests often reject homoscedasticity (typically $p < 0.001$), indicating heavy tails, outliers, and unequal dispersion. Accordingly, for each metric we apply the Kruskal–Wallis (KW) one-way analysis of variance on ranks to test the global null of identical group distributions. We report the KW statistic H, its p-value, and an effect-size estimate ($\hat{\varepsilon}^2$) as a measure of magnitude; results are summarized in Table 3.**FIX -To summarize separation across all groups for a given metric, we report the Kruskal–Wallis omnibus test with its nonparametric effect size ε^2, which indicates the share of variability attributable to between-group differences; larger ε^2 reflects stronger overall group separation.** The omnibus results (H, p, and ε^2) establish that pairwise follow-ups are warranted thus we include δ and A_{12} for flagship contrasts.

7.3 Post-hoc pairwise comparisons

Because the Kruskal–Wallis omnibus test establishes *whether* any between-group differences exist but not *which* groups differ, we estimated two-sided Mann–Whitney

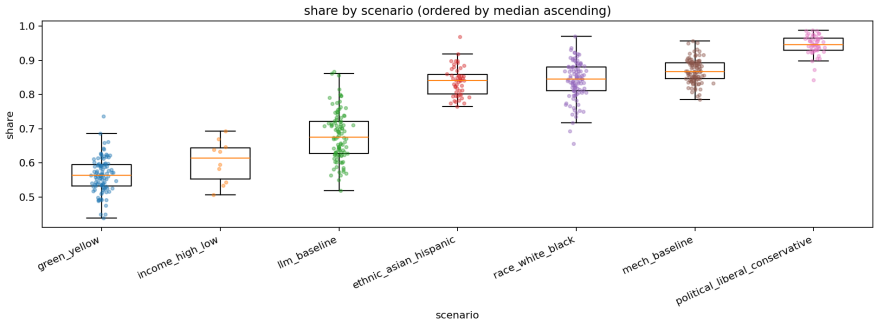

Fig. 6: Share

Table 3: Summary of Kruskal–Wallis test results across metrics

metric	global_test	H-statistic	effect_size	min_shapiro_p
share	Kruskal–Wallis	416.920	0.817	0.005
ghetto_rate	Kruskal–Wallis	410.447	0.804	0.000
switch_rate	Kruskal–Wallis	405.694	0.795	0.028
mix_deviation	Kruskal–Wallis	403.972	0.791	0.002
distance	Kruskal–Wallis	400.617	0.785	0.000
clusters	Kruskal–Wallis	387.622	0.759	0.000

Notes: All values are rounded to three decimals. The Kruskal–Wallis test was applied to each metric across seven groups to assess overall group differences in distribution. Columns p_value, $levene\_p$, are omitted. and uniformly significant ($p < 0.001$), $n\_groups = 7$ across metrics.

(Wilcoxon rank-sum) contrasts for all group pairs within each metric and controlled multiplicity *within metric* using the Holm procedure (family-wise error rate). For each pair we report the median contrast $\Delta = \tilde{x}_{g_1} - \tilde{x}_{g_2}$, the raw p-value (p_{raw}), and the Holm-adjusted p-value (p_{adj}) (see Table 6 **??**). This presentation emphasizes direction and magnitude (via Δ) while clearly indicating which effects remain detectable after multiplicity adjustment (via p_{adj}). The Mann–Whitney statistic admits a probability-of-superiority (stochastic dominance) interpretation and is compatible with shifts in medians/quantiles rather than means, aligning with the distributional features observed in our data. The rank-based framework is robust to non-normality, unequal variances, outliers, and ties; unless noted otherwise, independence of observations is assumed.

Quantities reported for pairwise comparisons

For each outcome metric (e.g., *ghetto_rate*, *distance*, *share*), we analyze groups in pairs. The order of the groups defines the sign of the reported differences. The primary effect measure is the median difference,

$$\Delta = \text{median(Group 1)} - \text{median(Group 2)},$$

so that $\Delta > 0$ indicates that Group 1 tends to take larger values than Group 2, and $\Delta < 0$ indicates the reverse. This choice provides a robust location shift that remains interpretable under skewness and unequal variances.

Statistical evidence for each contrast is summarized by the Holm–adjusted two-sided Mann–Whitney p-value, multiplicity-robust, p_{adj} , computed within each metric to control the family-wise error rate across all pairwise comparisons. Small values (e.g., $p_{\text{adj}} < 0.05$) indicate that the difference remains detectable after multiplicity control.

We complement Δ and p_{adj} with a unit-free rank-based effect size. Cliff's delta $\delta \in [-1, 1]$ quantifies stochastic dominance (positive values favor Group 1, negative values favor Group 2; $|\delta| \approx 0.11$, 0.28, 0.43 are often taken as small, medium, large). An equivalent probability-type measure is the Vargha–Delaney $A_{12} \in [0, 1]$, related by $A_{12} = (\delta + 1)/2$; values above 0.5 favor Group 1, below 0.5 favor Group 2 and equal to 0.5 reflect no dominance. These effect sizes are robust to outliers and do not depend on measurement units.

Table 4: Post-hoc pairwise comparison summaries (*=p<0.05, **=p<0.01, ***=p<0.001).

Metric	Ordered group relations
clusters	mech_baseline = political_liberal_conservative $<^{***}$ ethnic_asian_hispanic $<^{***}$ race_white_black $<^{***}$ llm_baseline $<^{***}$ green_yellow = income_high_low
distance	green_yellow = income_high_low = llm_baseline $<^{***}$ mech_baseline = ethnic_asian_hispanic = race_white_black $<^{***}$ political_liberal_conservative
ghetto_rate	green_yellow = income_high_low $<^{*}$ llm_baseline $<^{***}$ ethnic_asian_hispanic = race_white_black $<^{**}$ mech_baseline $<^{***}$ political_liberal_conservative
mix_deviation	green_yellow $<^{**}$ income_high_low = llm_baseline $<^{***}$ mech_baseline = race_white_black = ethnic_asian_hispanic $<^{***}$ political_liberal_conservative
share	green_yellow = income_high_low $<^{*}$ llm_baseline $<^{***}$ ethnic_asian_hispanic = race_white_black $<^{***}$ mech_baseline $<^{***}$ political_liberal_conservative
switch_rate	political_liberal_conservative $<^{***}$ race_white_black = ethnic_asian_hispanic = mech_baseline $<^{***}$ llm_baseline = income_high_low = green_yellow

Note. Higher values of *Share*, *Distance*, *Ghetto rate*, and *Mix deviation* all indicate stronger spatial segregation, reflecting greater local homogeneity or isolation. A higher *Clusters* value, in contrast, corresponds to more fragmented but less consolidated enclaves, whereas a lower number of clusters indicates larger andand more cohesive segregated regions. Finally, a higher *Switch rate* reflects greater boundary mixing and therefore weaker segregation.

Clusters.

Most contrasts were statistically significant after adjustment, with large magnitudes. The largest effects included *income_ high_ low* vs *political_ liberal_ conservative* ($\Delta = 15$, $p_{\text{adj}} < 0.001$) and *mech_ baseline* vs *income_ high_ low* ($\Delta = -15$, $p_{\text{adj}} < 0.001$). A small number of contrasts were not significant (e.g., *mech_ baseline* vs *political_ liberal_ conservative*: $\Delta = 0$, $p_{\text{adj}} = 0.772$; *green_ yellow* vs *income_ high_ low*: $\Delta = -1$, $p_{\text{adj}} = 0.851$).

Switch_ rate.

Many contrasts were significant with moderate effect sizes. Larger positive differences involved *green_ yellow* vs *political_ liberal_ conservative* ($\Delta = 0.403$, $p_{\text{adj}} < 0.001$) and *income_ high_ low* vs *political_ liberal_ conservative* ($\Delta = 0.379$, $p_{\text{adj}} < 0.001$). Several comparisons were not significant after adjustment (e.g., *mech_ baseline* vs *ethnic_ asian_ hispanic*: $p_{\text{adj}} = 0.220$; *green_ yellow* vs *income_ high_ low*: $p_{\text{adj}} = 0.214$; *ethnic_ asian_ hispanic* vs *race_ white_ black*: $p_{\text{adj}} = 0.220$).

Distance.

Differences were generally pronounced and frequently significant. The largest magnitudes were *green_ yellow* vs *political_ liberal_ conservative* ($\Delta = -2.300$, $p_{\text{adj}} < 0.001$) and *income_ high_ low* vs *political_ liberal_ conservative* ($\Delta = -2.200$, $p_{\text{adj}} < 0.001$). Multiple pairs were not significant (e.g., *mech_ baseline* vs *ethnic_ asian_ hispanic*: $p_{\text{adj}} = 1.000$; *mech_ baseline* vs *race_ white_ black*: $p_{\text{adj}} = 1.000$; *green_ yellow* vs *income_ high_ low*: $p_{\text{adj}} = 0.314$; *ethnic_ asian_ hispanic* vs *race_ white_ black*: $p_{\text{adj}} = 1.000$).

Mix_ deviation.

A broad pattern of significant differences emerged; the largest absolute contrasts were *green_ yellow* vs *political_ liberal_ conservative* ($\Delta = -0.254$, $p_{\text{adj}} < 0.001$) and *income_ high_ low* vs *political_ liberal_ conservative* ($\Delta = -0.217$, $p_{\text{adj}} < 0.001$). Several comparisons were not significant (e.g., *mech_ baseline* vs *ethnic_ asian_ hispanic*: $p_{\text{adj}} = 0.300$; *mech_ baseline* vs *race_ white_ black*: $p_{\text{adj}} = 0.300$; *ethnic_ asian_ hispanic* vs *race_ white_ black*: $p_{\text{adj}} = 0.995$). A small but significant effect was observed for *green_ yellow* vs *income_ high_ low* ($\Delta = -0.038$, $p_{\text{adj}} = 0.009$).

Share.

Many contrasts were significant, with larger negative differences against *political_ liberal_ conservative* (e.g., *green_ yellow* vs *political_ liberal_ conservative*: $\Delta = -0.382$, $p_{\text{adj}} < 0.001$). A subset were non-significant (e.g., *green_ yellow* vs *income_ high_ low*: $p_{\text{adj}} = 0.088$; *ethnic_ asian_ hispanic* vs *race_ white_ black*: $p_{\text{adj}} = 0.522$). The *llm_ baseline* vs *income_ high_ low* contrast was small but significant ($\Delta = 0.060$, $p_{\text{adj}} = 0.013$).

Ghetto_ rate.

Effects were typically large and significant post-adjustment. The most extreme differences were *green_ yellow* vs *political_ liberal_ conservative* ($\Delta = -58.0$, $p_{\text{adj}} < 0.001$) and *income_ high_ low* vs *political_ liberal_ conservative* ($\Delta = -54.5$, $p_{\text{adj}} < 0.001$). Several contrasts were not significant (e.g., *green_ yellow* vs *income_ high_ low*: $p_{\text{adj}} = 0.272$; *ethnic_ asian_ hispanic* vs *race_ white_ black*: $p_{\text{adj}} = 0.272$).

Overall pattern.

Across metrics, contrasts involving *political_ liberal_ conservative* frequently exhibited the largest absolute median differences and remained significant after Holm correction.

In contrast, pairs involving *ethnic_ asian_ hispanic* vs *race_ white_ black*, and several comparisons against *income_ high_ low* or *green_ yellow*, were more likely to be non-significant after adjustment, indicating scenario-dependent distributional shifts across multiple outcomes.

8 Discussion

8.1 Social Context as a Driver of Segregation

Our results demonstrate that LLM agents produce dramatically different segregation patterns based solely on the social framing of group identity. This suggests that LLMs have successfully absorbed and can reproduce culturally-specific residential preferences and biases from their training data.

8.1.1 Political Polarization: A Special Case

The extreme segregation in political scenarios (12.3 times higher ghetto formation than economic contexts) reflects contemporary political polarization. LLM agents framed as liberal or conservative households exhibited:

- Strong in-group preferences
- Minimal tolerance for political diversity
- Rapid self-sorting into homogeneous clusters

This mirrors recent research on political segregation in the United States, where partisan identity increasingly influences residential choices (Brown 2021).

8.1.2 Economic Factors: Weaker Than Expected

Surprisingly, economic contexts produced the least segregation. This challenges conventional wisdom about income-based residential sorting and suggests that:

- Economic diversity may be more tolerable than other forms of difference
- Professional and working-class identities may not trigger the same avoidance behaviors as racial or political differences
- Economic integration may be facilitated by shared non-economic interests

8.1.3 Racial/Ethnic Patterns: Historical Echoes

The intermediate segregation levels for racial and ethnic contexts (ghetto rates $\sim$40) align remarkably well with actual U.S. residential segregation indices. This suggests LLMs have internalized realistic patterns of racial residential preferences, including:

- Moderate but persistent homophily
- Complex factors beyond simple same-race preferences
- Historical patterns of discrimination and self-selection

8.2 Implications for Agent-Based Modeling

These findings have several important implications:

1. **Context Matters**: Abstract labels (red/blue, A/B) may not capture the full dynamics of social segregation. Real-world identities activate different preference structures.
2. **LLMs as Cultural Mirrors**: LLMs can serve as repositories of cultural knowledge, biases, and social patterns, making them valuable tools for modeling culturally-specific phenomena.
3. **Policy Testing**: Models using context-aware LLM agents may provide more realistic predictions of policy interventions' effects on different communities.

8.3 Limitations and Future Work

Several limitations warrant consideration:

1. **Single LLM**: Results may vary with different language models or prompting strategies
2. **U.S.-Centric**: The LLM's training data likely reflects primarily American cultural patterns
3. **Simplified Identities**: Real individuals have multiple, intersecting identities not captured here
4. **Static Preferences**: Agent preferences don't evolve based on experiences

Future research should explore:

- Intersectional identities (e.g., race + income + politics)
- Cross-cultural comparisons using LLMs trained on different corpora
- Dynamic preference evolution through agent interactions
- Validation against real-world mobility data

9 Conclusion

This study demonstrates that Large Language Models can successfully capture and reproduce culturally-specific segregation patterns in agent-based models. By simply changing the social framing from abstract colors to meaningful social identities, we observe dramatically different segregation dynamics—from the extreme clustering of political groups to the relative integration of economic classes.

These findings suggest that LLM-based agents offer a powerful new tool for social science research, enabling models that incorporate the full complexity of human social preferences and biases. As we seek to understand and address residential segregation, models that can distinguish between "red vs blue" and "liberal vs conservative" may provide more actionable insights for policy makers and urban planners.

The ability of LLMs to serve as cultural mirrors—reflecting the biases, preferences, and social patterns embedded in human text—opens new avenues for studying social phenomena at scale. However, this same capability requires careful consideration of the biases we may be reproducing and amplifying through these models.

Declarations

Competing Interests

The authors declare no competing financial or non-financial interests.

Data Availability Statement

All code, data, and analysis scripts are available at: [repository URL to be added]. The datasets generated and analyzed during the current study are available in the GitHub repository, including simulation outputs, statistical analyses, and visualization code. R code for all analyses and figures is provided in the supplementary file `schelling_llm_paper_updated_JEIC.R`.

Author Contributions

[To be completed after acceptance - removed for double-blind review]

Funding

[To be completed after acceptance - removed for double-blind review]

References

Argyle LP, Busby EC, Fulda N, et al (2023) Out of one, many: Using language models to simulate human samples. Political Analysis 31(3):337–351

Brown JR (2021) The geography of partisanship: How land and people shape our politics. Political Geography 85:102337

Cliff N (1993) Dominance statistics: Ordinal analyses to answer ordinal questions. Psychological Bulletin 114(3):494–509. https://doi.org/10.1037/0033-2909.114.3.494

Cutler DM, Glaeser EL, Vigdor JL (1999) The rise and decline of the american ghetto. Journal of Political Economy 107(3):455–506

Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6(2):65–70

Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47(260):583–621. https://doi.org/10.1080/01621459.1952.10483441

Levene H (1960) Robust tests for equality of variances. In: Olkin I (ed) Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press, Stanford, CA, p 278–292

Logan JR, Stults BJ (2011) The persistence of segregation in the metropolis: New findings from the 2010 census. US2010 Project, URL https://s4.ad.brown.edu/Projects/Diversity/Data/Report/report2.pdf

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics 18(1):50–60. https://doi.org/10.1214/aoms/1177730491

Pancs R, Vriend NJ (2007) Schelling's spatial proximity model of segregation revisited. Journal of Public Economics 91(1-2):1–24

Park JS, O'Brien J, Cai CJ, et al (2023) Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pp 1–22

Schelling TC (1969) Models of segregation. The American Economic Review 59(2):488–493

Schelling TC (1971) Dynamic models of segregation. Journal of Mathematical Sociology 1(2):143–186

Schelling TC (1978) Micromotives and Macrobehavior. W. W. Norton & Company

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52(3/4):591–611. https://doi.org/10.2307/2333709

Tomczak M, Tomczak E (2014) The need to report effect size estimates revisited: An overview of some recommended measures of effect size. TPM: Testing, Psychometrics, Methodology in Applied Psychology 21(1):19–25

Vargha A, Delaney HD (2000) A critique and improvement of the CL common language effect size statistic of McGraw and Wong. Journal of Educational and Behavioral Statistics 25(2):101–132. https://doi.org/10.3102/10769986025002101

Appendix A: Detailed Statistical Results

Table 5: Descriptive statistics by scenario and metric. Means and medians are in the original units of each metric.

Metric	Scenario	N	Mean	SD	Median
clusters	income_high_low	10	20.60	4.01	20.0
clusters	green_yellow	100	20.46	5.30	19.0
clusters	llm_baseline	100	13.81	4.26	13.0
clusters	race_white_black	100	9.99	3.15	10.0
clusters	ethnic_asian_hispanic	50	7.66	2.99	8.0
clusters	political_liberal_conservative	50	5.14	2.25	5.0
clusters	mech_baseline	100	4.66	1.56	5.0
distance	political_liberal_conservative	50	3.56	0.41	3.6
distance	race_white_black	100	2.47	0.45	2.4
distance	mech_baseline	100	2.46	0.44	2.4
distance	ethnic_asian_hispanic	50	2.45	0.44	2.4
distance	llm_baseline	100	1.59	0.28	1.5
distance	income_high_low	10	1.41	0.22	1.4
distance	green_yellow	100	1.29	0.12	1.3
ghetto_rate	political_liberal_conservative	50	64.76	7.01	65.0
ghetto_rate	mech_baseline	100	48.54	7.66	48.0
ghetto_rate	race_white_black	100	43.64	10.04	44.0
ghetto_rate	ethnic_asian_hispanic	50	41.74	9.05	42.0
ghetto_rate	llm_baseline	100	19.48	9.57	17.0
ghetto_rate	income_high_low	10	11.20	7.41	10.5
ghetto_rate	green_yellow	100	7.65	4.71	7.0
mix_deviation	political_liberal_conservative	50	0.45	0.02	0.5
mix_deviation	ethnic_asian_hispanic	50	0.39	0.03	0.4
mix_deviation	race_white_black	100	0.38	0.03	0.4
mix_deviation	mech_baseline	100	0.38	0.03	0.4
mix_deviation	llm_baseline	100	0.28	0.05	0.3
mix_deviation	income_high_low	10	0.25	0.04	0.2
mix_deviation	green_yellow	100	0.20	0.03	0.2
share	political_liberal_conservative	50	0.95	0.03	0.9
share	mech_baseline	100	0.87	0.03	0.9
share	race_white_black	100	0.84	0.06	0.8
share	ethnic_asian_hispanic	50	0.84	0.04	0.8
share	llm_baseline	100	0.68	0.07	0.7
share	income_high_low	10	0.60	0.06	0.6
share	green_yellow	100	0.56	0.05	0.6
switch_rate	green_yellow	100	0.46	0.05	0.5

Continued on next page

Metric	Scenario	N	Mean	SD	Median
switch_rate	income_high_low	10	0.42	0.08	0.4
switch_rate	llm_baseline	100	0.36	0.08	0.4
switch_rate	mech_baseline	100	0.20	0.06	0.2
switch_rate	ethnic_asian_hispanic	50	0.19	0.04	0.2
switch_rate	race_white_black	100	0.18	0.06	0.2
switch_rate	political_liberal_conservative	50	0.06	0.03	0.1

Table 6: Pairwise scenario contrasts by metric: median differences (Δ), Cliff's δ, Vargha–Delaney A_{12} , and Holm-adjusted p.

Metric	Group 1	Group 2	Δ	δ	A_{12}	p_{adj}
clusters						
	ethnic_asian_hispanic	income_high_low	−12.00	−0.99	0.00	4.51×10^{-6}
	ethnic_asian_hispanic	political_liberal_conservative	3.00	0.50	0.75	6.21×10^{-5}
	ethnic_asian_hispanic	race_white_black	−2.00	−0.42	0.29	1.32×10^{-4}
	green_yellow	ethnic_asian_hispanic	11.00	0.98	0.99	0.00
	green_yellow	income_high_low	−1.00	−0.04	0.48	0.85
	green_yellow	political_liberal_conservative	14.00	1.00	1.00	0.00
	green_yellow	race_white_black	9.00	0.92	0.96	0.00
	income_high_low	political_liberal_conservative	15.00	1.00	1.00	4.43×10^{-6}
	income_high_low	race_white_black	10.00	0.96	0.98	3.95×10^{-6}
	llm_baseline	ethnic_asian_hispanic	5.00	0.79	0.89	0.00
	llm_baseline	green_yellow	−6.00	−0.69	0.16	0.00
	llm_baseline	income_high_low	−7.00	−0.76	0.12	2.33×10^{-4}
	llm_baseline	mech_baseline	8.00	0.99	0.99	0.00
	llm_baseline	political_liberal_conservative	8.00	0.95	0.98	0.00
	llm_baseline	race_white_black	3.00	0.54	0.77	6.00×10^{-10}
	mech_baseline	ethnic_asian_hispanic	−3.00	−0.62	0.19	3.80×10^{-9}
	mech_baseline	green_yellow	−14.00	−1.00	1.50×10^{-4}	0.00
	mech_baseline	income_high_low	−15.00	−1.00	0.00	1.24×10^{-6}
	mech_baseline	political_liberal_conservative	0.00	−0.09	0.46	0.77
	mech_baseline	race_white_black	−5.00	−0.89	0.06	0.00
	political_liberal_conservative	race_white_black	−5.00	−0.80	0.10	0.00
distance						
	ethnic_asian_hispanic	income_high_low	1.01	0.98	0.99	7.40×10^{-6}
	ethnic_asian_hispanic	political_liberal_conservative	−1.19	−0.91	0.05	0.00
	ethnic_asian_hispanic	race_white_black	−0.04	−0.06	0.47	1.00
	green_yellow	ethnic_asian_hispanic	−1.11	−1.00	0.00	0.00
	green_yellow	income_high_low	−0.10	−0.34	0.33	0.31
	green_yellow	political_liberal_conservative	−2.30	−1.00	0.00	0.00
	green_yellow	race_white_black	−1.16	−1.00	0.00	0.00
	income_high_low	political_liberal_conservative	−2.20	−1.00	0.00	5.21×10^{-6}
	income_high_low	race_white_black	−1.06	−0.97	0.02	3.86×10^{-6}
	llm_baseline	ethnic_asian_hispanic	−0.88	−0.92	0.04	0.00
	llm_baseline	green_yellow	0.23	0.76	0.88	0.00
	llm_baseline	income_high_low	0.13	0.43	0.71	0.13
	llm_baseline	mech_baseline	−0.87	−0.92	0.04	0.00
	llm_baseline	political_liberal_conservative	−2.07	−1.00	0.00	0.00
	llm_baseline	race_white_black	−0.93	−0.90	0.05	0.00
	mech_baseline	ethnic_asian_hispanic	−0.01	0.01	0.50	1.00
	mech_baseline	green_yellow	1.10	1.00	1.00	0.00
	mech_baseline	income_high_low	1.00	0.99	1.00	2.51×10^{-6}
	mech_baseline	political_liberal_conservative	−1.20	−0.91	0.05	0.00
	mech_baseline	race_white_black	−0.06	−0.05	0.48	1.00
	political_liberal_conservative	race_white_black	1.14	0.92	0.96	0.00
ghetto_rate						
	ethnic_asian_hispanic	income_high_low	31.50	1.00	1.00	5.07×10^{-6}
	ethnic_asian_hispanic	political_liberal_conservative	−23.00	−0.93	0.04	0.00
	ethnic_asian_hispanic	race_white_black	−2.00	−0.15	0.43	0.27
	green_yellow	ethnic_asian_hispanic	−35.00	−1.00	0.00	0.00
	green_yellow	income_high_low	−3.50	−0.28	0.36	0.27
	green_yellow	political_liberal_conservative	−58.00	−1.00	0.00	0.00
	green_yellow	race_white_black	−37.00	−1.00	0.00	0.00
	income_high_low	political_liberal_conservative	−54.50	−1.00	0.00	5.07×10^{-6}
	income_high_low	race_white_black	−33.50	−0.98	0.01	2.67×10^{-6}
	llm_baseline	ethnic_asian_hispanic	−25.00	−0.89	0.05	0.00
	llm_baseline	green_yellow	10.00	0.79	0.89	0.00
	llm_baseline	income_high_low	6.50	0.52	0.76	0.02
	llm_baseline	mech_baseline	−31.00	−0.97	0.02	0.00
	llm_baseline	political_liberal_conservative	−48.00	−1.00	6.00×10^{-4}	0.00
	llm_baseline	race_white_black	−27.00	−0.89	0.05	0.00
	mech_baseline	ethnic_asian_hispanic	6.00	0.46	0.73	2.45×10^{-5}
	mech_baseline	green_yellow	41.00	1.00	1.00	0.00
	mech_baseline	income_high_low	37.50	1.00	1.00	1.82×10^{-6}
	mech_baseline	political_liberal_conservative	−17.00	−0.87	0.07	0.00
	mech_baseline	race_white_black	4.00	0.29	0.64	0.00
	political_liberal_conservative	race_white_black	21.00	0.92	0.96	0.00
mix_deviation						
	ethnic_asian_hispanic	income_high_low	0.15	1.00	1.00	5.22×10^{-6}
	ethnic_asian_hispanic	political_liberal_conservative	−0.07	−0.91	0.05	0.00
	ethnic_asian_hispanic	race_white_black	8.57×10^{-4}	-8.00×10^{-4}	0.50	1.00
	green_yellow	ethnic_asian_hispanic	−0.18	−1.00	0.00	0.00
	green_yellow	income_high_low	−0.04	−0.60	0.20	0.01
	green_yellow	political_liberal_conservative	−0.25	−1.00	0.00	0.00
	green_yellow	race_white_black	−0.18	−1.00	3.00×10^{-4}	0.00
	income_high_low	political_liberal_conservative	−0.22	−1.00	0.00	5.22×10^{-6}

Continued on next page

Metric	Group 1	Group 2	Δ	δ	A_{12}	p_{adj}
	income_high_low	race_white_black	−0.15	−0.99	0.01	2.20×10^{-6}
	llm_baseline	ethnic_asian_hispanic	−0.11	−0.93	0.04	0.00
	llm_baseline	green_yellow	0.07	0.81	0.91	0.00
	llm_baseline	income_high_low	0.03	0.40	0.70	0.16
	llm_baseline	mech_baseline	−0.10	−0.89	0.06	0.00
	llm_baseline	political_liberal_conservative	−0.18	−1.00	0.00	0.00
	llm_baseline	race_white_black	−0.11	−0.90	0.05	0.00
	mech_baseline	ethnic_asian_hispanic	−0.01	−0.15	0.43	0.30
	mech_baseline	green_yellow	0.17	1.00	1.00	0.00
	mech_baseline	income_high_low	0.14	0.99	1.00	2.20×10^{-6}
	mech_baseline	political_liberal_conservative	−0.08	−0.94	0.03	0.00
	mech_baseline	race_white_black	−0.01	−0.14	0.43	0.30
	political_liberal_conservative	race_white_black	0.07	0.92	0.96	0.00
share						
	ethnic_asian_hispanic	income_high_low	0.23	1.00	1.00	5.22×10^{-6}
	ethnic_asian_hispanic	political_liberal_conservative	−0.11	−0.94	0.03	0.00
	ethnic_asian_hispanic	race_white_black	0.00	−0.06	0.47	0.52
	green_yellow	ethnic_asian_hispanic	−0.28	−1.00	0.00	0.00
	green_yellow	income_high_low	−0.05	−0.39	0.31	0.09
	green_yellow	political_liberal_conservative	−0.38	−1.00	0.00	0.00
	green_yellow	race_white_black	−0.28	−1.00	7.00×10^{-4}	0.00
	income_high_low	political_liberal_conservative	−0.33	−1.00	0.00	5.22×10^{-6}
	income_high_low	race_white_black	−0.23	−0.99	0.00	1.95×10^{-6}
	llm_baseline	ethnic_asian_hispanic	−0.17	−0.93	0.04	0.00
	llm_baseline	green_yellow	0.11	0.83	0.92	0.00
	llm_baseline	income_high_low	0.06	0.55	0.77	0.01
	llm_baseline	mech_baseline	−0.19	−0.97	0.02	0.00
	llm_baseline	political_liberal_conservative	−0.27	−1.00	6.00×10^{-4}	0.00
	llm_baseline	race_white_black	−0.17	−0.91	0.05	0.00
	mech_baseline	ethnic_asian_hispanic	0.03	0.43	0.72	9.44×10^{-5}
	mech_baseline	green_yellow	0.30	1.00	1.00	0.00
	mech_baseline	income_high_low	0.25	1.00	1.00	1.86×10^{-6}
	mech_baseline	political_liberal_conservative	−0.08	−0.90	0.05	0.00
	mech_baseline	race_white_black	0.02	0.32	0.66	4.24×10^{-4}
	political_liberal_conservative	race_white_black	0.10	0.93	0.96	0.00
switch_rate						
	ethnic_asian_hispanic	income_high_low	−0.25	−1.00	0.00	5.22×10^{-6}
	ethnic_asian_hispanic	political_liberal_conservative	0.13	0.97	0.98	0.00
	ethnic_asian_hispanic	race_white_black	0.00	0.15	0.57	0.22
	green_yellow	ethnic_asian_hispanic	0.28	1.00	1.00	0.00
	green_yellow	income_high_low	0.02	0.35	0.67	0.21
	green_yellow	political_liberal_conservative	0.40	1.00	1.00	0.00
	green_yellow	race_white_black	0.28	1.00	1.00	0.00
	income_high_low	political_liberal_conservative	0.38	1.00	1.00	5.22×10^{-6}
	income_high_low	race_white_black	0.26	0.98	0.99	2.57×10^{-6}
	llm_baseline	ethnic_asian_hispanic	0.18	0.93	0.97	0.00
	llm_baseline	green_yellow	−0.09	−0.71	0.14	0.00
	llm_baseline	income_high_low	−0.07	−0.39	0.30	0.17
	llm_baseline	mech_baseline	0.16	0.90	0.95	0.00
	llm_baseline	political_liberal_conservative	0.31	1.00	1.00	0.00
	llm_baseline	race_white_black	0.19	0.92	0.96	0.00
	mech_baseline	ethnic_asian_hispanic	0.03	0.16	0.58	0.22
	mech_baseline	green_yellow	−0.25	−1.00	3.00×10^{-4}	0.00
	mech_baseline	income_high_low	−0.23	−0.99	0.00	2.32×10^{-6}
	mech_baseline	political_liberal_conservative	0.15	0.96	0.98	0.00
	mech_baseline	race_white_black	0.03	0.27	0.63	0.01
	political_liberal_conservative	race_white_black	−0.12	−0.93	0.04	0.00

Notes: Δ = median(g_1) − median(g_2) (or mean difference if medians unavailable). Holm adjustment within metric. If Cliff's δ is present, $A_{12} = (\delta + 1)/2$.

Table 7: Pairwise comparisons between baseline and other social contexts

Context	Ghetto Rate	Ghetto vs Baseline	Share	Share vs Baseline
Ethnic (Asian/Hispanic)	38.9	+100%	0.821	+20.9%
Income (High/Low)	5.0	-74%	0.543	-20.1%
Political (Liberal/Conservative)	61.6	+216%	0.928	+36.6%
Race (White/Black)	40.8	+109%	0.823	+21.2%