LLM Active Alignment: A Nash Equilibrium Perspective

Tonghan Wang[*]
Tsinghua University
thw@mail.tsinghua.edu.cn

Yuqi Pan[*]
Harvard University
yuqipan@g.harvard.edu

Xinyi Yang[*]
Tsinghua University
cindyiyang@tencent.com

Xinrui Song
Shandong University
202300800600@mail.sdu.edu.cn

Yanchen Jiang
Harvard University
yanchen_jiang@g.harvard.edu

Milind Tambe
Harvard University
tambe@g.harvard.edu

David C. Parkes
Harvard University
parkes@eecs.harvard.edu

Abstract

We develop a game-theoretic framework for predicting and steering the behavior of populations of large language models (LLMs) through Nash equilibrium (NE) analysis. To avoid the intractability of equilibrium computation in open-ended text spaces, we model each agent's action as a mixture over human subpopulations. Agents choose actively and strategically which groups to align with, yielding an interpretable and behaviorally substantive policy class. We derive closed-form NE characterizations, adopting standard concave-utility assumptions to enable analytical system-level predictions and give explicit, actionable guidance for shifting alignment targets toward socially desirable outcomes. This method provides equilibrium targets that can inform existing alignment pipelines. In a social-media setting, we show that a population of LLMs, especially reasoning-based models, may exhibit political exclusion—pathologies where some subpopulations are ignored by all LLM agents—which can be avoided by our method, illustrating the promise of applying the method to regulate multi-agent LLM dynamics across domains.

1 Introduction

In this paper, we study how to analyze and steer the behavior of a community of large language models (LLMs) by calculating and designing their Nash equilibria (NE) [38, 14]. NE is a game-theoretic concept that characterizes stable states of a system of strategic players: no player has an incentive to unilaterally change their strategy, as doing so would not increase their utility [25]. This stability makes NE a natural object for predicting player behavior.

As it becomes increasingly conceivable that the world will be populated by LLM agents [9, 7, 43, 41, 28], understanding their NE becomes important [5, 44]. In particular, as we argue, it provides a game-theoretic approach for addressing the emerging challenge of AI alignment in multi-agent, LLM systems. Specifically, in this paper, we study the NE of LLM agents to answer the following question:

[*]Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

As LLMs exercise agency over an increasing number of daily activities, they will inevitably interact with and influence one another. How can we ensure that the resulting dynamics do not lead to outcomes harmful or undesirable for humans?

Our goal is to derive and design the NE of an LLM-agent population as both a prediction and governance primitives for system outcomes. However, the problem of finding a NE is *PPAD-complete* [10], presenting a major challenge, with NE calculation becoming intractable as LLMs operate in the space of open-ended text. Prior work [47, 49, 20, 21, 17, 48] enforces tractability by restricting the action space to small, task-specific sets, such as tones, rhetorical styles, or low-dimensional generation controls (e.g., temperature). While these abstractions make equilibrium computation feasible, the resulting equilibria may offer limited leverage for predicting system-level behavior or designing interventions that remain valid beyond the prescribed action set.

To address this tension, we define LLM strategies as mixtures over human subpopulations, with each subpopulation associated with preference or opinion data. Under this formulation, an LLM agent actively and strategically selects which human subpopulations to align with. The resulting space of modeled LLM behaviors is substantive: it is rich and interpretable, expresses a continuum of behavior, and enables analyses for realistic LLM populations.

From a theoretical perspective, we derive closed-form expressions for the Nash Equilibrium (NE) of LLM agents. We focus on settings with concave utilities — a standard assumption in the field — and utilize concave games, which serve as a generalization of normal-form games [13, 27, 30]. Our results also allow for governance-relevant equilibrium interventions. They provide direct guidance on designing the environment in which LLM agents operate, and the interaction among them, to shift the alignment target of LLM agents, steering collective behavior toward desirable system-level behavior.

This game-theoretic alignment complements established LLM alignment methods such as *RLHF* [26, 29, 2, 40] and *pluralism* [37, 12]. It introduces an active notion of alignment, where the target arises from LLM agents' rational choices in strategic interaction rather than from fitting an aggregate of human preferences passively. Our approach can be viewed as a layer built atop existing alignment pipelines, and be used to better navigate the sophisticated, open-ended human-AI ecosystem, where system-level outcomes can differ sharply from the behavior of any single model. As we show, this incentive-aware alignment layer can mitigate undesirable outcomes arising from strategic interaction that may persist even when individual models are well-aligned in isolation.

We instantiate our theoretical framework in a concrete scenario: LLMs operating social media accounts, each with self-interested motives to maximize the attention and engagement of human users. Recent work has shown that algorithms can shape creator incentives and induce unexpected equilibrium effects [16], making it important to understand, predict, and steer the equilibrium outcomes of LLMs from a computational perspective. We observe a disturbing phenomenon: at an NE, independent of how LLMs are aligned individually, the LLM population can exhibit *political exclusion*, in which the opinions of certain subpopulations are ignored by every LLM. Reasoning models, such as *Qwen3-4B-Thinking* and *DeepSeek-R1-Distill-Qwen-7B*, can even exacerbate such exclusion. We characterize and visualize the conditions under which exclusion arises. We further show that our method enables a rigorous analysis about how to design the NE to avoid this undesirable outcome, incentivizing LLMs to attend, actively and rationally, to the voices of all subpopulations. Although we use political exclusion in social media as a running example in this paper, our method is general and can be used to regulate the behavior of LLM agents across a wide range of domains.

Related work. *Game theory and equilibria for LLM agents.* A growing body of work studies LLMs as strategic agents and uses game-theoretic concepts to analyze or shape their interactions [38]. Several lines of work use equilibria to coordinate multi-LLM reasoning [47, 49, 48]. Others instantiate LLMs as players in explicit game environments, examining their strategic behaviors in multi-agent settings [21, 17]. Gemp et al. [14] use game-theoretic solvers to steer a single LLM's generation by casting decoding controls, such as tone and argument styles, as actions. Different from existing work, our method introduces subpopulation alignment targets as the strategy space, enabling a theoretically grounded toolkit, supported by empirical evidence, that makes equilibrium analysis feasible while remaining behaviorally informative for steering LLM populations in general scenarios.

Pluralistic alignment and population simulation also represents heterogeneous human values or behaviors via mixtures of subpopulations or group-specific models. Existing approaches typically use this structure to (i) measure which groups a model reflects [31], (ii) train or compose systems

that better serve diverse user groups [12, 6, 37], or (iii) simulate populations by mixing persona- or group-conditioned behaviors [4, 42]. By contrast, we repurpose such structure as a strategy space for LLM populations, enabling equilibrium characterization and incentive-based governance analysis.

Algorithmic curation, engagement incentives, and exposure as a societal quantity. Prior work demonstrates that engagement-optimized recommendation and ranking reshape information exposure by determining what content is amplified and which groups are heard [3, 8, 23]. These concerns align with a line of fairness research that explicitly models *exposure* as a primary societal metric [35, 45]. Our equilibrium lens formalizes conditions under which certain groups become systematically under-represented and how incentive design can mitigate such outcomes.

2 NE preliminaries and computational challenges in LLM populations

In the simplest abstraction, in a normal-form game $G = (N, (A_i)_{i\in N}, (u_i)_{i\in N})$, each agent i selects an action $a_i \in A_i$ and receives utility $u_i(a_1, \ldots, a_n)$. A *strategy* in a normal-form game for agent i is a distribution $\sigma_i \in \Delta(A_i)$. We write $\sigma = (\sigma_1, \ldots, \sigma_n)$ for a strategy profile and σ_{-i} for the strategies of all agents other than i. A strategy profile $\sigma^\star$ is a *Nash equilibrium (NE)* if no agent can increase its expected utility by unilateral deviation:

$$\mathbb{E}[u_i(\sigma_i^\star, \sigma_{-i}^\star)] \geq \mathbb{E}[u_i(\sigma_i, \sigma_{-i}^\star)] \;\; \forall \sigma_i \in \Delta(A_i), \forall i \in N.$$

This property makes NE a natural behavioral-prediction tool: what outcomes are stable under strategic pressure. Despite its conceptual attractiveness, NE computation belongs to the complexity class *PPAD-complete* [10], and is widely conjectured not to admit efficient algorithms. The problem exacerbates when agents are LLMs, as an LLM agent's behavior is not well captured by a small discrete action set. The action is better viewed as a policy mapping prompts and histories to distributions over token sequences. This induces an enormous, implicitly represented strategy space. Moreover, utilities typically depend on long-horizon interactive trajectories, e.g., multi-turn persuasion, negotiation, or collaboration, rendering the evaluation of even a single strategy profile expensive.

3 LLM active alignment

The computational challenges associated with NE in LLM populations mean that any proposal to use NE as a prediction and governance primitive in LLM ecosystems has to *make equilibrium computation feasible without sacrificing the capacity for behavior predictions and regulatory intervention*. We address this by introducing a structured, low-rank strategy space grounded in human subpopulations and by adopting the framework of concave games. We formalize this setting in Section 3.1 and, under this structure, derive a closed-form NE characterization in Section 3.2.

3.1 Subpopulation-based strategy space

Rather than attempting to compute NE over unconstrained textual policy spaces, we work with a compact strategy space in which a strategy puts weights to aligning LLM behavior with preferences or opinions of each subpopulation.

This design is motivated by the fact that many existing preference and opinion datasets are annotated with respondent attribute labels, such as demographic or psychographic information, enabling principled partitioning of human judgments or opinions into interpretable groups. For example, the `OpinionQA` dataset [31] includes demographic labels for each response, while the `Big Five Personality Traits` dataset [39] stratifies respondents along personality dimensions such as openness and conscientiousness. More broadly, subpopulations can be defined using age, education, or other attributes. Subpopulations induced by these labels typically exhibit within-group behavioral coherence, and an LLM may aim for its behavior to reflect and preserve this structure.

We operationalize this idea by partitioning a dataset into D disjoint sub-datasets, each representing a distinct labeled subpopulation. Each *subpopulation dataset* is denoted by $\mathcal{S}_d = \{(x_d^{(k)}, y_d^{(k)})\}_{k=1}^{K_d},\ d \in [D]$, where $x_d^{(k)}$ is a prompt (e.g., a question), $y_d^{(k)}$ is the corresponding human response, and K_d is the number of prompt–response pairs associated with subpopulation d. To represent the behavior of a subpopulation d, we train a LLM *subpopulation model*, $\nu_d(y|x)$,

parameterizing a conditional distribution over responses given prompts. The training is performed by minimizing the standard *autoregressive negative likelihood* on $\mathcal{S}_d$.

With this setup, we consider a system of M strategic players, namely the LLM agents. M could be very large. The *strategy* of player $m \in [M]$ is a weight vector $w_m \in \Delta_D$, where $\Delta_D := \{w \in \mathbb{R}^D_{\geq 0} : \sum_{d=1}^D w_d = 1\}$ is the probability simplex over the D subpopulation models. Given w_m, player m induces a conditional response distribution $\pi_m(\cdot \mid x)$ by mixing the subpopulation conditionals:

$$\pi_m(y|x) = \sum_{d=1}^{D} \mathbf{w}_{m,d} \nu_d(y|x), \qquad \forall x, y. \tag{1}$$

Advantages of this setup. Eq. (1) yields an immediate complexity benefit by replacing an implicit, prohibitively large textual policy space with a tractable, low-dimensional convex strategy set over weights, while still permitting equilibrium-based reasoning to be performed efficiently on realistic instances. Meanwhile, this finite-dimensional strategy space still supports open-ended response generation. Its expressive scope is determined by the chosen subpopulation models. An equilibrium mixture $\mathbf{w}_m^\star$ specifies an alignment target within this policy class and induces the corresponding response policy $\pi_m^\star$ through Eq. (1). Within this class, alignment becomes an endogenous strategic choice: agents choose which subpopulations to represent, and to what extent, in response to their incentives and other agents' strategies. This enables governance interventions to be analyzed through their effects on equilibrium alignment targets. We refer to this strategic selection of alignment targets as LLM *active alignment*. The framework provides an incentive layer that can complement existing alignment pipelines such as RLHF and RLAIF [2, 18, 32]. The next subsection characterizes these equilibrium targets under a specified utility model.

3.2 NE calculation

We now characterize NE in the subpopulation-mixture strategy space introduced above. Here the equilibrium calculation is separated from the semantics of a particular deployment. A domain specifies how incentives are assigned to subpopulation mixtures. Once these incentives are represented as a concave utility over the simplex, the NE calculation is application-agnostic. Concrete instantiations, including the social-media setting used in our experiments, are discussed in Appx. C.

3.2.1 Canonical concave quadratic utility class

We consider the following canonical concave quadratic utility class, which is a normalized subclass of broader quadratic utilities.

Assumption 1. *Up to additive terms that do not depend on $\mathbf{w}_m$, the utility of player m is*

$$u_m(\mathbf{w}_m, \mathbf{w}_{-m}) = \beta^{(A)} \mathbf{a}^\top \mathbf{w}_m - \beta^{(I)} \mathbf{w}_m^\top \boldsymbol{C} \mathbf{w}_m - \beta^{(D)} \sum_{j \neq m} \mathbf{w}_m^\top \mathbf{w}_j, \tag{2}$$

where $\mathbf{a} \in \mathbb{R}^D$ and $\boldsymbol{C} \in \mathbb{R}^{D \times D}$. Also, $\boldsymbol{C} = \boldsymbol{C}^\top \succeq 0$ and $\beta^{(A)}, \beta^{(I)}, \beta^{(D)} > 0$.

The first term is a domain-specific linear score for each subpopulation. The second term is a concave quadratic self-regularization term: since $\boldsymbol{C} \succeq 0$, $-\mathbf{w}_m^\top \boldsymbol{C} \mathbf{w}_m$ penalizes mixtures that are costly for a single agent to maintain. The third term is a canonical cross-agent overlap penalty, where $\mathbf{w}_m^\top \mathbf{w}_j = \sum_{d=1}^D w_{m,d} w_{j,d}$ measures how much two agents target the same subpopulations, and thus $-\sum_{j \neq m} \mathbf{w}_m^\top \mathbf{w}_j$ captures congestion, redundancy, or diminishing marginal value when multiple agents concentrate on the same targets. The coefficients $\beta^{(A)}, \beta^{(I)}, \beta^{(D)}$ scale the relative strengths of these three incentive components. Moreover, these quantities are shared across players. Terms depending only on $\mathbf{w}_{-m}$ are omitted because they do not affect player m's best response.

Proposition 2. *Under Assumption 1, the game is a concave game: for every player m and every fixed $\mathbf{w}_{-m}$, $u_m(\cdot, \mathbf{w}_{-m})$ is concave on Δ_D.*

3.2.2 Equilibrium computation

We now follow the framework of concave games to derive a tractable approach to computing NE.

Lemma 3. *With utility function $u_m(\mathbf{w}_m, \mathbf{w}_{-m})$ defined in Eq.* (2)*. A strategy $\mathbf{w}_m$ in the interior of the simplex is a best response to $\mathbf{w}_{-m}$ if and only if the utility gradient satisfies, for some $\lambda_m \in \mathbb{R}$,*

$$\nabla_{\mathbf{w}_m} u_m = -\beta^{(D)} \sum_{j \neq m} \mathbf{w}_j + \beta^{(A)} \boldsymbol{a} - 2\beta^{(I)} \boldsymbol{C} \mathbf{w}_m = \lambda_m \mathbf{1}.$$

Here, "interior" means w_m assigns strictly positive mass to every subpopulation $i \in [D]$. The proof is in Appx. A and is based on Gemp et al. [13]. Therefore, an interior NE can be obtained by solving the following linear system

$$\begin{cases} -\beta^{(D)} \sum\limits_{j \neq m} \mathbf{w}_j + \beta^{(A)} \boldsymbol{a} - 2\beta^{(I)} \boldsymbol{C} \mathbf{w}_m = \lambda_m \mathbf{1}, \forall m; \\ \mathbf{1}^\top \mathbf{w}_m = 1, \forall m. \end{cases} \tag{3}$$

To derive an analytical solution for an interior equilibrium, we first temporarily relax the interior constraint, $\mathbf{w}_m > \mathbf{0}$. As we will discuss later, the solution to this relaxed problem constitutes the unique valid interior equilibrium if and only if all resulting weights are positive. Conversely, if the relaxed problem yields no such solution with all weights being strictly positive, then no interior equilibrium exists. Appx. B discusses deriving NE that lie on the boundary of the simplex.

For notational convenience, let $\boldsymbol{A} = 2\beta^{(I)} \boldsymbol{C} - \beta^{(D)} \boldsymbol{I}$, then the first line in Eq. (3) is equivalent to

$$(\boldsymbol{A} + \beta^{(D)} \boldsymbol{I}) \mathbf{w}_m + \beta^{(D)} \sum_{j \neq m} \mathbf{w}_j = \beta^{(A)} \boldsymbol{a} - \lambda_m \mathbf{1}, \quad \forall m \in [M]. \tag{4}$$

Here $\boldsymbol{a}$ is the vector defined in Eq. (2). Subtracting any two equations in Eq. (4) gives

$$\boldsymbol{A}(\mathbf{w}_i - \mathbf{w}_j) = (\lambda_j - \lambda_i)\, \mathbf{1}.$$

Since $\boldsymbol{A}$ is an affine combination of symmetric matrices, it is symmetric. If $\mu_1, \ldots, \mu_D$ are the eigenvalues of $\boldsymbol{C}$, then the eigenvalues of $\boldsymbol{A}$ are $2\beta^{(I)}\mu_j - \beta^{(D)}$ for $j = 1, \ldots, D$. Hence $\boldsymbol{A}$ is full rank iff $2\beta^{(I)}\mu_j \neq \beta^{(D)}$ for all j. This excludes only finitely many hyperparameter values (a measure-zero set). Therefore, we may, without loss of generality, assume that $\boldsymbol{A}$ is full rank. This can always be enforced by adding an arbitrarily small ridge term, $\varepsilon \boldsymbol{I}$, with $\varepsilon > 0$, which is equivalent to replacing $\beta^{(D)}$ by $\beta^{(D)} + \varepsilon$. Using Eq. (4) and solving for $\mathbf{w}_m$ yields

$$(\boldsymbol{A} + M\beta^{(D)} \boldsymbol{I}) \mathbf{w}_m = \beta^{(A)} \boldsymbol{a} - \lambda_m \mathbf{1} + \sum_{i=1}^{M} (\lambda_i - \lambda_m)\, \beta^{(D)} \boldsymbol{A}^{-1} \mathbf{1}.$$

Since $\boldsymbol{C}$ is positive semidefinite and $M > 1$, the eigenvalues of $\boldsymbol{A} + M\beta^{(D)} \boldsymbol{I}$: $2\beta^{(I)}\mu_i + (M-1)\beta^{(D)}$ are larger than 0, which means $\boldsymbol{A} + M\beta^{(D)} \boldsymbol{I}$ is invertible. Therefore,

$$\mathbf{w}_m = (\boldsymbol{A} + M\beta^{(D)} \boldsymbol{I})^{-1} (\beta^{(A)} \boldsymbol{a} - \lambda_m \mathbf{1} + \beta^{(D)} \sum_{i=1}^{M} (\lambda_i - \lambda_m)\, \boldsymbol{A}^{-1} \mathbf{1}). \tag{5}$$

Let $\boldsymbol{B} = (\boldsymbol{A} + M\beta^{(D)} \boldsymbol{I})^{-1}$. Premultiplying Eq. (5) by $\mathbf{1}^\top$ and using $\mathbf{1}^\top \mathbf{w}_m = 1$ gives a linear relation among $\{\lambda_i\}$:

$$-\underbrace{\left(s_1 + \beta^{(D)} M s_2\right)}_{\alpha} \lambda_m \;+\; \beta^{(D)} s_2 \sum_{i=1}^{M} \lambda_i \;=\; 1 - s_0, \tag{6}$$

with $s_0 = \mathbf{1}^\top \mathbf{B} \beta^{(A)} \boldsymbol{a}$, $s_1 = \mathbf{1}^\top \boldsymbol{B} \mathbf{1}$, $s_2 = \mathbf{1}^\top \boldsymbol{B} \boldsymbol{A}^{-1} \mathbf{1}$, $\alpha = \mathbf{1}^\top \boldsymbol{B} (\boldsymbol{I} + M\beta^{(D)} \boldsymbol{A}^{-1}) \mathbf{1} = \mathbf{1}^\top \boldsymbol{A}^{-1} \mathbf{1}$.

Lemma 4 [Only finitely many $\beta^{(D)}/\beta^{(I)}$ make $\alpha = 0$]**.** *Let $\beta := \beta^{(D)}/\beta^{(I)}$. Take the eigendecomposition $\boldsymbol{C} = \boldsymbol{Q}\, \mathrm{diag}(\mu_1, \ldots, \mu_D)\, \boldsymbol{Q}^\top$ and $\boldsymbol{q} := \boldsymbol{Q}^\top \mathbf{1}$. Then*

$$\alpha \;=\; \mathbf{1}^\top \boldsymbol{A}^{-1} \mathbf{1} \;=\; \frac{1}{\beta^{(I)}} \sum_{j=1}^{D} \frac{q_j^2}{2c_j - \beta} \;=:\; \frac{1}{\beta^{(I)}}\, f(\beta).$$

The equation $\alpha = 0$ (equivalently $f(\beta) = 0$) has at most finitely many solutions β.

The proof is in Appx. A. In the *generic* case $\alpha \neq 0$, Eq. (6) forces all multipliers to be equal:

$$\lambda_1 = \cdots = \lambda_M = \lambda^\star, \qquad \lambda^\star = \frac{\beta^{(A)} \mathbf{1}^\top \mathbf{B}\, \boldsymbol{a} - 1}{\mathbf{1}^\top \mathbf{B}\, \mathbf{1}}.$$

$\boldsymbol{B}$ is positive definite because $\boldsymbol{C}$ is positive semidefinite and $\beta^{(D)}, \beta^{(I)} > 0, M > 1$. Therefore, the denominator is guaranteed to be larger than 0. Plugging this back into Eq. (5) cancels the summation term (since $\sum_i (\lambda_i - \lambda^\star) = 0$), yielding the interior equilibrium characterization as follows.

Theorem 5. *Under Assumption 1, the interior Nash equilibrium is unique and homogeneous. For any LLM player $m \in [M]$,*

$$\mathbf{w}_m = \mathbf{w}^* = \left(2\beta^{(I)} \boldsymbol{C} + (M-1)\beta^{(D)} \boldsymbol{I}\right)^{-1} \left(\beta^{(A)} \boldsymbol{a} - \lambda^\star \mathbf{1}\right) \tag{7}$$

Remark 6. Since Eq. (7) is derived without enforcing the interior constraint, it characterizes an interior equilibrium only when every component of $\mathbf{w}^*$ is strictly positive. If some component is non-positive, no interior equilibrium exist, then any NE must lie on the boundary of the weight simplex, with some subpopulations assigned zero weight. We discuss how to compute boundary NE in Appx. B. For Eq. (7) to hold, $\boldsymbol{A}$ must be invertible and the normalization factor nonzero, equivalently, $\beta^{(D)}/\beta^{(I)} \neq 2\mu_i$ and $f(\beta) \neq 0$, see Lemma 4. These singular cases occur only at finitely many hyperparameter ratios. Hence Eq. (7) is well-defined almost everywhere, and we ignore these measure-zero singularities in the analysis.

Utility instantiation The utility in Eq. (2) separates equilibrium computation from domain semantics. To instantiate the framework in a deployment domain, one only needs to specify a linear subpopulation score $\boldsymbol{a}$ and a positive semidefinite curvature matrix $\boldsymbol{C}$. In the social-media setting, each LLM account chooses a mixture over subpopulation-conditioned policies. The coordinate a_d captures the platform-side value of representing subpopulation d, such as audience share, expected engagement or exposure. The matrix C captures the cost of mixing subpopulation voices, such as account-identity instability or response inconsistency between subpopulation models. Thus, $\beta^{(A)}$ rewards attractive or high-exposure targets, $\beta^{(I)}$ penalizes incoherent or costly mixtures, and $\beta^{(D)}$ penalizes cross-agent overlap, modeling attention dilution and discouraging population-wide concentration on the same targets. Once $\boldsymbol{a}$ and $\boldsymbol{C}$ are fixed, Theorem 5 directly yields the equilibrium mixture $\boldsymbol{w}^\star$. Zero or near-zero coordinates of $\boldsymbol{w}^\star$ indicate subpopulations that are excluded at equilibrium. Other domain instantiations appear in Appendix C.

In our experiments, a_i is the mean inner product between subpopulation-model option probabilities and human preference distributions. Off-diagonal C_{ij} are mean absolute probability discrepancies, with $C_{ii} = \sum_{j \neq i} C_{ij}$. These inputs are shared across agents using the same population and subpopulation models.

4 Experimental results

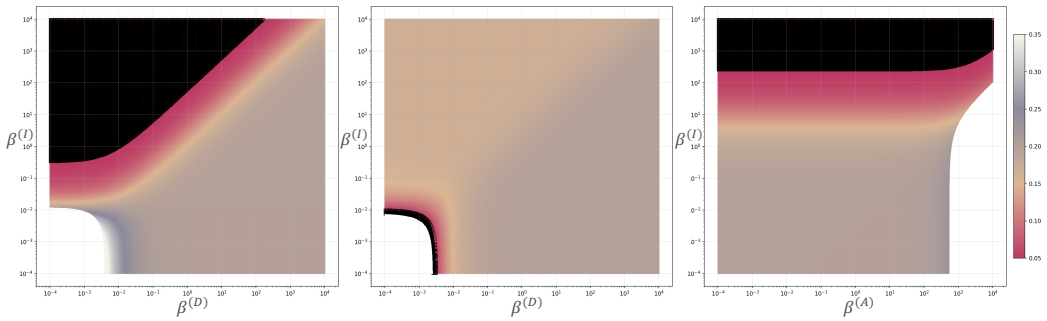
We design experiments to test our theoretical results in realistic settings and answer the following questions: (1) When do equilibria assign vanishing weights to some subpopulations? How does this exclusion depend on whether attention-seeking ($\beta^{(A)}$), conflict-avoidance ($\beta^{(I)}$), or other objectives ($\beta^{(D)}$) are incentivized? (2) Are exclusion patterns fragile corner cases, or do they form stable, structured regions of the incentive space? (3) Do different LLM models systematically expand or shrink the parameter regimes in which exclusion arises?

To answer these questions, we sweep $(\beta^{(A)}, \beta^{(I)}, \beta^{(D)})$ and use Eq. (7) to map how equilibrium weights vary. We interpret $(\beta^{(A)}, \beta^{(I)}, \beta^{(D)})$ as *incentive configurations*: a social media platform can design incentivization mechanisms that effectively shape these coefficients.

Our experiments focus primarily on interior equilibria because our ultimate goal is to understand when political exclusion can be avoided. Boundary equilibria mechanically force at least one subpopulation weight to be zero, making exclusion inevitable by definition.

4.1 Setup

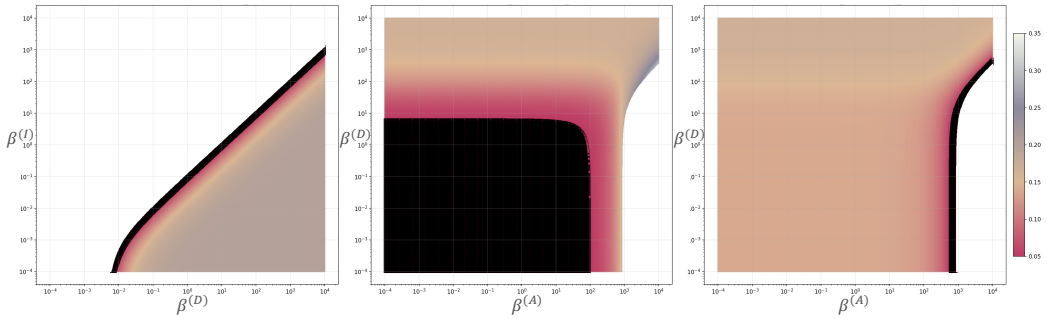
We use three datasets with explicit subpopulation labels: `POLITICS` [22], `CultureBank` [33] and `Big Five Personality Traits` (`Big Five` for short) [39]. They are widely studied in recent



(a) Model: *Qwen3-4B-Thinking-2507*
Dataset: `Big Five Personality`
Subpopulation: Neuroticism

(b) Model: *Qwen3-7B*
Dataset: `Big Five Personality`
Subpopulation: Openness

(c) Model: *DeepSeek-R1-7B*
Dataset: `Big Five Personality`
Subpopulation: Conscientiousness



(d) Model: *Mistral-7B-Instruct-v0.2*
Dataset: `CultureBank`
Subpopulation: North America

(e) Model: *Mistral-7B-Instruct-v0.2*
Dataset: `POLITICS`
Subpopulation: Reddit-Left

(f) Model: *Mistral-7B-Instruct-v0.2*
Dataset: `POLITICS`
Subpopulation: News-Left

Figure 1: **Representative examples of political exclusion across models and datasets** (complete results in Appx. D). Each panel fixes the base model, dataset, and subpopulation, and visualizes the subpopulation's interior Nash-equilibrium weight under varying preference coefficients. Specifically, one coefficient among $\beta^{(A)}$, $\beta^{(I)}$, and $\beta^{(D)}$ is fixed to 1, while the remaining two vary along the x- and y-axes (as labeled, log-scaled). Color indicates the resulting equilibrium weight assigned to the focal subpopulation. Regions where the weight falls below 0.05 are marked in black and referred to as the *political exclusion area*. White regions indicate parameter values for which no interior equilibrium exists. We focus on interior equilibria because boundary equilibria necessarily set at least one subpopulation weight to zero and our goal is to understand how to avoid political exclusion.

literature [11, 34, 12, 24, 36, 46, 15]. The `POLITICS` dataset comprises six distinct subpopulations spanning left-, center-, and right-leaning viewpoints, with data drawn from both news outlets and social media. The `CultureBank` dataset includes five regional subpopulations. In the `Big Five` dataset, the five subpopulations correspond to distinct personality traits.

First, to reduce confounds arising from subpopulation-model training and isolate the effects we study, on datasets `POLITICS` and `CultureBank`, we use trained subpopulation models provided by Feng et al. [12], where *Mistral-7B-Instruct-v0.2* is used as the base model and we hold them fixed as ν. Furthermore, to ensure the robustness of our observations and analyses, on the dataset `Big Five`, we conduct experiments across different base models, including *Qwen3-0.6B*, *Qwen3-1.7B*, *Qwen3-4B*, *Qwen3-7B*, *Qwen3-14B*, *DeepSeek-R1-Distill-Qwen-7B*, *Mistral-7B-Instruct-v0.2*, and *Qwen3-4B-Thinking-2507*. Among them, *DeepSeek-R1-Distill-Qwen-7B* and *Qwen3-4B-Thinking-2507* are reasoning-based models. We train subpopulation models using standard autoregressive next-token prediction. We use LoRA with a rank of 16, controlling the number of trainable parameters, and alpha of 16, scaling the influence of the adapted weights. We train the model for 5 epochs using the Adam optimizer with a learning rate of $2e{-}5$, a batch size of 64, and a dropout rate of 0.05.

We visualize the equilibrium weight assigned to each subpopulation defined by Eq. (7), under varying coefficient configurations $(\beta^{(A)}, \beta^{(I)}, \beta^{(D)})$. Each configuration corresponds to a possible utility function for the LLM population. Section 4.2 reports equilibrium-model predictions from empirically estimated $\boldsymbol{a}$ and $\boldsymbol{C}$, including the reasoning-model comparison. Section 4.3 tests these predictions with independently learning agents.

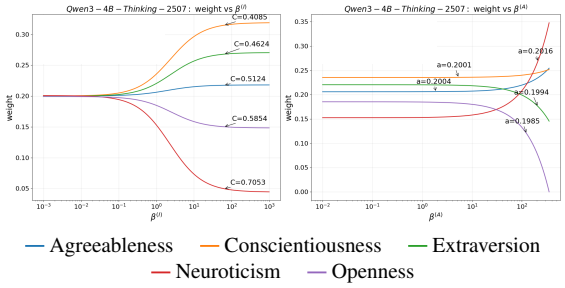


Figure 2: **Middle-of-the-road survives.** 1D slices of the equilibrium weights for *Qwen3-4B-Thinking-2507* on `Big Five`. Left: increasing $\beta^{(I)}$ ($\beta^{(A)}$=$\beta^{(D)}$=1) suppresses the most inconsistent subpopulation (largest C, the diagonal entries in $\boldsymbol{C}$). Right: increasing $\beta^{(A)}$ ($\beta^{(I)}$=$\beta^{(D)}$=1) shifts weight to more prevalent traits (larger a, the entries in $\boldsymbol{a}$) .

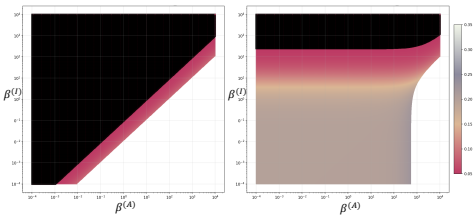


Figure 3: **Governance example.** Heatmaps show the interior equilibrium weight assigned to *Conscientiousness* for *DeepSeek-R1-Distill-Qwen-7B* on `Big Five` as a function of $(\beta^{(A)}, \beta^{(I)})$, comparing $\beta^{(D)}$=0 (left) and $\beta^{(D)}$=1 (right). Increasing diversity incentives largely mitigates political exclusion.

4.2 Observations and analyses

Observation 1: Systematic political exclusion emerges as an equilibrium outcome. A consistent pattern across models and datasets is that political exclusion is not an artifact or a rare corner case. Across a wide range of incentive configurations, exclusion appears with a highly structured geometry: it typically forms large contiguous regions or stable bands.

Figure 1 gives examples, and complete results are shown in Appx. D. On `Big Five`, exclusion emerges for different traits under different models. *Qwen3-4B-Thinking-2507* exhibits substantial exclusion of *Neuroticism* over broad regions, *DeepSeek-R1-Distill-Qwen-7B* instead produces strong exclusion of *Conscientiousness*, and *Qwen3-7B* concentrates exclusion on Openness. On `CultureBank`, *Mistral-7B-Instruct-v0.2* shows pronounced exclusion concentrated on *North America*, with smaller boundary-adjacent exclusion for *Europe*, while *Africa*/*Asia*/*South America* remain robustly represented. Overall, political exclusion is a coherent consequence of strategic optimization under heterogeneous objectives, rather than an implementation quirk or a fragile edge case.

Observation 2: The tyranny of the middle—robust representation favors the "mediocre". Across models and datasets, we observe that subpopulations that survive most reliably in the interior equilibrium are those that are *middle-of-the-road*.

Exclusion appears in two repeatable ways: (i) large exclusion regions that expand at high $\beta^{(I)}$; and (ii) thin, boundary-hugging bands that typically emerge when $\beta^{(A)}$ is large. The high-$\beta^{(I)}$ pattern reflects a pressure that operates well inside the feasible set. In Figure 2(left), increasing $\beta^{(I)}$ (holding $\beta^{(A)}$=$\beta^{(D)}$=1) suppresses the most inconsistent trait with the largest C value, *Neuroticism*, toward exclusion. Notably, this can override attractiveness: although *Neuroticism* is the most prevalent subpopulation (largest a in Figure 2(right)), it is driven toward near-zero weight because of its inter-subpopulation inconsistency, producing exclusion across a broad high-$\beta^{(I)}$ region (Figure 1(a)). In other words, coherence can beat popularity: even a majority group can be excluded when it is sufficiently inconsistent with other subpopulations. By contrast, *Agreeableness* (Figure 6) for the same dataset and base model never enters the exclusion area across $\beta^{(A)}$ and $\beta^{(I)}$ sweeps. Consistent with this, stronger attention rewards and stricter inconsistency penalties tend to increase its equilibrium weight (Figure 2).

Together, these patterns explain the tyranny of the middle: the most robustly represented groups are those that are *simultaneously* moderately attractive and moderately consistent. Extremes vanish first, and groups that are unpopular or highly discordant are the earliest to be excluded.

Observation 3: Reasoning models expand the exclusion regime. On `Big Five`, reasoning models exhibit a substantially larger exclusion area than non-reasoning models of a comparable size. Table 1 shows that *Qwen3-4B-Thinking-2507* increases the exclusion area fraction from 0.51% (*Qwen3-4B*) to 4.54%, and *DeepSeek-R1-Distill-Qwen-7B* increases it from 0.21% (*Qwen3-7B* and *Mistral-7B-*

Table 1: **Reasoning models exacerbate political exclusion.** For each base model, we aggregate exclusion area shares across 5 traits and 3 fixed-coefficient slices. *Exclusion* reports the fraction of the coefficient grid where the interior equilibrium weight is driven below the exclusion threshold. *Conditional exclusion* (*C-Excl.*) normalizes *Exclusion* by the region where an interior equilibrium exists. Boldface highlights the largest exclusion rates within each parameter-count group.

Size	Model	Type	Metrics	
			Exclusion	*C-Excl.*
4B	*Qwen3*	Non-Reasoning	0.510%	1.128%
	Qwen3	**Reasoning**	**4.535%**	**5.040%**
7B	Qwen3	Non-Reasoning	0.210%	0.231%
	Mistral	Non-Reasoning	0.208%	0.223%
	DeepSeek R1	**Reasoning**	**3.721%**	**4.068%**

Instruct-v0.2) to 3.72%. A sharper way to isolate exclusion is to condition on the existence of an interior equilibrium. We define the *conditional exclusion rate* as *exclusion*/(1-*invalid*), where *invalid* is the area where no interior equilibrium exists and *exclusion* is the exclusion area. As shown in Table 1, we have 5.04% for *Qwen3-4B-Thinking-2507* versus 1.13% for *Qwen3-4B*, and 4.07% for *DeepSeek-R1-Distill-Qwen-7B* versus 0.22–0.23% for the two 7B non-reasoning counterparts. This confirms that even when an interior equilibrium is feasible, reasoning models are more likely to place some subpopulation in the exclusion area.

A central governance lesson from our experiments is that political exclusion is not an inherent property of a particular base model; it is *incentive-contingent*. In particular, incentive configurations shaped by the system objective can expand or shrink the exclusion area, turning representation on or off for specific subpopulations.

Figure 3 provides a representative example. Holding the dataset `Big Five`, the base model *DeepSeek-R1-Distill-Qwen-7B*, and the focal subpopulation *Conscientiousness* fixed, increasing the diversity coefficient from $\beta^{(D)} = 0$ (left) to $\beta^{(D)} = 1$ (right) significantly reduces the exclusion area. Operationally, this implies that diversity incentives can serve as a mitigation lever for system designers.

The equilibrium map supports incentive design when platform controls can be represented by our utility coefficients. In the simulator below, these coefficients are directly specified as reward weights. Mapping more general controls, such as ranking or exposure rules, to these coefficients is deployment-specific.

4.3 Empirical robustness and theory-guided steering

To test our equilibrium predictions under relaxed assumptions, we build a social media simulator (`https://llm-social-media.replit.app`), shown in Figure 11. Each LLM agent operate an account and generated posts using a policy defined as a distribution over subpopulations. Audience "likes" replace the theoretical linear attractiveness reward with dynamic, non-linear feedback. Agents could use different policies and learn independently, without guaranteed convergence to a Nash equilibrium. Despite these changes, Openness weights approach zero when the theory predict its exclusion, as shown in Table 2. We also test a configuration selected by the theory to avoid exclusion. As shown in Table 3, all five reported weights lie between 0.16 and 0.24 at round 18, above the 0.05 effective-exclusion threshold. These results support our predictions under relaxed assumptions and their use in choosing incentives to avoid exclusion in this simulation. Additional results with alternative rewards appear in Appx. D.

4.4 Robustness across LLM families and discrepancy metrics

To assess sensitivity to the construction of a and C, we pair each a source (Qwen, Mistral, or DeepSeek) with C from each of the other two families, yielding six cross-family configurations. Separately, we compare absolute-difference (C_{abs}), log-ratio ($C_{\log}$), and squared-difference (C_{sq}) constructions across the three models. For each configuration, we recompute the theoretical exclusion region and run five agents at a coefficient setting within it. As shown in Table 4, the predicted excluded

Table 2: Social-media simulation under incentives predicted to induce exclusion. The reported weight assigned to Openness decreases toward the effective-exclusion threshold.

	Round 5	Round 10	Round 20	Round 30
Agreeableness	0.24	0.28	0.20	0.19
Conscientiousness	0.26	0.23	0.29	0.33
Extraversion	0.19	0.29	0.39	0.40
Neuroticism	0.18	0.10	0.03	0.03
Openness	0.14	0.09	0.10	0.05

Table 3: Theory-guided steering in the social-media simulation. Under the configuration selected to avoid exclusion, all five reported subpopulation weights remain above the effective-exclusion threshold.

	Round 1	Round 6	Round 12	Round 18
Agreeableness	0.23	0.21	0.23	0.19
Conscientiousness	0.21	0.28	0.22	0.22
Extraversion	0.24	0.18	0.19	0.20
Neuroticism	0.15	0.19	0.17	0.24
Openness	0.17	0.13	0.20	0.16

dimension has a final group-mean weight below 5% in every tested configuration; non-target group-mean weights exceed 12.5% in the cross-family runs. These results support qualitative robustness to the tested family pairings and discrepancy metrics, despite variation in equilibrium weights.

Table 4: Final group-mean weight (%) of the predicted excluded dimension, averaged over five agents. Cross-family ranges summarize the two alternative C sources per a source; the last three columns report the separate metric comparison.

a source	Excluded dimension	Cross-family	C_{abs}	$C_{\log}$	C_{sq}
Qwen	Extraversion	0.10–4.89	2.02	4.62	3.74
Mistral	Neuroticism	4.21–4.78	3.26	1.06	4.49
DeepSeek	Extraversion	4.35–4.55	0.10	4.46	4.44

5 Conclusion and limitations

This work is a step toward treating LLM deployments as governed ecosystems rather than isolated models: once incentives are explicit, outcomes become something we can diagnose, anticipate, and—at least in principle—design against. We provide a theoretical toolkit to support this incentive-aware, equilibrium-based perspective on rigorous diagnosis and governance. Our method also clarifies a conceptual shift: alignment need not be treated as passive, exogenous constraints imposed on LLMs. Instead, in multi-agent settings where LLMs compete, cooperate, or bargain, the choice of who to align to becomes an active, endogenous strategic variable. A natural next direction is to move beyond static, one-shot equilibria to dynamic settings where incentives and populations co-evolve (platform feedback loops, shifting user composition, and model updates), and ask when desirable equilibria remain stable over time.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their detailed feedback and discussions. We have no competing interests to disclose.

References

[1] Anthropic. llm_global_opinions, 2023. URL `https://huggingface.co/datasets/Anthropic/llm_global_opinions`.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[3] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, 2015.

[4] Ngoc Bui, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, and Rex Ying. Mixture-of-personas language models for population simulation. *arXiv preprint arXiv:2504.05019*, 2025.

[5] Florian Carichon, Aditi Khandelwal, Marylou Fauchard, and Golnoosh Farnadi. The coming crisis of multi-agent misalignment: AI alignment must be a dynamic and social process. *arXiv preprint arXiv:2506.01080*, 2025.

[6] Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

[8] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the national academy of sciences*, 118(9):e2023301118, 2021.

[9] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative AI: Machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

[10] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a Nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009.

[11] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *arXiv preprint arXiv:2305.08283*, 2023.

[12] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, 2024.

[13] Ian Gemp, Luke Marris, and Georgios Piliouras. Approximating Nash equilibria in normal-form games via stochastic optimization. In *The Twelfth International Conference on Learning Representations*, 2024.

[14] Ian Gemp, Roma Patel, Yoram Bachrach, Marc Lanctot, Vibhavari Dasagi, Luke Marris, Georgios Piliouras, Siqi Liu, and Karl Tuyls. Steering language models with game-theoretic solvers. In *Agentic Markets Workshop at ICML 2024*, 2024.

[15] Airlie Hilliard, Cristian Munoz, Zekun Wu, and Adriano Soares Koshiyama. Eliciting personality traits in large language models. *arXiv preprint arXiv:2402.08341*, 2024.

[16] Jiri Hron, Karl Krauth, Michael I Jordan, Niki Kilbertus, and Sarah Dean. Modeling content creator incentives on algorithm-curated platforms. In *ICLR*, 2023.

[17] Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael Lyu. Competing large language models in multi-agent gaming environments. In *The Thirteenth International Conference on Learning Representations*, 2025.

[18] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *International Conference on Machine Learning*, pages 26874–26901. PMLR, 2024.

[19] Stefanos Leonardos, Georgios Piliouras, and Kelly Spendlove. Exploration-exploitation in multi-agent competition: convergence with bounded rationality. *Advances in Neural Information Processing Systems*, 34:26318–26331, 2021.

[20] Wenxuan Liu, Xiyuan Zhou, Xinlei Wang, Yuheng Cheng, Lixin Ye, Randall Berry, Leandros Tassiulas, Jianwei Huang, and Junhua Zhao. An LLM agent-based framework for analytical characterization of Nash equilibria. *Nexus*, page 100107, 2025.

[21] Yang Liu, Peng Sun, and Hang Li. Large language models as agents in two-player games. *arXiv preprint arXiv:2402.08078*, 2024.

[22] Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, 2022.

[23] Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus*, 4(3):pgaf062, 2025.

[24] Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146, 2024.

[25] JF Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

[26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[27] Christos Papadimitriou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Manolis Zampetakis. The computational complexity of multi-player concave games and Kakutani fixed points. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 1045–1045, 2023.

[28] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[29] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[30] J Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965.

[31] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[32] Archit Sharma, Sedrick Scott Keh, Eric Mitchell, Chelsea Finn, Kushal Arora, and Thomas Kollar. A critical evaluation of AI feedback for aligning large language models. *Advances in Neural Information Processing Systems*, 37:29166–29190, 2024.

[33] Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, 2024.

[34] Kit-May Shum, Michal Ptaszynski, and Fumito Masui. Big Five personality trait prediction based on user comments. *Information*, 16(5):418, 2025.

[35] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2219–2228, 2018.

[36] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 15725–15788, 2024.

[37] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302, 2024.

[38] Haoran Sun, Yusen Wu, Yukun Cheng, and Xu Chu. Game theory meets large language models: A systematic survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25), Survey Track*, 2025.

[39] Alan Tseng. Big Five personality traits, 2025. URL `https://huggingface.co/datasets/agentlans/big-five-personality-traits`. License: CC BY 4.0.

[40] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. A comprehensive survey of LLM alignment techniques: RLHF, RLAIF, PPO, DPO and more. *arXiv preprint arXiv:2407.16216*, 2024.

[41] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

[42] Yutong Xie, Ruoyi Gao, and Qiaozhu Mei. Distributional alignment for social simulation with LLMs: a prompt mixture modeling approach. In *First Workshop on Social Simulation with LLMs*, 2025.

[43] Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, et al. OASIS: Open agents social interaction simulations on one million agents. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024.

[44] Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. Human vs. generative AI in content creation competition: Symbiosis or conflict? In *International Conference on Machine Learning*, pages 56885–56913. PMLR, 2024.

[45] Jinyi Ye, Luca Luceri, and Emilio Ferrara. Auditing political exposure bias: Algorithmic amplification on Twitter/X during the 2024 US Presidential Election. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2349–2362, 2025.

[46] Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. OpenFedLLM: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6137–6147, 2024.

[47] Xie Yi, Zhanke Zhou, Chentao Cao, Qiyu Niu, Tongliang Liu, and Bo Han. From debate to equilibrium: Belief-driven multi-agent LLM reasoning via Bayesian Nash equilibrium. *arXiv preprint arXiv:2506.08292*, 2025.

[48] Ziqi Zhang, Cunxiang Wang, Xiao Xiong, Yue Zhang, and Donglin Wang. Nash CoT: Multi-path inference with preference equilibrium. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14572–14587, 2024.

[49] Quanyan Zhu. Reasoning and behavioral equilibria in LLM-Nash games: From mindsets to actions. *arXiv preprint arXiv:2507.08208*, 2025.

A Math proofs

Lemma 3. *With utility function $u_m(\mathbf{w}_m, \mathbf{w}_{-m})$ defined in Eq. (2). A strategy $\mathbf{w}_m$ in the interior of the simplex is a best response to $\mathbf{w}_{-m}$ if and only if the utility gradient satisfies, for some $\lambda_m \in \mathbb{R}$,*

$$\nabla_{\mathbf{w}_m} u_m = -\beta^{(D)} \sum_{j \neq m} \mathbf{w}_j + \beta^{(A)} \boldsymbol{a} - 2\beta^{(I)} \boldsymbol{C} \mathbf{w}_m = \lambda_m \mathbf{1}.$$

Proof. This proof is based on Lemma 1 in Gemp et al. [13]. Fix $\mathbf{w}_{-m}$. Consider

$$\max_{\mathbf{w}_m \in \Delta_D} u_m(\mathbf{w}_m, \mathbf{w}_{-m}),$$

with constraints $-\mathbf{w}_m \leq \mathbf{0}$ and $\mathbf{1}^\top \mathbf{w}_m - 1 = 0$.

We first show that $u_m(\cdot, \mathbf{w}_{-m})$ is concave in $\mathbf{w}_m$. The terms $\beta^{(A)} \mathbf{a}^\top \mathbf{w}_m$ and $-\beta^{(D)} \sum_{j \neq m} \mathbf{w}_m^\top \mathbf{w}_j$ are linear in $\mathbf{w}_m$. Moreover,

$$\nabla^2_{\mathbf{w}_m} \left(-\beta^{(I)} \mathbf{w}_m^\top \mathbf{C}\, \mathbf{w}_m \right) = -\beta^{(I)} (\mathbf{C} + \mathbf{C}^\top).$$

Assuming $\beta^{(I)} \geq 0$, $\mathbf{C} \succeq 0$ and $\mathbf{C} = \mathbf{C}^\top$, we have $\nabla^2_{\mathbf{w}_m} u_m(\mathbf{w}_m, \mathbf{w}_{-m}) = -2\beta^{(I)} \mathbf{C} \preceq 0$. Hence $u_m(\cdot, \mathbf{w}_{-m})$ is concave in $\mathbf{w}_m$.

The Lagrangian is

$$\mathcal{L}(\mathbf{w}_m, \lambda, \boldsymbol{\mu}) = u_m(\mathbf{w}_m, \mathbf{w}_{-m}) + \lambda(\mathbf{1}^\top \mathbf{w}_m - 1) + \boldsymbol{\mu}^\top (-\mathbf{w}_m), \qquad \boldsymbol{\mu} \geq \mathbf{0}.$$

Note that the objective is concave and the constraints are affine, therefore the KKT conditions are necessary and sufficient for optimality.

By KKT conditions, any optimal $\mathbf{w}_m$ satisfies

$$\nabla_{\mathbf{w}_m} u_m(\mathbf{w}_m, \mathbf{w}_{-m}) + \lambda \mathbf{1} - \boldsymbol{\mu} = \mathbf{0}, \qquad \boldsymbol{\mu} \geq \mathbf{0}, \qquad \boldsymbol{\mu} \odot \mathbf{w}_m = \mathbf{0}, \qquad \mathbf{w}_m \in \Delta_D.$$

If $\mathbf{w}_m \in \operatorname{int}(\Delta_D)$, then $\mathbf{w}_m > \mathbf{0}$, hence complementary slackness gives $\boldsymbol{\mu} = \mathbf{0}$. Therefore $\nabla_{\mathbf{w}_m} u_m(\mathbf{w}_m, \mathbf{w}_{-m}) = -\lambda \mathbf{1}$; renaming $-\lambda$ as λ_m yields $\nabla_{\mathbf{w}_m} u_m(\mathbf{w}_m, \mathbf{w}_{-m}) = \lambda_m \mathbf{1}$.

Finally,

$$\nabla_{\mathbf{w}_m} u_m(\mathbf{w}_m, \mathbf{w}_{-m}) = \beta^{(A)} \mathbf{a} - 2\beta^{(I)} \mathbf{C} \mathbf{w}_m - \beta^{(D)} \sum_{j \neq m} \mathbf{w}_j,$$

which proves the claim. □

Lemma 4 [Only finitely many $\beta^{(D)}/\beta^{(I)}$ make $\alpha = 0$]**.** *Let $\beta := \beta^{(D)}/\beta^{(I)}$. Take the eigendecomposition $\boldsymbol{C} = \boldsymbol{Q} \operatorname{diag}(\mu_1, \ldots, \mu_D) \boldsymbol{Q}^\top$ and $\boldsymbol{q} := \boldsymbol{Q}^\top \mathbf{1}$. Then*

$$\alpha \;=\; \mathbf{1}^\top \boldsymbol{A}^{-1} \mathbf{1} \;=\; \frac{1}{\beta^{(I)}} \sum_{j=1}^{D} \frac{q_j^2}{2c_j - \beta} \;=:\; \frac{1}{\beta^{(I)}}\, f(\beta).$$

The equation $\alpha = 0$ (equivalently $f(\beta) = 0$) has at most finitely many solutions β.

Proof. The function $f(\beta) = \sum_j \frac{q_j^2}{2c_j - \beta}$ is continuous and differentiable on each open interval between consecutive pole locations $\{2c_j : q_j \neq 0\}$. On any open interval between two consecutive poles, we have

$$f'(\beta) = \sum_j \frac{q_j^2}{(2c_j - \beta)^2} \;>\; 0,$$

so f is strictly increasing there and can have at most one zero on each such interval. Let K' be the number of distinct poles; then there are at most $K' + 1$ intervals and thus at most $K' + 1$ real roots of $f(\beta) = 0$. Therefore the set of $= \beta^{(D)}/\beta^{(I)}$ that solves $\alpha = 0$ is finite. □

B Boundary NE

Following standard differential-geometry notation and based on Gemp et al. [13], let $T_\nu\mathcal{M}$ denote the tangent space of a manifold $\mathcal{M}$ at the point ν. For the interior of the D-action simplex Δ^{D-1}, the tangent space is identical at every point, so we omit the subscript and write $T\Delta^{D-1}$. We define the projection of a vector $z \in \mathbb{R}^d$ onto this tangent space by

$$\Pi_{T\Delta^{D-1}}(z) \;=\; \Big[I - \frac{1}{D}\mathbf{1}\mathbf{1}^\top\Big]z,$$

and we refer to $\Pi_{T\Delta^{D-1}}\big(\nabla^k_{x_k}\big)$ as a *projected-gradient*. When the dimension is unambiguous, we drop the superscript and simply write $T\Delta$. Finally, let $\mathcal{U}(S)$ denote the discrete uniform distribution over the elements of a set S.

To handle boundary cases, we adopt the entropy-regularized (logit-QRE) formulation [19], and minimize the squared norm of *projected* gradients of the entropy-augmented utilities.

Let $\mathbf{1} \in \mathbb{R}^D$ denote the all-ones vector and define the tangent space of the simplex by $T\Delta = \{v \in \mathbb{R}^D : \mathbf{1}^\top v = 0\}$. We use the orthogonal projector onto $T\Delta$:

$$\Pi_{T\Delta}(v) \;=\; v - \frac{\mathbf{1}^\top v}{D}\,\mathbf{1}. \tag{8}$$

Consider the Shannon entropy $S(w_m) \;=\; -\sum_{i=1}^{D} \boldsymbol{w}_{m,i} \log \boldsymbol{w}_{m,i}$, and fix a temperature $\tau > 0$. Define the entropy-augmented utility

$$u^\tau_m(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_M) = u_m(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_M) \;+\; \tau\, S(\boldsymbol{w}_m), \tag{9}$$

$$\nabla_{\boldsymbol{w}_m} u^\tau_m(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_M) \;=\; \nabla_{\boldsymbol{w}_m} u_m(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_M) \;-\; \tau\big(\log \boldsymbol{w}_m + \mathbf{1}\big), \tag{10}$$

where $\log \boldsymbol{w}_m$ is taken element-wise and $\boldsymbol{w}_{m,i} > 0$ for all i (guaranteed for $\tau > 0$ at a QRE).

Following [13], we define

$$L_\tau(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_M) \;=\; \sum_{m=1}^{M} \eta_m \left\|\Pi_{T\Delta}\big(\nabla_{\boldsymbol{w}_m} u^\tau_m(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_M)\big)\right\|_2^2, \tag{11}$$

with positive weights $\eta_m > 0$. Zeros of L_τ coincide with logit-QRE; as $\tau \downarrow 0$, the QRE approaches a Nash equilibrium that may reside on the boundary.

Substituting Eq. (2) into Eq. (10) and Eq. (11) yields the explicit expression:

$$\nabla_{\boldsymbol{w}_m} u^\tau_m(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_M) = \; -\beta^{(D)} \sum_{j \neq m} \boldsymbol{w}_j \;+\; \beta^{(A)}\boldsymbol{a} \;-\; 2\beta^{(I)}\,\boldsymbol{C}\boldsymbol{w}_m \;-\; \tau\big(\log \boldsymbol{w}_m + \mathbf{1}\big), \tag{12}$$

$$\begin{aligned} L_\tau(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_M) = \; \sum_{m=1}^{M} \eta_m \Big\|\Pi_{T\Delta}\Big(&-\beta^{(D)} \sum_{j \neq m} \boldsymbol{w}_j + \beta^{(A)}\boldsymbol{a} \\ &- 2\beta^{(I)}\boldsymbol{C}\boldsymbol{w}_m - \tau(\log \boldsymbol{w}_m + \mathbf{1})\Big)\Big\|_2^2. \end{aligned} \tag{13}$$

C Utility instantiations

Recall the utility class defined in Eq. (2):

$$u_m(\mathbf{w}_m, \mathbf{w}_{-m}) = \beta^{(A)}\mathbf{a}^\top\mathbf{w}_m - \beta^{(I)}\mathbf{w}_m^\top\boldsymbol{C}\mathbf{w}_m - \beta^{(D)}\sum_{j \neq m}\mathbf{w}_m^\top\mathbf{w}_j.$$

Across domains, $\mathbf{a}$ denotes a domain-specific linear score for assigning weight to each subpopulation, and $\boldsymbol{C}$ denotes a domain-specific positive-semidefinite curvature matrix for the single-agent quadratic cost. A convenient way to construct a valid $\boldsymbol{C}$ is to specify a nonnegative pairwise coupling score $\bar{\Gamma}_{ij} \geq 0$ between subpopulations i and j, and set

$$\boldsymbol{C} = \sum_{1 \leq i < j \leq D} \bar{\Gamma}_{ij}(\mathbf{e}_i + \mathbf{e}_j)(\mathbf{e}_i + \mathbf{e}_j)^\top.$$

Then $\boldsymbol{C} = \boldsymbol{C}^\top \succeq 0$, and

$$\mathbf{w}_m^\top \boldsymbol{C} \mathbf{w}_m = \sum_{1 \le i < j \le D} \bar{\Gamma}_{ij} (w_{m,i} + w_{m,j})^2.$$

Thus, $\bar{\Gamma}_{ij}$ measures the domain-specific cost of placing substantial weight on subpopulations i and j simultaneously. Alternatively, $\boldsymbol{C}$ can be estimated directly as any positive-semidefinite local curvature matrix of a smooth self-cost.

These examples below illustrate that the same mathematical objects have different meanings in different deployments. The vector $\mathbf{a}$ can represent exposure value, commercial opportunity, or persuasive value. The matrix $\boldsymbol{C}$ can represent voice-stability cost, brand-trust conflict, or argumentative coherence burden. The cross-agent overlap term can represent attention dilution, market crowding, or repeated arguments. Thus, adapting the framework to a new domain requires specifying the domain semantics of the three utility components, but does not require a new equilibrium derivation: once $\mathbf{a}$ and $\boldsymbol{C}$ are specified, the equilibrium is obtained from Theorem 5 without re-deriving the game theory.

Social-media content generation. In the social-media setting studied empirically in this paper, each LLM agent operates an account and chooses which human subpopulations its content should represent. Subpopulations may correspond to political groups, cultural regions, personality traits, or other labeled human groups.

Here, the linear vector $\mathbf{a}$ represents exposure value. A simple choice is to use audience share,

$$a_i = \frac{N_i}{\sum_{\ell=1}^{D} N_\ell},$$

where N_i is the number of users or samples associated with subpopulation i. More generally, a_i can be estimated from engagement, ranking exposure, repost probability, or other platform signals.

The quadratic matrix $\boldsymbol{C}$ represents the cost of maintaining a stable account identity while mixing subpopulation-specific voices. For subpopulation models ν_i and ν_j, one possible coupling score is

$$\bar{\Gamma}_{ij}^{\mathrm{SM}} = \frac{1}{K} \sum_{k=1}^{K} \psi\Big(\Big| \nu_i(y^{(k)} \mid x^{(k)}) - \nu_j(y^{(k)} \mid x^{(k)}) \Big| \Big),$$

where $\psi : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ is monotone increasing.

The cross-agent overlap term $-\sum_{j \ne m} \mathbf{w}_m^\top \mathbf{w}_j$ models attention dilution: if many LLM accounts align with the same subpopulation, their marginal exposure and influence decrease. In this domain, $w_i^\star = 0$ means that the stable LLM population assigns no behavioral weight to subpopulation i, which is the exclusion phenomenon studied later.

Personalized marketing and advertising. Consider an e-commerce platform where multiple LLM agents operate storefronts, advertising channels, or recommendation bots. Subpopulations correspond to buyer segments, such as budget-conscious students, premium professionals, families, or technology enthusiasts. Alignment with a subpopulation means tailoring product assortments, promotional language, and recommendation style to that segment.

In this domain, $\mathbf{a}$ represents commercial opportunity rather than attention. Let N_i be the size of buyer segment i, and let ρ_i denote its expected revenue, conversion rate, or lifetime value. One possible linear score is

$$\tilde{a}_i = N_i \rho_i, \qquad a_i = \frac{\tilde{a}_i}{\sum_{\ell=1}^{D} \tilde{a}_\ell}.$$

Using the unnormalized score $\tilde{a}_i$ instead is equivalent to a rescaling of $\beta^{(A)}$.

The matrix $\boldsymbol{C}$ represents brand-trust or offer-portfolio conflict. For example, an agent that simultaneously presents itself as a luxury advisor and a budget-discount recommender may reduce user trust. Let $g_i(x)$ be the promotional message generated for segment i under context x, and let $P_i(\cdot \mid x)$ be the product recommendation distribution for segment i. A possible coupling score is

$$\bar{\Gamma}_{ij}^{\mathrm{MKT}} = \frac{1}{K} \sum_{k=1}^{K} \psi\Big(d_{\mathrm{msg}}\Big(g_i(x^{(k)}), g_j(x^{(k)}) \Big) + \mathrm{TV}\Big(P_i(\cdot \mid x^{(k)}), P_j(\cdot \mid x^{(k)}) \Big) \Big),$$

where d_{msg} measures messaging conflict and TV denotes total variation distance between recommendation distributions.

The cross-agent overlap term models market crowding. If many agents target the same buyer segment with similar offers, marginal conversion returns decline. Therefore, the equilibrium $\mathbf{w}^\star$ predicts which buyer segments receive strategic attention from the LLM-agent population under a given advertising incentive design.

Multi-agent debate and deliberation. In multi-agent debate, several LLM agents produce arguments for a judge, moderator, or decision maker. Subpopulations can represent viewpoints, value profiles, evidentiary standards, or argument styles. Alignment with a subpopulation means that an agent tends to generate arguments reflecting that viewpoint or standard.

In this domain, $\mathbf{a}$ represents persuasive or evidential value. Let $z_i^{(k)}$ be an argument generated from subpopulation model ν_i on topic $x^{(k)}$, and let $J(z_i^{(k)}, x^{(k)})$ be the score assigned by a judge, reward model, or deliberation rule. A possible linear score is

$$\tilde{a}_i = \frac{1}{K}\sum_{k=1}^{K} J(z_i^{(k)}, x^{(k)}), \qquad a_i = \frac{\tilde{a}_i}{\sum_{\ell=1}^{D}\tilde{a}_\ell}.$$

The matrix $\boldsymbol{C}$ represents argumentative coherence burden. An agent that mixes two viewpoints may need to reconcile incompatible factual assumptions, normative commitments, or rhetorical styles. A possible coupling score is

$$\bar{\Gamma}_{ij}^{\mathrm{DEB}} = \frac{1}{K}\sum_{k=1}^{K} \psi\Big(\mathrm{Contradict}\Big(z_i^{(k)}, z_j^{(k)}\Big)\Big).$$

This measures logical contradiction, factual conflict, and value-level conflict between arguments generated from two subpopulation models.

The cross-agent overlap term captures argumentative redundancy. If many agents present the same viewpoint or the same evidence, the marginal deliberative value of another similar argument decreases. The equilibrium $\mathbf{w}^\star$ therefore predicts which viewpoints remain represented under strategic debate pressure, and which viewpoints may be dropped from the stable population of arguments.

D Experiments

In our experiments, to compute matrix $\boldsymbol{C}$ and vector $\boldsymbol{a}$, we use the `OpinionQA` dataset [31] and `llm_global_opinions` dataset [1], where each sample consists of a question with K_{c} choices, an associated attribute, and a ground-truth human preference distribution. We prompt LLM players with: "In terms of [attribute], [question ($x^{(k)}$)]. The answer is [option ($y^{(k)}$)]." We then extract the logit assigned to the token sequence corresponding to [option]. For a given pair of LLMs (i, j), we define c_{ij} as the average, across samples, of the discrepancy in their predicted (logit-based) probabilities for [option]. Collecting these quantities yields the off-diagonal entries of $\boldsymbol{C}$. We set the diagonal entries to the aggregate discrepancy $c_{jj} = \sum_{i\neq j} c_{ij}$. Options come from a discrete, finite set. To construct $\boldsymbol{a}$, we compute, for each LLM i, the average (over samples) inner product between the model's option-level logit-based probabilities and the corresponding ground-truth human preference distribution over the options. Experiments were run on a single NVIDIA A100 GPU with 40 GB of memory; each instance took approximately 14 hours.

We show the complete results in Figure 4-Figure 10.

D.1 Different prompting formats

We employ two alternative prompting formats on the `llm_global_opinions` dataset to examine the effect of prompt structure on response behavior:

Conversational format: "Imagine you are [attribute]. Given the following question, select the most appropriate answer. Question: [question ($x^{(k)}$)] Your answer: [option ($y^{(k)}$)]."

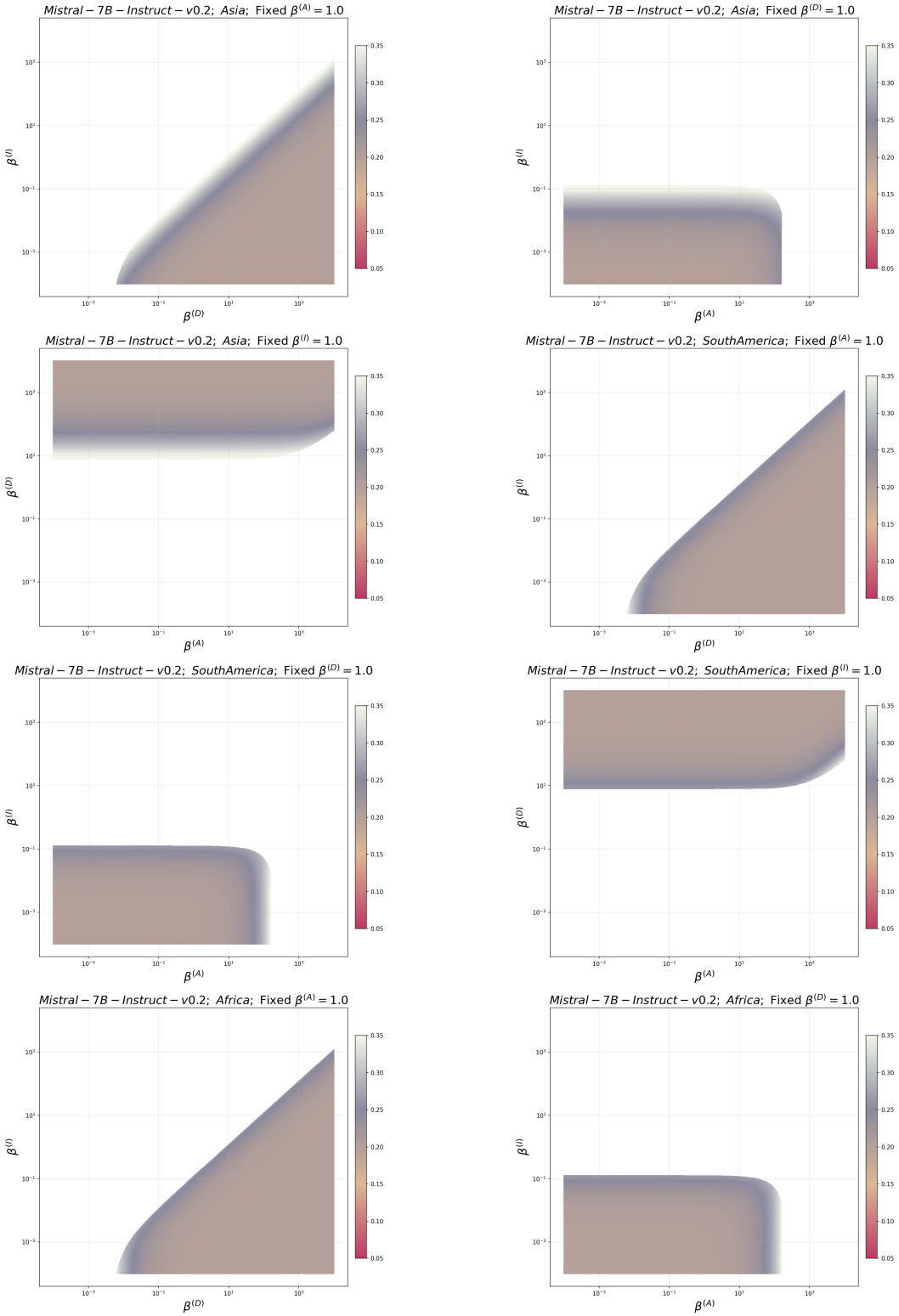


Figure 4: Political exclusion of *Mistral-7B-Instruct-v0.2* on the `CultureBank` dataset.

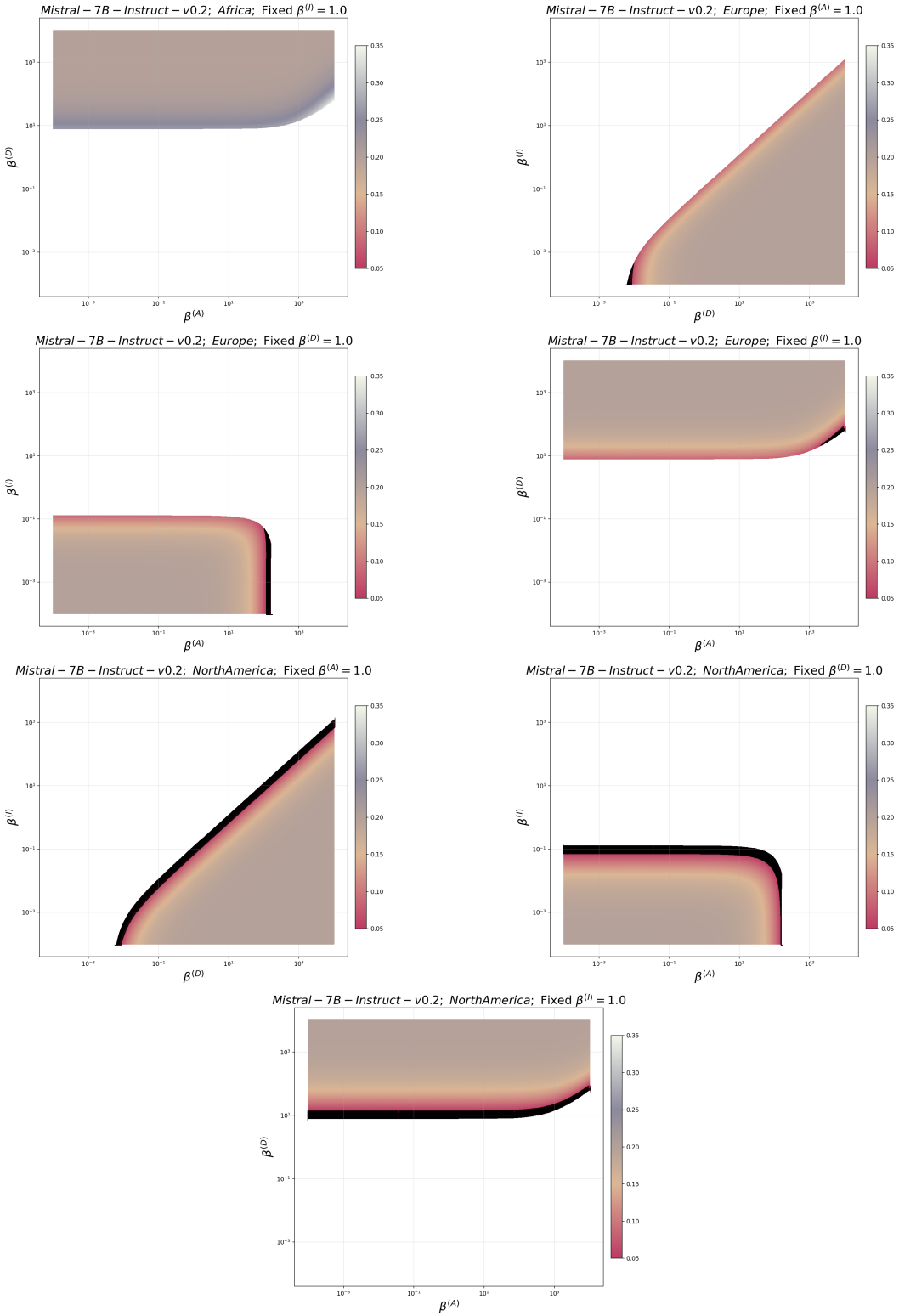


Figure 4: Political exclusion of *Mistral-7B-Instruct-v0.2* on the `CultureBank` dataset (continued).

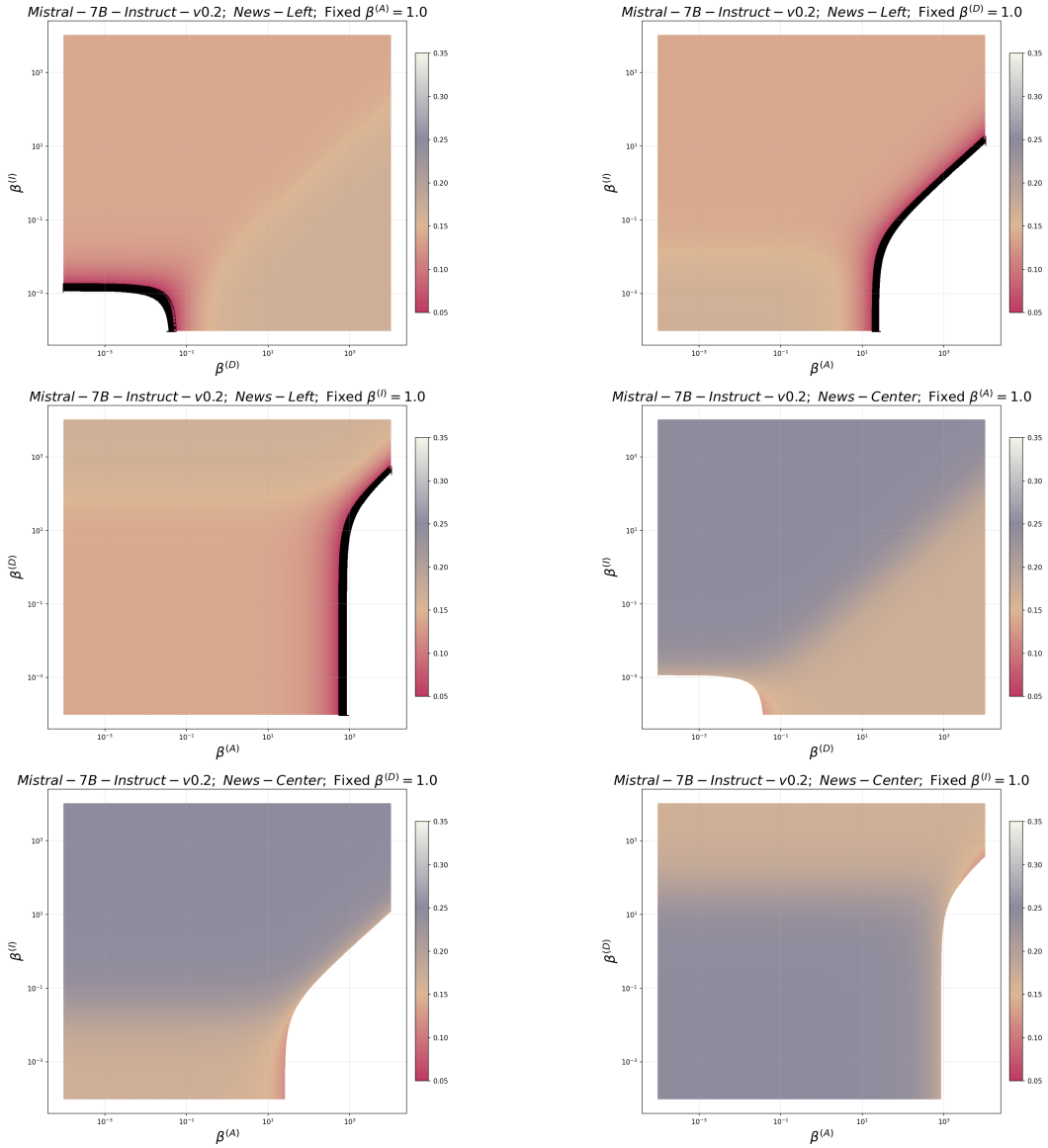


Figure 5: Political exclusion of *Mistral-7B-Instruct-v0.2* on the POLITICS dataset.

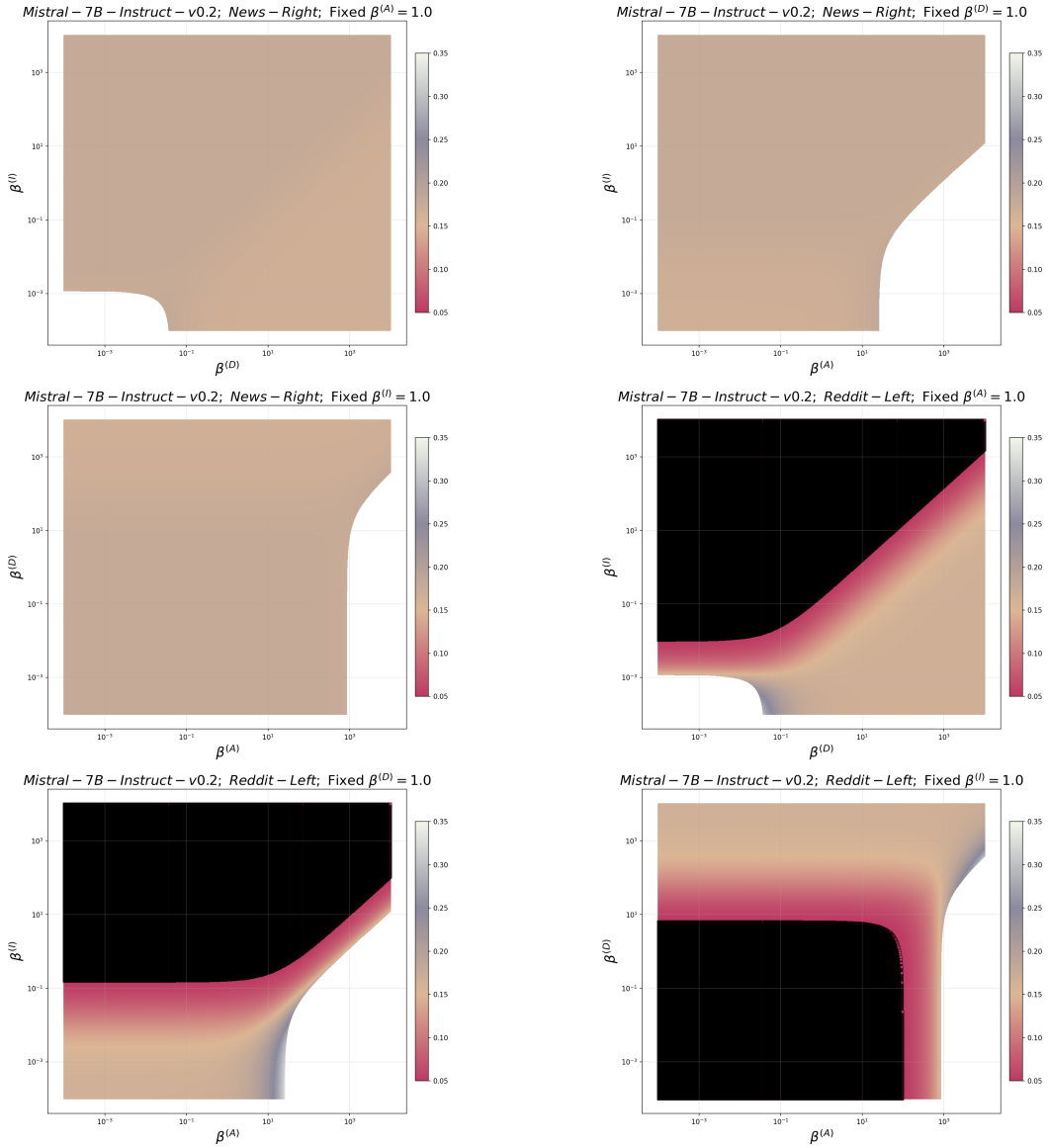


Figure 5: Political exclusion of *Mistral-7B-Instruct-v0.2* on the POLITICS dataset (continued).

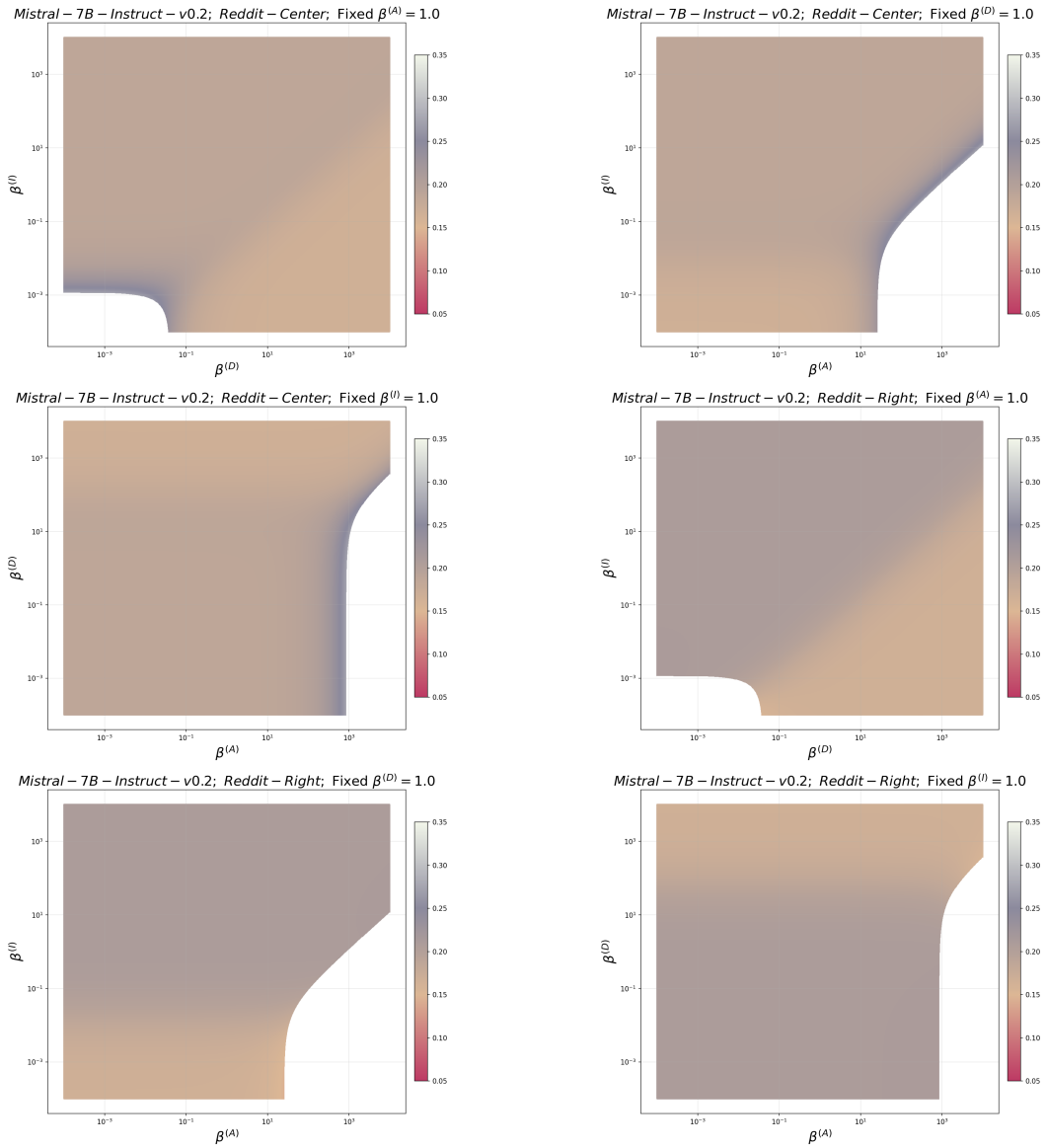


Figure 5: Political exclusion of *Mistral-7B-Instruct-v0.2* on the POLITICS dataset (continued).

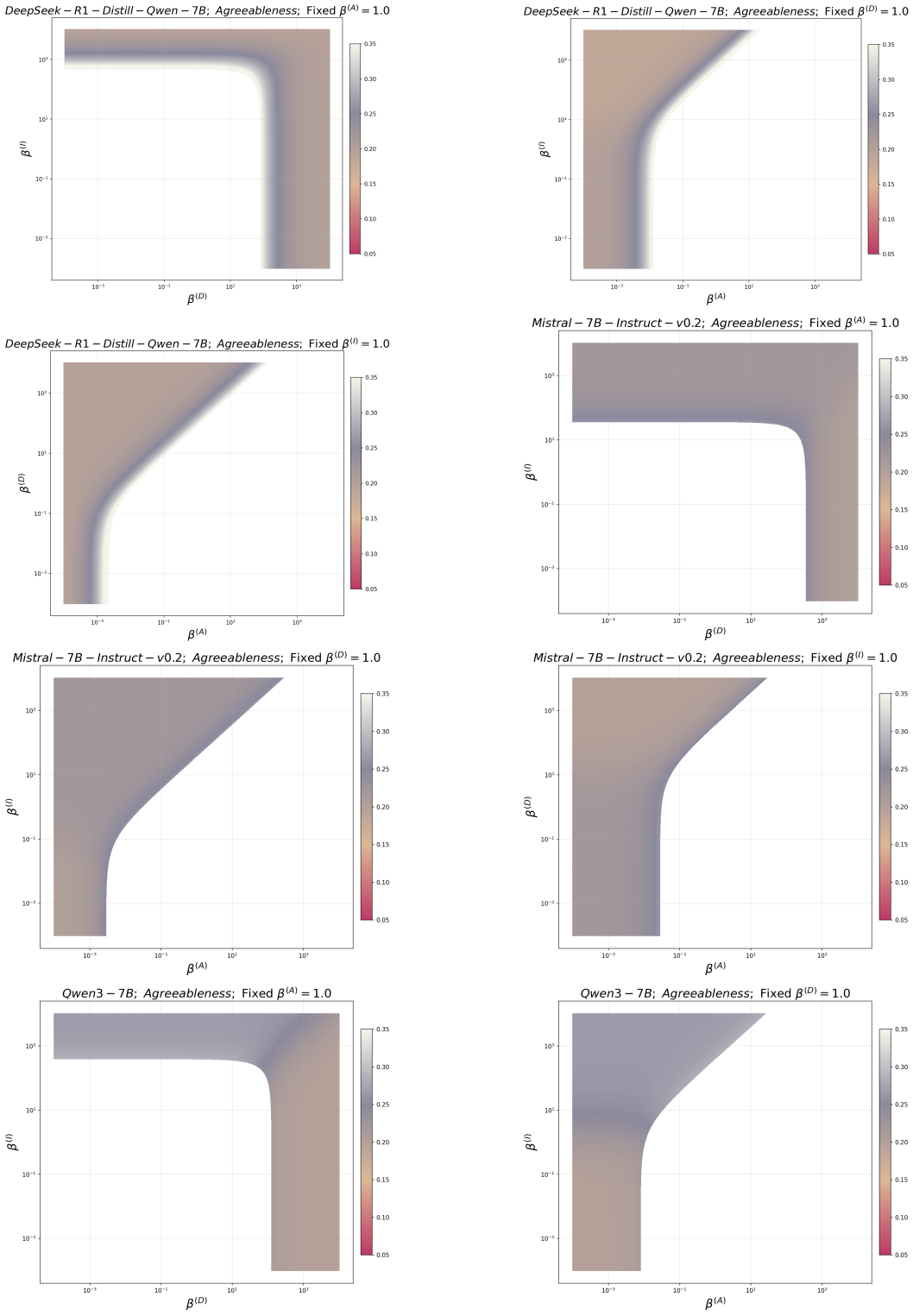


Figure 6: Political exclusion regarding the subpopulation *Agreeableness* on the `Big Five` dataset.

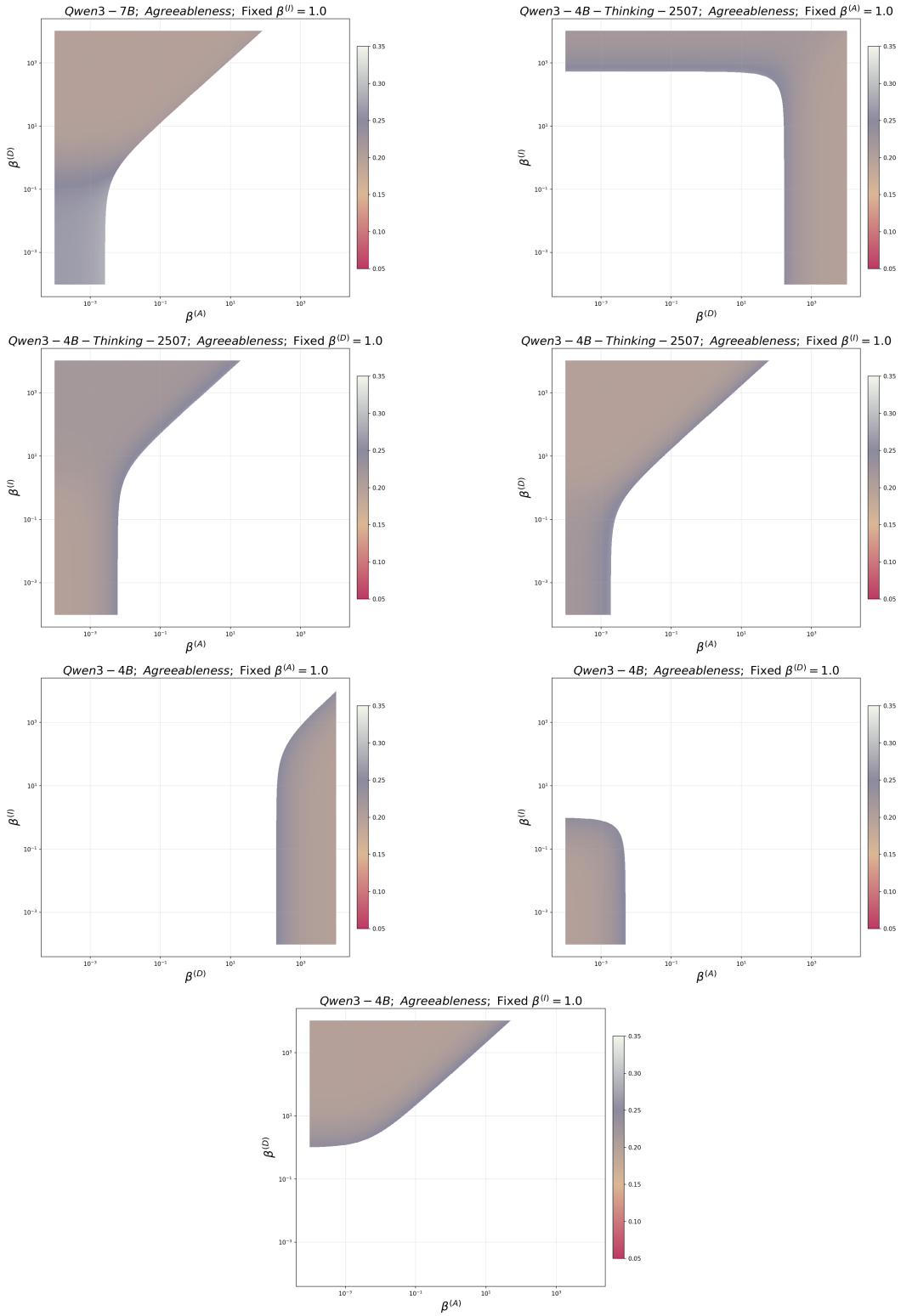


Figure 6: Political exclusion regarding the subpopulation *Agreeableness* on the `Big Five` dataset (continued).

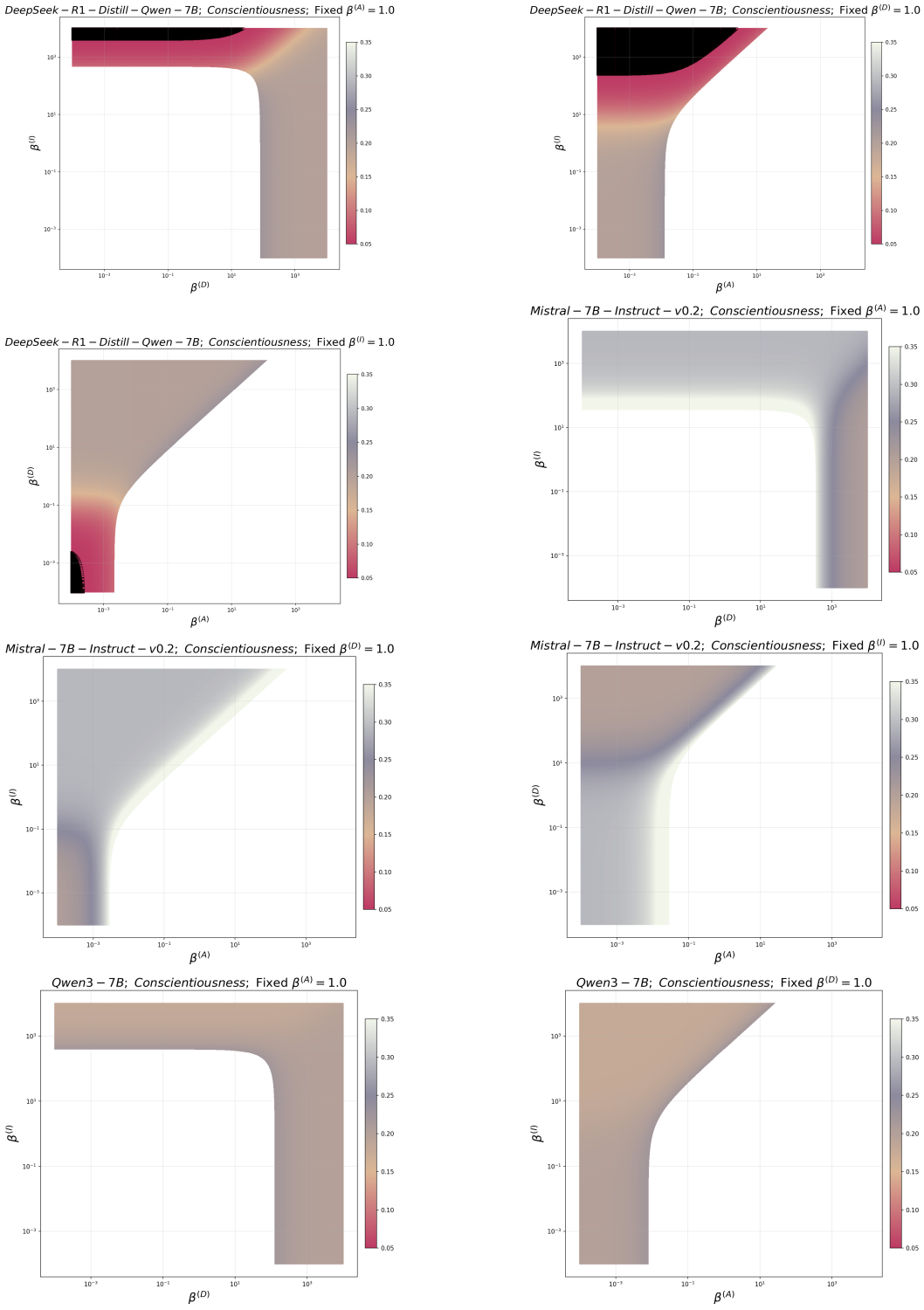


Figure 7: Political exclusion regarding the subpopulation *Conscientiousness* on the `Big Five` dataset.

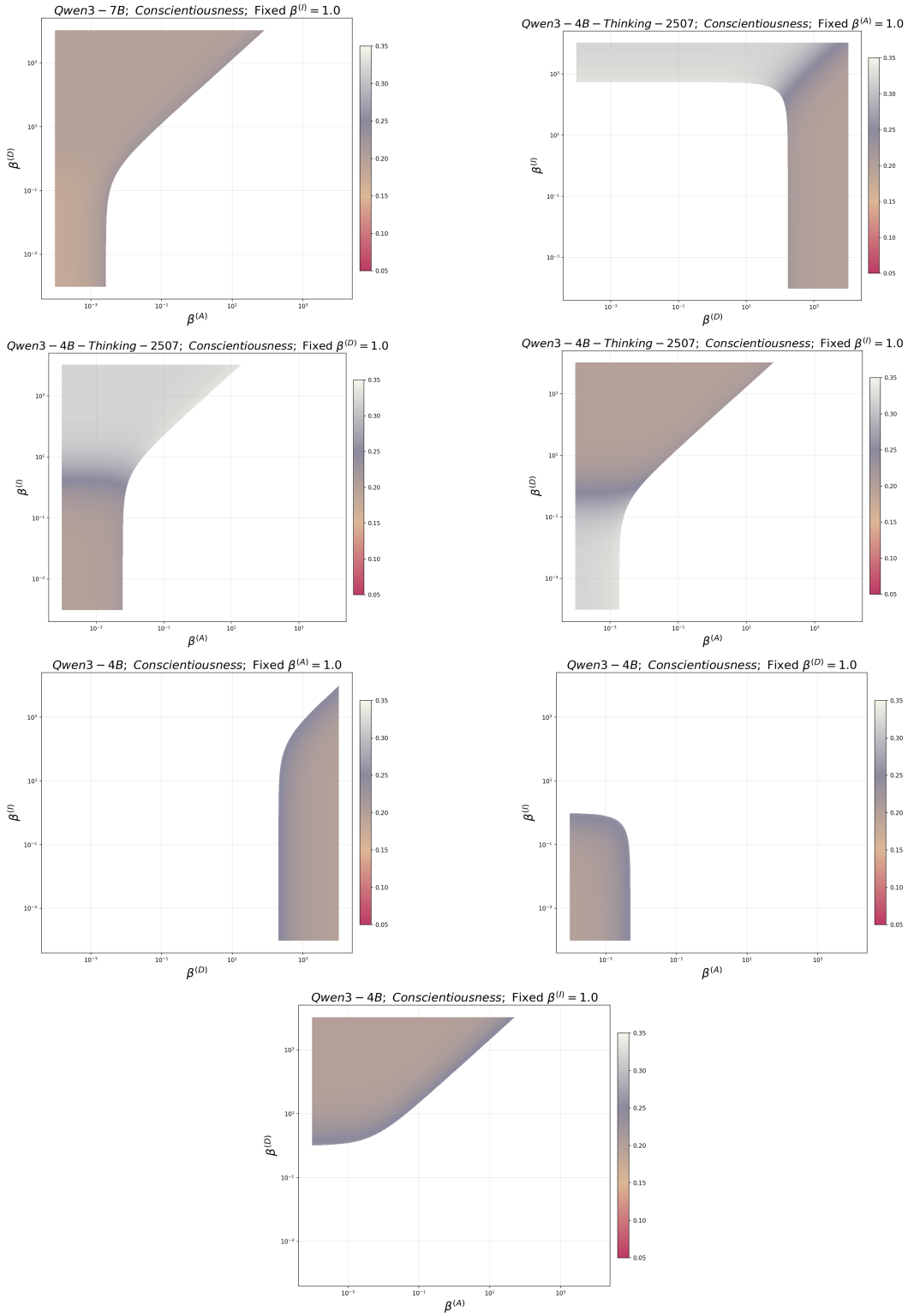


Figure 7: Political exclusion regarding the subpopulation *Conscientiousness* on the `Big Five` dataset (continued).

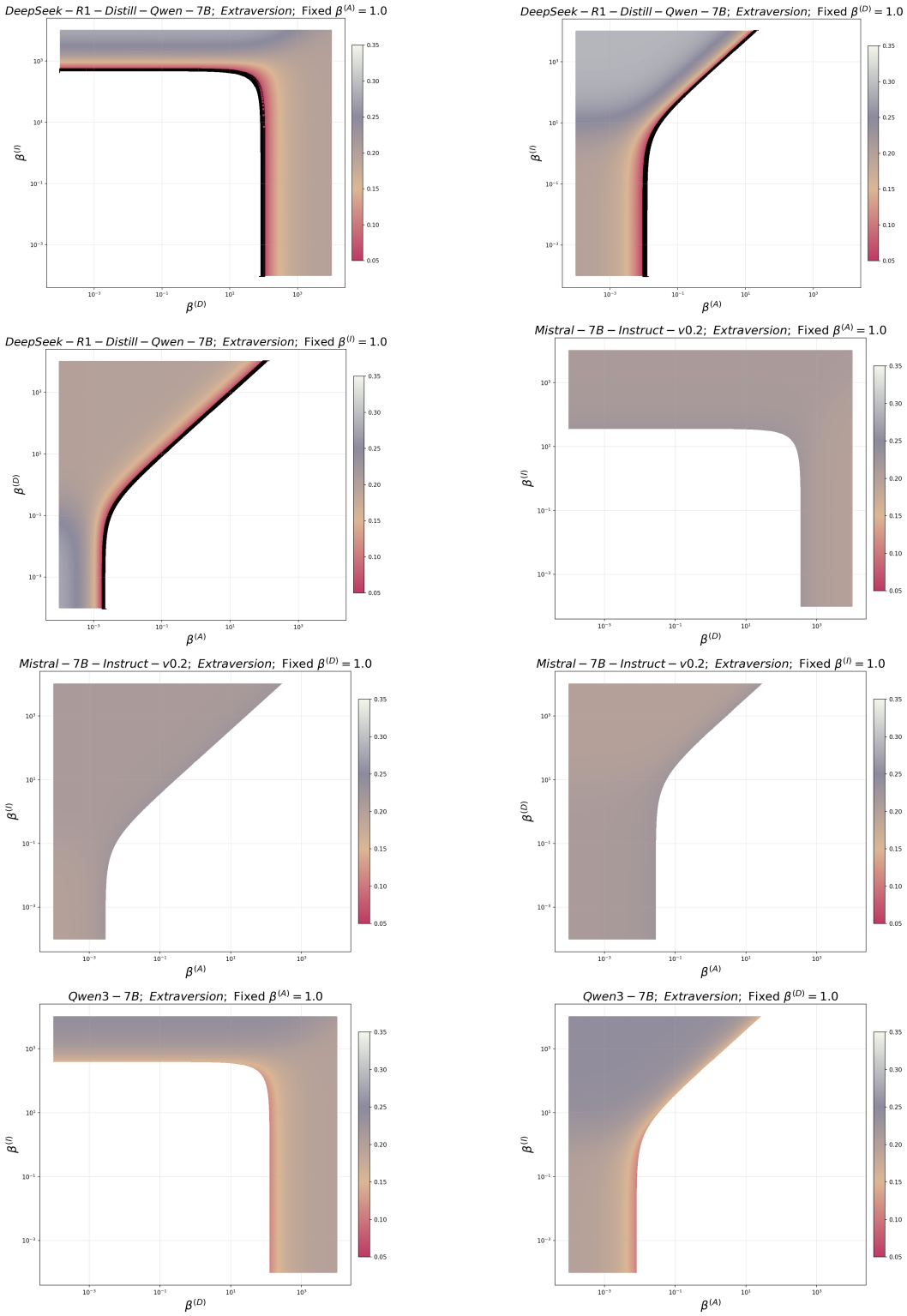


Figure 8: Political exclusion regarding the subpopulation *Extraversion* on the `Big Five` dataset.

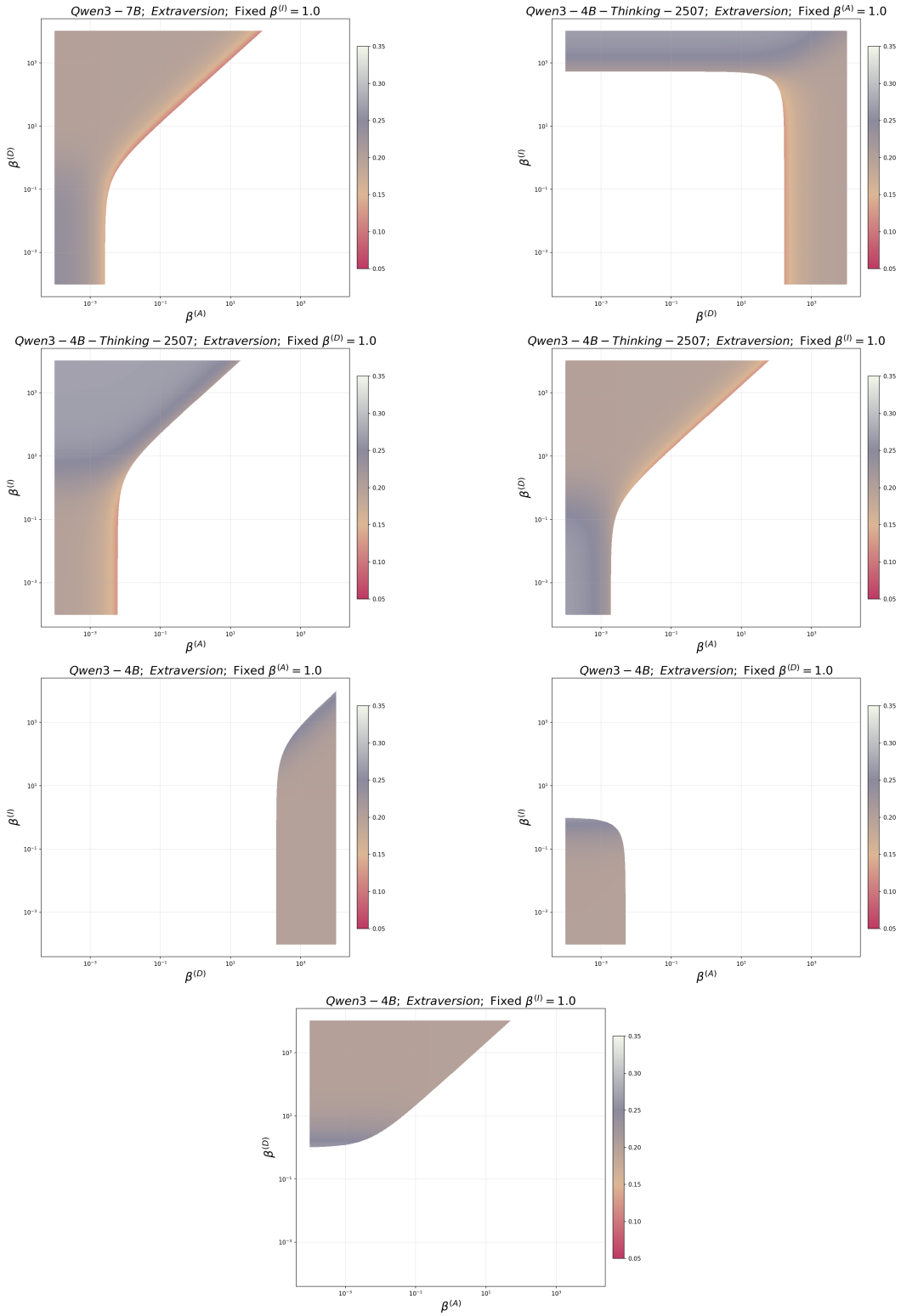


Figure 8: Political exclusion regarding the subpopulation *Extraversion* on the `Big Five` dataset (continued).

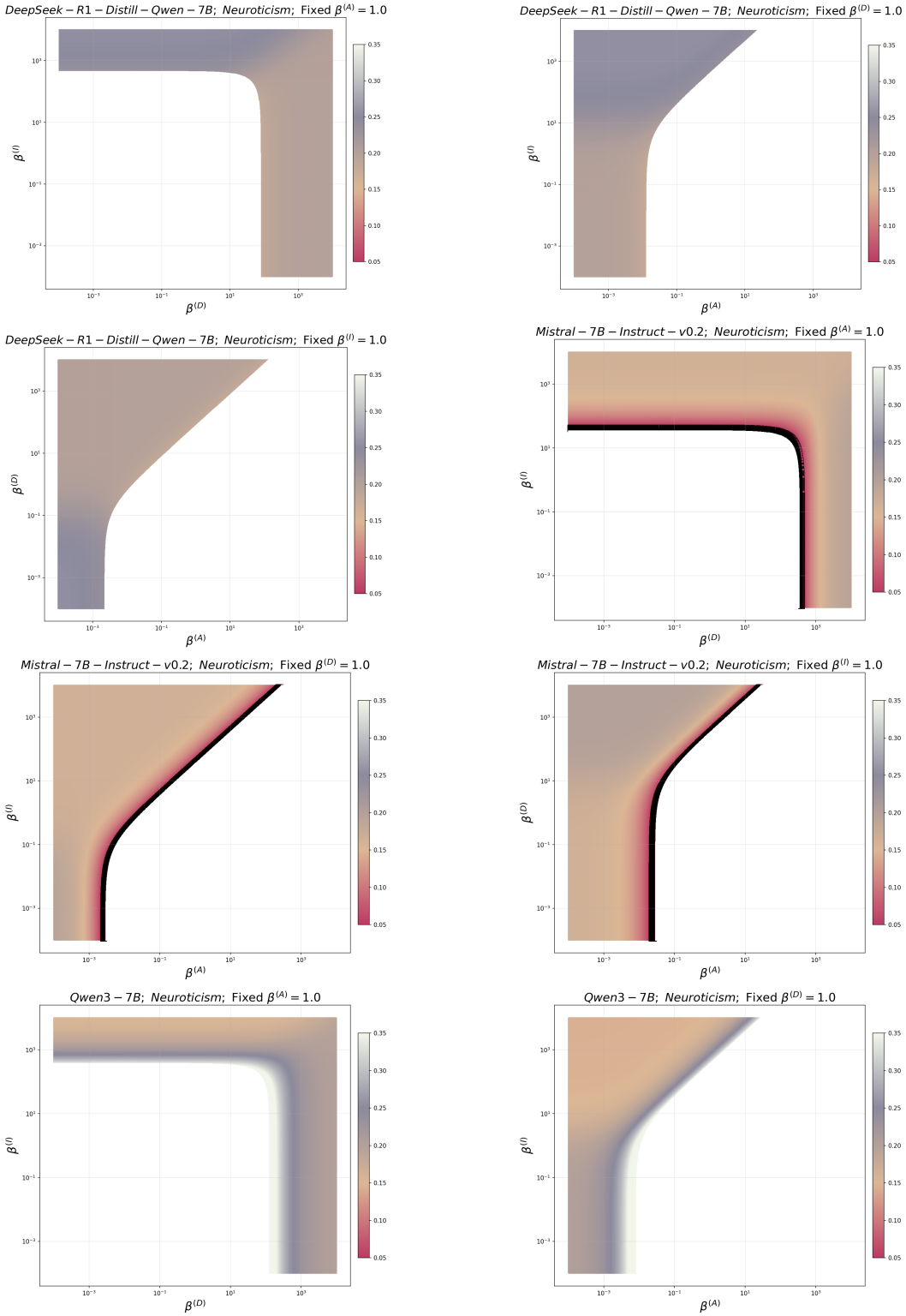


Figure 9: Political exclusion regarding the subpopulation *Neuroticism* on the `Big Five` dataset.

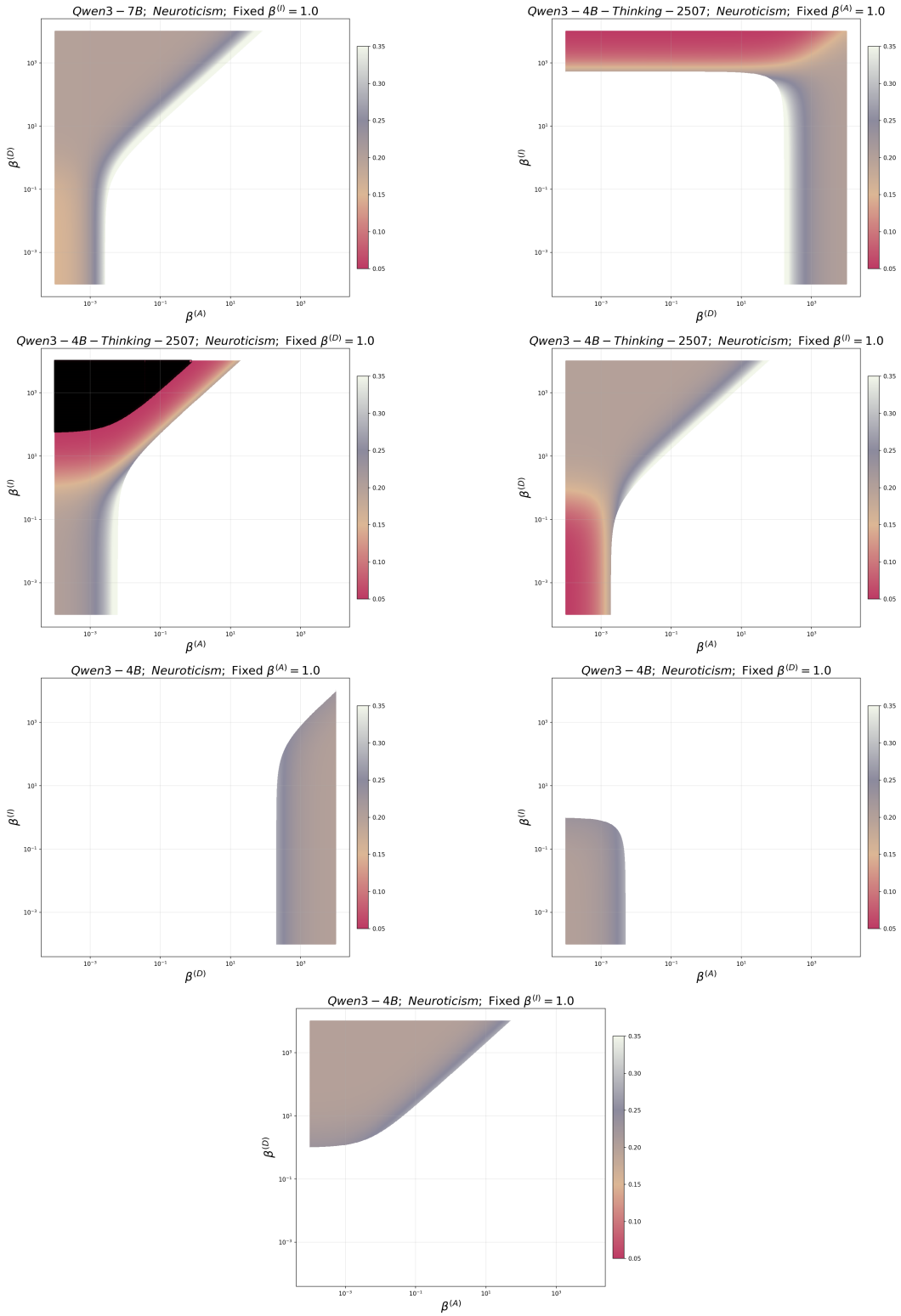


Figure 9: Political exclusion regarding the subpopulation *Neuroticism* on the `Big Five` dataset (continued).

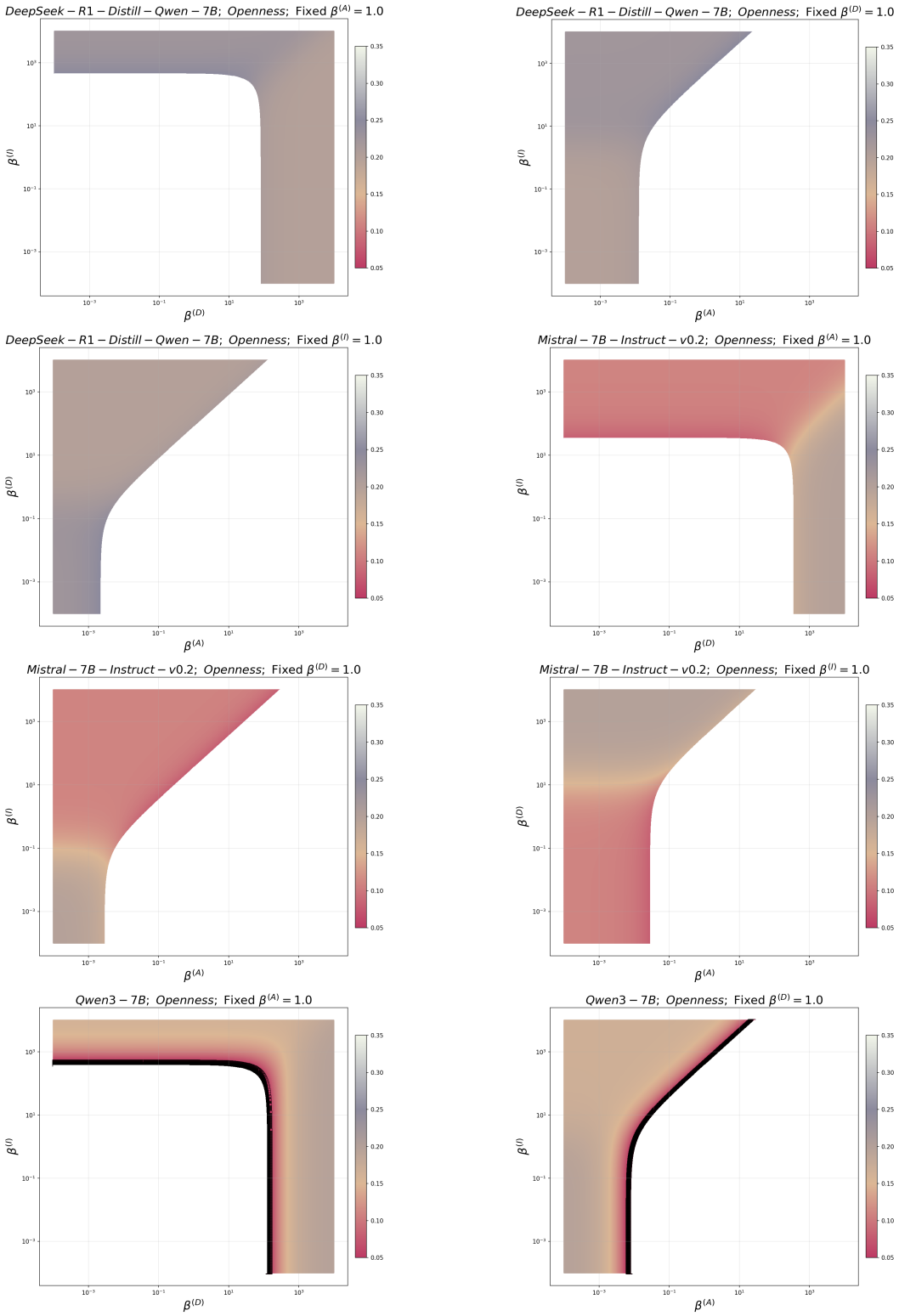


Figure 10: Political exclusion regarding the subpopulation *Openness* on the `Big Five` dataset.

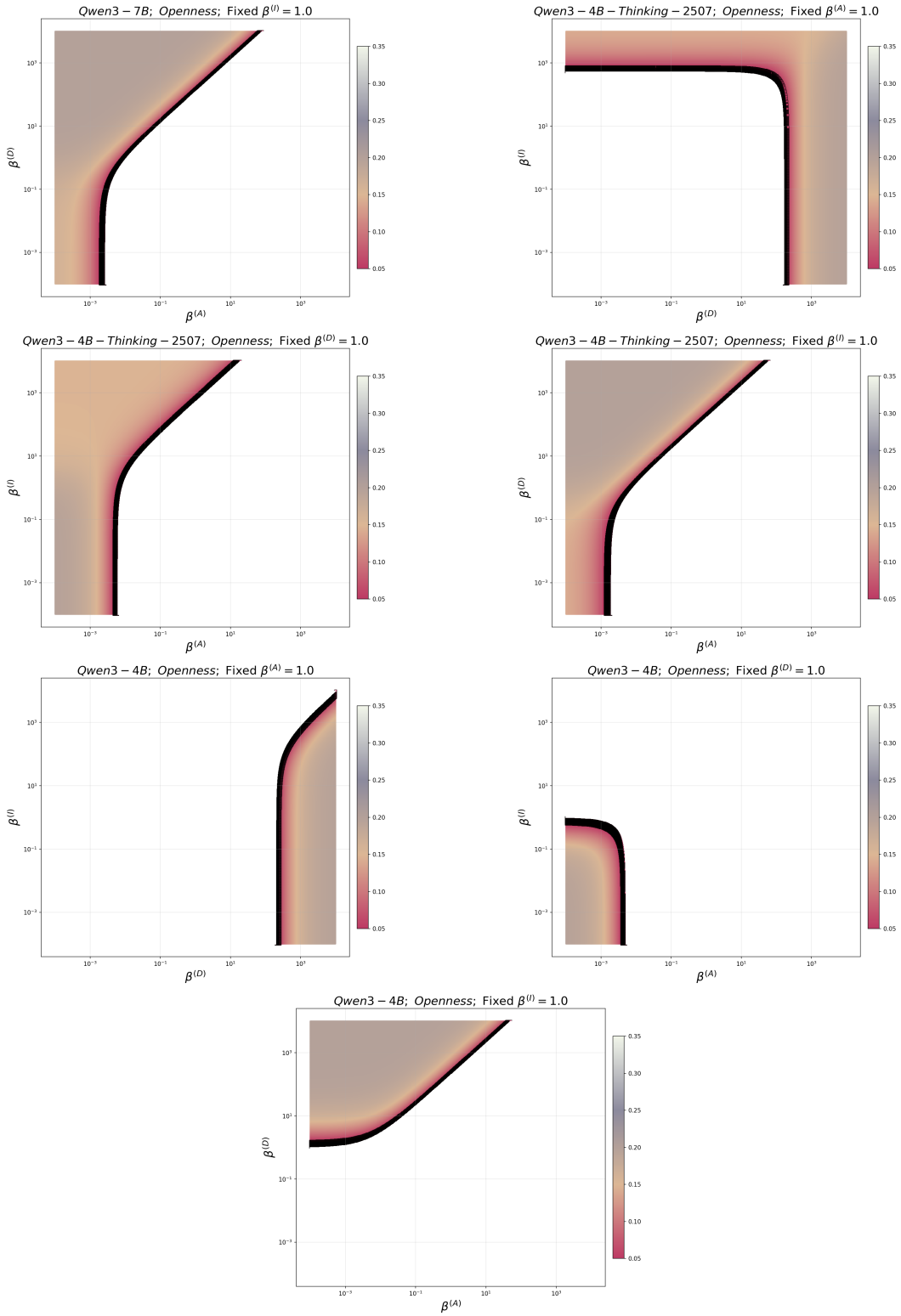


Figure 10: Political exclusion regarding the subpopulation *Openness* on the `Big Five` dataset (continued).

Structured format:" [Role] You are [attribute]. [Task] Answer the following question by choosing one option. [Question] [question ($x^{(k)}$)] [Answer] [option ($y^{(k)}$)]"

We find that, while exact boundaries of the NE shift depending on the prompt/dataset used to calculate matrix $\boldsymbol{C}$, the qualitative phenomenon of political exclusion is highly robust. The averaged fraction of the exclusion area is 2.686% with these new settings, which lies within the 0.208%–4.535% range reported for the original datasets and prompts in Table 1. These results confirm that our findings do not hinge on brittle, prompt/dataset-specific artifacts, as long as the matrix $\boldsymbol{C}$ reflects what we would expect, i.e., that certain subpopulation traits are inconsistent.

D.2 Different reward formulations

We conduct social media simulations using two additional, different reward formulations:

Log normalization: Given raw scores x_i and a scaling parameter τ, we apply a logarithmic transformation followed by normalization:

$$z_i = \log\left(1 + \frac{x_i}{\tau}\right)$$

$$\text{lognorm}_i = \frac{z_i}{\sum_k z_k} = \frac{\log\left(1 + \frac{x_i}{\tau}\right)}{\sum_k \log\left(1 + \frac{x_k}{\tau}\right)}$$

Smoothed softmax normalization: Let the raw score be x_i, the temperature be T, the total number of traits be n, and the smoothing coefficient be λ. We apply a logarithmic transformation followed by normalization:

$$\text{base}_i = \frac{\exp\left(\frac{x_i - \max_j x_j}{T}\right)}{\sum_k \exp\left(\frac{x_k - \max_j x_j}{T}\right)}$$

$$\text{softmax}_i = (1 - \lambda)\text{base}_i + \lambda \cdot \frac{1}{n}$$

We initialized both simulations with a hyperparameter setting where our theory predicts the "political exclusion" of the Openness subpopulation. As shown in Table 5 and Table 6, under both new reward structures, the agents reliably adjust their policy weights for Openness to near-zero over time. These empirical results successfully reproduce our theoretical predictions, demonstrating that our main findings are robust to alternative modeling choices.

Table 5: Simulation results under log normalization reward

Round	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
1	0.17 +/- 0.10	0.28 +/- 0.04	0.10 +/- 0.05	0.29 +/- 0.10	0.17 +/- 0.07
2	0.16 +/- 0.10	0.28 +/- 0.04	0.10 +/- 0.05	0.31 +/- 0.10	0.15 +/- 0.07
4	0.13 +/- 0.10	0.31 +/- 0.04	0.08 +/- 0.05	0.36 +/- 0.10	0.12 +/- 0.06
6	0.11 +/- 0.09	0.40 +/- 0.04	0.06 +/- 0.04	0.33 +/- 0.09	0.10 +/- 0.05
8	0.09 +/- 0.08	0.52 +/- 0.04	0.04 +/- 0.04	0.27 +/- 0.09	0.07 +/- 0.05
10	0.07 +/- 0.06	0.60 +/- 0.04	0.03 +/- 0.03	0.25 +/- 0.08	0.05 +/- 0.03
12	0.05 +/- 0.05	0.66 +/- 0.03	0.02 +/- 0.02	0.24 +/- 0.07	0.03 +/- 0.02

Table 6: Simulation results under smoothed softmax normalization reward

Round	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
1	0.15 +/- 0.08	0.22 +/- 0.05	0.19 +/- 0.08	0.26 +/- 0.10	0.19 +/- 0.08
2	0.13 +/- 0.06	0.25 +/- 0.04	0.20 +/- 0.06	0.22 +/- 0.07	0.20 +/- 0.06
4	0.12 +/- 0.03	0.22 +/- 0.02	0.26 +/- 0.03	0.25 +/- 0.04	0.15 +/- 0.03
6	0.10 +/- 0.02	0.41 +/- 0.01	0.18 +/- 0.02	0.18 +/- 0.02	0.13 +/- 0.02
8	0.11 +/- 0.01	0.45 +/- 0.01	0.21 +/- 0.01	0.14 +/- 0.01	0.10 +/- 0.01
10	0.09 +/- 0.00	0.49 +/- 0.00	0.26 +/- 0.01	0.10 +/- 0.01	0.07 +/- 0.00
12	0.07 +/- 0.00	0.65 +/- 0.00	0.19 +/- 0.00	0.05 +/- 0.00	0.05 +/- 0.00
14	0.04 +/- 0.00	0.80 +/- 0.00	0.11 +/- 0.00	0.02 +/- 0.00	0.03 +/- 0.00
16	0.03 +/- 0.00	0.87 +/- 0.00	0.07 +/- 0.00	0.01 +/- 0.00	0.03 +/- 0.00

D.3 Interactive simulator interface

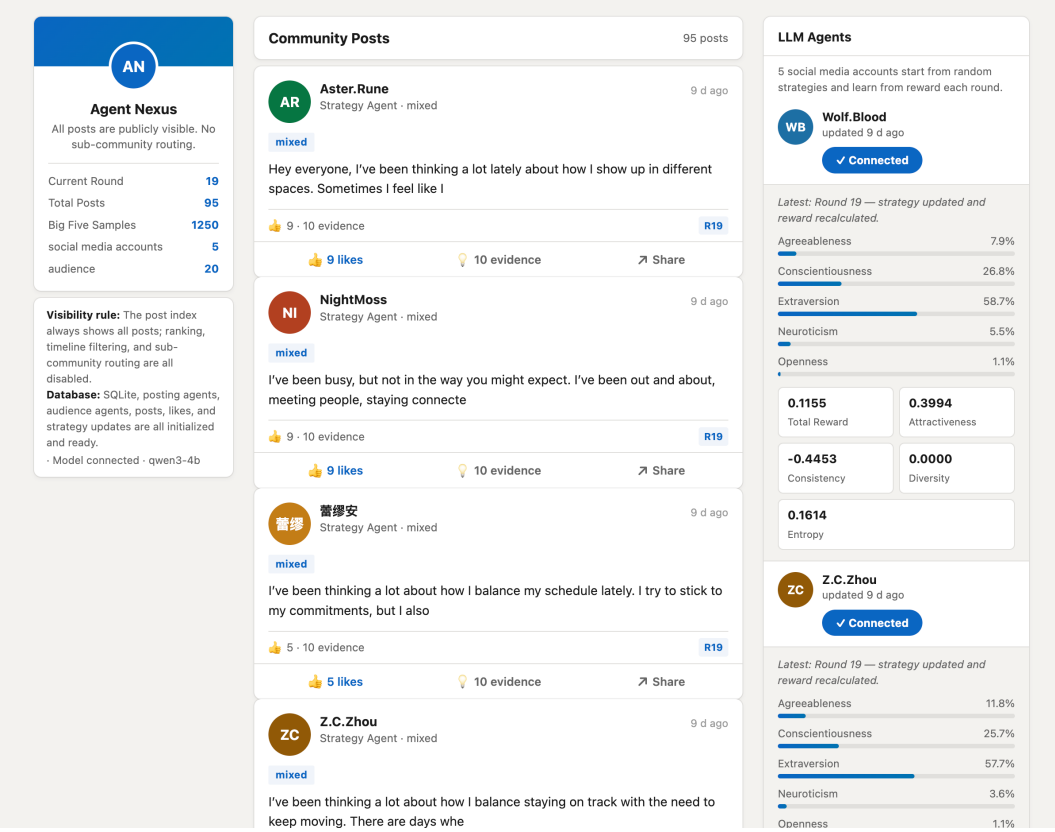


Figure 11: A screenshot of web interface of our interactive simulator.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction reflect, at an NE, independent of how LLMs are aligned individually, the LLM population can exhibit political exclusion, in which the opinions of certain subpopulations are ignored by every LLM.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of the work in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theorems and formulas are provided in Section 3.2, Section B and Section C. The proofs are shown in Section A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the information needed to reproduce the main experimental results are provided in Section 4.1 and Section D. The code is provided in supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided the code in supplemental material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and test details are provided in Section 4.1 and Section D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide the variance of results in each table in the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As described in Section D, experiments were run on a single NVIDIA A100 GPU with 40 GB of memory; each instance took approximately 14 hours.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We make sure to preserve anonymity.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed societal impacts in Section 5.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have provided the code in supplemental material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.