PERSONALIZATION AIDS PLURALISTIC ALIGNMENT UNDER COMPETITION

Natalie Collina, Surbhi Goel, Aaron Roth & Mirah Shi
University of Pennsylvania

ABSTRACT

Can competition among misaligned AI providers yield aligned outcomes for a diverse population of users, and what role does model personalization play? We study a setting where multiple competing AI providers interact with multiple users who must make downstream decisions but differ in preferences. Providers have their own objectives over users' actions and strategically deploy AI models to advance them. We model the interaction as a Stackelberg game with multiple leaders (providers) and followers (users): providers commit to conversational policies, and users choose which model to use, how to converse, and how to act. With user-specific personalization, we show that under a Weak Market Alignment condition, every equilibrium gives each user outcomes comparable to those from a perfectly aligned common model—so personalization can induce pluralistically aligned outcomes even with self-interested providers. In contrast, when providers must deploy a single anonymous policy, there exist equilibria with uninformative behavior under the same condition. We then give a stronger alignment condition that guarantees each user their optimal utility in the anonymous setting.

1 INTRODUCTION

AI alignment is often framed as a single-developer problem: how to encode a particular set of goals and values into a model. A core challenge is that users disagree about what "alignment" should mean. This has motivated work on *pluralistic* alignment guarantees for a single trained model (Sorensen et al., 2024; Shirali et al., 2025; Huang et al., 2024; Klassen et al., 2024). We take a broader view and ask when the *marketplace* for AI models itself can induce pluralistic alignment for a diverse population of users.

Crucially, providers may be misaligned with many of their users and cannot be relied on to align in isolation. For example, an AI system produced (or sponsored) by a drug manufacturer for diagnosis and treatment may be incentivized to recommend its own products, diverging from the goals of doctors using the system. Technical advances in alignment do not resolve this incentive mismatch if the provider has little reason to align to the user in the first place.

But providers do not operate in a vacuum: users can choose among competing models. Collina et al. (2025b) study competition with a single downstream user and show that even if every provider is substantially misaligned, if providers are sufficiently *differently* misaligned—in the sense that the user's utility lies in the non-negative span of providers' utilities—then competition forces perfect alignment in equilibrium. They explicitly leave pluralistic alignment with multiple downstream users as an open question.

We extend and generalize Collina et al. (2025b) to many providers and many users, each with different objectives. Providers commit to AI systems modeled as arbitrary *conversation rules* mapping private information and a conversation prefix to a next message. Users then choose which model to consult, how to interact, and what action to take. Because different users can select different models and use them differently, pluralistic alignment may emerge endogenously; we study when it does.

A central distinction in our framework is *personalization* versus *anonymity*. Personalized providers can commit to different conversation rules for different users, capturing account-based systems with memory and profiling (e.g., via browser history). Anonymous providers must deploy a single rule for all users, modeling open-weight models and anonymous API access.

We give a simple market alignment condition—roughly, a decomposition across users and providers of the condition in (Collina et al., 2025b)—under which, in *all* equilibria of the personalized commitment game, every user attains the benefits of a fully aligned model, despite heterogeneous user utilities and provider misalignment. In contrast, in the anonymous game, even under the same condition, alignment need not obtain: we construct an example where a Nash equilibrium yields *no* non-trivial utility from using the deployed models. Thus, personalization can be *helpful* for pluralistic alignment, even when no provider is aligned with any user. Finally, we introduce a stronger, in a sense dual, alignment condition—that there exists a subset of providers whose utilities lie in the non-negative span of their users' utilities—which is strictly stronger than our initial condition. Under this stronger condition (and additional structural assumptions), pluralistic alignment is guaranteed in all equilibria even when providers must deploy anonymous models.

1.1 Our model and results

We study a marketplace where multiple AI providers commit to conversational policies (as a function of private information), and downstream users choose whom to consult and how to use the interaction to select an action. We formalize this as a multi-leader, multi-follower Stackelberg game: providers commit first; each user then best-responds by choosing a provider, interacting, and acting. Our key distinction is whether providers can condition on user identity. In the *personalized game*, each provider commits to a potentially different policy for each user, and users observe only their own available policies. In the *anonymous game*, each provider commits to a single publicly observable policy used for all users. Informally, these capture account-based personalization versus anonymous/open-access deployment.

A benchmark based on common information. Since providers may hold different private information, not all policies are feasible for all providers. We therefore benchmark welfare using only providers' *common information*. We formalize this via *shared conversation rules* and define each user's optimal shared rule $C^*_S(i)$. Our theorems lower-bound equilibrium utility by $u_i(C^*_S(i))$, i.e., the best utility user i could achieve from a perfectly aligned provider informed only by what all providers commonly know.

Market-level alignment conditions. We introduce two structural assumptions linking provider and user objectives. *Weak Market Alignment* (Definition 2.9) requires providers' utilities to be (approximately) additively separable across users, with each user's utility (approximately) a nonnegative combination of the corresponding components. *Strong Market Alignment* (Definition 2.10) requires each provider's utility to be (approximately) a nonnegative combination of users' utilities; we show this implies Weak Market Alignment. In this framework, we prove several results:

- In the personalized game, competition plus Weak Market Alignment suffices to guarantee that *every* user receives near-benchmark utility in *every* Nash equilibrium (Theorem 3.1). In this sense, personalization "decomposes" competition across users: even if no provider is globally aligned, equilibrium behavior must still be approximately user-optimal for each user.
- In contrast, under the same Weak Market Alignment condition, the anonymous game can have equilibria in which all users receive very low utility: we construct such an example where equilibrium reveals essentially no useful information (Theorem 4.1). This captures a core tension of anonymous/public policies: a provider may be pushed to withhold information to avoid outcomes it dislikes for some users, even when that hurts all users.
- We then study when anonymous competition can still deliver strong user guarantees. Under Strong Market Alignment plus an additional assumptions on the marginal value of an additional round of conversation, we prove anonymous-game utility guarantees (Theorem E.1). We also prove a clean special case: if each relevant provider has an *user-dominant* conversation rule (e.g., when it is possible for the providers to fully reveal their private observations), then each user attains near-benchmark utility in every anonymous-game equilibrium (Theorem 4.3).
- Finally, we study how guarantees change when a new user with arbitrary preferences — possibly breaking the alignment conditions — enters the market and perturbs providers' utilities. In the personalized setting, the original users' guarantees persist (Theorem A.1); in the anonymous setting, we show that adding a user can break previously good guarantees (Theorem A.2).
- We empirically test our alignment conditions (Section 5) using OpinionQA Santurkar et al. (2023), where demographic groups are users and LLMs are providers. Weak Market Alignment holds: user utilities are well-approximated by nonnegative combinations of provider utilities, improving

with market size, though some groups are harder to align than others. Strong Market Alignment is more demanding: single-provider coverage is poor, but adding one-two more providers dramatically improves it. However, as theory predicts, approximation errors for Strong alignment are roughly 50% larger than for Weak.

1.2 Related Work

Bayesian Persuasion and Economic Competition. Bayesian Persuasion was introduced by Kamenica & Gentzkow (2011) — in the canonical model, there is a single informed "sender" and an uninformed "receiver" who share a common prior. The sender commits to a "signaling scheme", which is a mapping from observations to messages sent to the receiver, who conditions on the message and takes their best response action under their posterior. We adopt the basics of this model, but extend it by allowing that both parties be differently informed, and that interaction involve a multi-round conversation rather than a single message. Multi-sender Bayesian Persuasion was introduced by Gentzkow & Kamenica (2016) and studies the standard Bayesian Persuasion model with multiple senders who simultaneously commit to a signaling scheme. Subsequently a number of papers have studied multi-sender Bayesian Persuasion (Gentzkow & Kamenica, 2017; Li & Norman, 2018; Au & Kawai, 2020; Wu, 2023). There is also work studying multiple *receiver* Bayesian persuasian, but with a single sender. This work highlights a fundamental tension: when restricted to public signals (Alonso & Câmara, 2016), a single sender must often degrade information quality to balance conflicting receiver preferences—a dynamic we recover in our 'Anonymous' setting. Conversely, when private signals are allowed (Arieli & Babichenko, 2019), a single sender can tailor disclosures to exploit each receiver individually. Our work extends this to the multi-sender setting, showing that competition resolves these failures: in a personalized market, competition prevents senders from exploiting the privacy of the channel, forcing them instead to reveal full information to capture the user. Our model has multiple participants on both sides of the market, and is a direct extension of the model of Collina et al. (2025b) from a single user to many users. Our finding that personalized competition leads to user-optimal outcomes parallels the classic result in economic theory by Thisse & Vives (1988), who showed that competitive price discrimination forces firms into a Prisoner's Dilemma that drives prices to marginal cost—improving outcomes for consumers.

AI Alignment. Our work fits broadly into the study of multi-agent AI systems (Guo et al., 2024). We present a game theoretic approach in which "alignment" emerges from the competitive interaction of many mis-aligned agents. Recent work has explored cooperative multi-agent approaches to AI safety, where multiple AI systems work together to improve alignment outcomes. Constitutional AI (Bai et al., 2022) uses AI feedback to train more helpful and harmless models, with one AI system providing critiques and revisions of another's outputs. Similarly, approaches using AI systems to evaluate and improve other AI systems (Leike et al., 2018) rely on cooperative dynamics where the evaluating system is assumed to be sufficiently aligned to provide useful feedback. These approaches typically assume that at least some components of the multi-agent system are well-aligned or that the agents share compatible objectives. Our work differs by studying strategic rather than cooperative multi-agent settings.

Several recent papers with alignment motivations (Collina et al., 2025a; 2026; Nayebi, 2025) have studied *agreement protocols* through which conversational agents can come to agreement about their beliefs through short interactions. These should be viewed as protocols for cooperative agents, as they are assumed to express their true beliefs at each iteration. We adopt the conversational framework of these papers but like (Collina et al., 2025b) study *strategic* agents who do not have the same goals. Our work can likewise be viewed as a strategic generalization of the agreement literature (Aaronson, 2005; Frongillo et al., 2021; Collina et al., 2026; 2025a; Nayebi, 2025; Kearns et al., 2026).

2 Preliminaries

We study a strategic market of AI providers who interact with many downstream users. Each user is trying to gain information in order to make an optimal decision. AI providers aim to steer users towards actions that they prefer.

Players and information. There is a finite set of *users* $N = \{1, \ldots, n\}$ and a finite set of *AI providers* $K = \{1, \ldots, k\}$. A common state $y \in \mathcal{Y}$ is drawn along with private features $(x_1^U, \ldots, x_n^U, x_1^P, \ldots, x_k^P)$ from a common prior[1] $\phi \in \Delta(\mathcal{Y} \times \prod_{i \in N} \mathcal{X}_{U_i} \times \prod_{j \in K} \mathcal{X}_{P_j})$. Each user i observes x_i^U and each provider j observes x_j^P. The private information of each provider captures the vast amount of data used to train AI models and their ability to parse large numeric datasets that may be illegible to human users; meanwhile, the private information of each user captures the individual expertise and observations of users interacting with AI systems.

Actions and utilities. Each user i has a utility function $u_i^U : \mathcal{A}_i \times \mathcal{Y} \to [0, 1]$ that depends on her chosen action a_i and the state. Each provider j has a utility function $u_j^P : \prod_{i \in N} \mathcal{A}_i \times \mathcal{Y} \to [0, 1]$ that depends on all users' actions $a_{1:n} = (a_1, ..., a_n)$ and the state.

Conversations and decision rules. We model a setting where an AI model and a user interact over multiple rounds of conversation, exemplified by the usage of AI chatbots. In particular, providers deploy models in the form of *conversation rules* that can be multi-round and adaptive, and users interact via their own conversation rules. The providers' conversation rules naturally represent model weights (together with any accompanying guardrails), which dictate the next sequence of tokens given a context and conversation history. Meanwhile, users' conversation rules represent a strategy for querying an AI, similarly mapping context and a conversation history to a next message.

Let $\mathcal{M}$ be a message space. Let $\mathcal{M}^{<R}$ be the space of conversation transcripts of length $< R$ and $\mathcal{M}^R$ be the space of conversation transcripts of length R. A *conversation rule for provider j* is a mapping $C_{P_j} : \mathcal{X}_{P_j} \times \mathcal{M}^{<R} \to \Delta(\mathcal{M})$ from his private features and the transcript of his conversation with some user to a distribution over next messages. A *conversation rule for user i* is a mapping $C_{U_i} : \mathcal{X}_{U_i} \times \mathcal{M}^{<R} \to \Delta(\mathcal{M})$ from her private features and the transcript of her conversation with some provider to a distribution over next messages. After the conversation ends, user i applies a *decision rule* $D_{U_i} : \mathcal{X}_{U_i} \times \mathcal{M}^R \to \Delta(\mathcal{A}_i)$ mapping her private features and the full transcript of her conversation with some provider to a distribution over actions.

The game. We model the interaction as a multi-leader, multi-follower Stackelberg game: providers commit to conversation rules, then users select which provider to consult, how to converse, and how to act. Commitment models the fact that training a new LLM requires substantial time and investment and happens over a longer time-frame than user decisions. After these weights are deployed, a user can try different AI models and learn how helpful each one is for her. After this experimentation period, each user will choose the model (provider conversation rule) to interact with which is best for her interests. This setup mirrors the "Best-AI Selection Game" defined in Collina et al. (2025b)).

A key modeling choice is the strategy space available to providers — in particular, whether they can tailor rules to each user or must deploy a single rule used anonymously by all users. We consider two versions of the game that capture this distinction. In the *personalized* game, each provider j commits to a separate conversation rule $C_{P_j}(i)$ for each user i. Each user i observes only the rules chosen for her $C_{\vec{P}}(i) = (C_{P_1}(i), \ldots, C_{P_k}(i))$. In the *anonymous* game, each provider j must commit to a single conversation rule C^{P_j} that is publicly observable and applied across all Alices. The personalized game captures account-based AI models, which can interact with different users differently based upon their profiles. Meanwhile, the anonymous game captures open-access AI models, which have no information about a user beyond their prompt.

In both games, upon observing providers' chosen rules, each user chooses a particular provider j to interact with, as well as a conversation rule C_{U_i} and decision rule D_{U_i}. Then, a a state y and players' private information $(x_1^U, \ldots, x_n^U, x_1^P, \ldots, x_k^P)$ are drawn from ϕ. Every user i engages in the conversation defined by C_{U_i} and $C_{P_j}(i)$ (C_{P_j} in the anonymous game) of her chosen provider and selects an action a_i according to D_{U_i}. Each user i receives her utility $u_i^U(a_i, y)$. Each provider j receives his utility $u_j^P(a_{1:n}, y)$. In both settings, each user observes only her own conversational transcript when applying her decision rule. The two games are fully defined in Appendix C.1.

The user's interaction with a provider, under chosen conversation and decision rules, induces a joint distribution over the user's action and the state of the world.

[1]For simplicity of notation we assume that all users and providers have access to this joint distribution. However, for all results in this paper, it is sufficient for user i to know only the joint distribution of (x_i^u, x_j^p, y) marginally for each provider j.

Definition 2.1 (Induced distribution $\mathcal{I}_i(C_P, C_U, D_U; j)$)**.** *We write $\mathcal{I}_i(C_P, C_U, D_U; j) \in \Delta(\mathcal{A}_\rangle \times \mathcal{Y})$ for the distribution over user i's actions and states induced by her interaction with provider j, given her conversation rule C_U and decision rule D_U, and the provider's conversation rule C_P.*

The user has three choices to make: (i) which provider she should converse with, (ii) how to converse with him, and (iii) how to map the learned information to a final decision. We assume that users make these choices optimally, i.e. *best respond* (formalized in Appendix C.1). Since we assume that every user best responds, the induced distributions depend on the providers' conversation rules alone. For every user i, we write $\mathcal{I}_i(C_{\vec{P}})$ for the induced distribution over (a_i, y) when user i best responds. We write $\mathcal{I}(C_{\vec{P}})$ for the induced joint distribution over over $(a_{1:n}, y)$.

Now, given that users best respond, this defines a game among providers. We will be interested in Nash equilibria of the game in both personalized and anonymous settings:

Definition 2.2 (Nash Equilibrium in the Personalized Game)**.** *A vector of providers' conversation rules $C_{\vec{P}}$ is a Nash equilibrium in the personalized game if for every provider j and every alternative vector of personalized rules $C'_{\vec{P}_j}$: $\mathbb{E}_{\mathcal{I}(C_{\vec{P}})}[u_j^P(a_{1:n}, y)] \geq \mathbb{E}_{\mathcal{I}(C'_{\vec{P}_j}, C_{\vec{P}_{-j}})}[u_j^P(a_{1:n}, y)]$*

Definition 2.3 (Nash Equilibrium in the Anonymous Game)**.** *A vector of providers' conversation rules $C_{\vec{P}}$ is a Nash equilibrium in the anonymous game if for every provider j and every alternative vector of rules C'_{P_j}: $\mathbb{E}_{\mathcal{I}(C_{\vec{P}})}[u_j^P(a_{1:n}, y)] \geq \mathbb{E}_{\mathcal{I}(C'_{P_j}, C_{P_{-j}})}[u_j^P(a_{1:n}, y)]$*

Benchmark. Our goal is to lower-bound the individual utility of each user in *every* possible Nash equilibrium. As each provider has potentially different information, there is a large and asymmetric set of defections each provider could make from any potential equilibrium. We will compare equilibrium outcomes to a benchmark that captures what *any* provider can implement given his private information. Accordingly, we are interested in the common information among all providers, a notion we call a *common garbling*.

Before formally defining this notion, we provide an illustrative setting where our notion of common information is straightforward. Suppose $y \in \mathbb{R}^d$ is a vector where each element is distributed independently, and each x_j^P is a subset of these coordinates. Here, the information that every provider has about y is captured precisely by the intersection of their sets. In other words, the "common information" among all providers is $z = \bigcap_{j \in K} x_j^P$. Importantly, since every provider has access to z, all providers can implement any conversation rule C_{P_j} depending only on z.

In general, y and $x_{1:k}^P$ may be arbitrarily correlated random variables, and thus the "common information" of $x_{1:k}^P$ may not be representable by a set intersection. Common garblings allow us to generalize this notion to random variables which can be simulated by any provider while preserving their correlation with y. In the special case where all providers observe the same feature, the shared feature itself is a common garbling. This is the setting of Collina et al. (2025b), and so our model is a strict generalization.

Definition 2.4 (Common garbling)**.** *A random variable $z \in \mathcal{Z}$ is a common garbling of $(x_1^P, \ldots, x_k^P)$ with respect to y if for every provider $j \in K$, there exists a (randomized) mapping $f_j : \mathcal{X}_{P_j} \to \Delta\mathcal{Z}$ such that, for all $\hat{y}$ and $\hat{z}$: $\Pr_{y \sim P}(z = \hat{z} | y = \hat{y}) = \mathbb{E}_{x_j^P, y \sim \phi}[\Pr_{f_j}\left(f_j(x_j^P) = \hat{z}\right) | y = \hat{y}]$. In other words, conditional on a fixed value of y, the marginals of z and $f_j(x_j^P)$ are equivalent.*

We can now define the set of conversation rules that can be implemented by any provider in some set T.

Definition 2.5 (T-Shared Conversation Rules $\mathcal{C}_S(x_T^P)$)**.** *Given a set of random variables $x_{1:k}^P, y$, a* T-Shared Conversation Rule *is a conversation rule of the form*

$$C_S : \mathcal{Z} \times \mathcal{M}^{<R} \to \Delta(\mathcal{M})$$

where $z \in \mathcal{Z}$ is a realization of a random variable that is a common garbling of x_T^P with respect to y, where $T \subseteq K$. We denote the set of all such R-round T -shared conversation rules $\mathcal{C}_{S,R}(x_T^P)$.

We define user i's utility when she interacts with a particular provider who deploys some conversation rule.

Definition 2.6 (User i's utility when interacting with provider j who deploys C_{P_j})**.** *We write* $u_i^U(C_{P_j}; j) = \mathbb{E}_{\mathcal{I}_i(C_{P_j}, C^*_{U_i}, D^*_{U_i}; j)}[u_i^U(a_i, y)]$ *to denote user i's expected utility when she interacts with the conversation rule C_{P_j} deployed by provider j.*

A special case is when a provider deploys a shared conversation rule C_G taking as input a common garbling z. Since *any* provider can generate z and implement C_G, the distribution induced by the user's interaction with C_G — and hence the user's expected utility — is the same no matter which provider deployed it. Thus, we will drop the provider's index when speaking about the user's utility when interacting with a shared conversation rule.

Definition 2.7 (User i's utility when interacting with C_S)**.** *We write* $u_i^U(C_S) = \mathbb{E}_{\mathcal{I}_i(C_S, C^*_{U_i}, D^*_{U_i})}[u_i^U(a_i, y)]$ *to denote user i's expected utility over the induced distribution of interacting with conversation rule C_S which depends on random variable z.*

This leads us to the following benchmark which we will use throughout the paper — the outcome the user can obtain if she were interacting with a single perfectly aligned provider who can implement any shared conversation rule. As we will frequently consider subsets of providers who satisfy market alignment conditions, it will be useful to define the optimal shared conversation rule for any subset of providers.

Definition 2.8 (Optimal T-Shared Conversation Rule $C^*_{S,T,R}(i)$)**.** *User i's optimal R-round T-shared conversation rule $C^*_{S,T,R}(i)$ is the R-round T-shared conversation rule which maximizes her utility, assuming she best responds* [2]*. Formally:* $C^*_{S,T,R}(i) := \arg\max_{C_S \in \mathcal{C}_{S,R}(x_T^P)} u_i^U(C_S)$*. We drop the subscript R when the number of rounds is clear from context.*

Market Alignment Conditions. We now give alignment conditions formalizing how AI providers' preferences relate to users'. These are *market-level* conditions: they do not require any single AI provide to be aligned with any user, but rather offer enough structure such that competition can lead to user-aligned outcomes. In our results, it will only be necessary for a *subset* of the providers to satisfy our alignment conditions; thus, we write the below definitions for a set of providers $T \subseteq K$.

Definition 2.9 ($(T, \epsilon_U, \epsilon_P)$-Weak Market Alignment)**.** *We say a set $T \subseteq K$ of providers satisfy $(T, \epsilon_U, \epsilon_P)$-Weak Market Alignment if:*

1. *(Provider separability) Every provider $j \in T$ has an additively separable utility. That is, for every provider $j \in T$, there exist non-negative weights $(\lambda_{j,i})_{i \in N}$, a constant $c_j \in \mathbb{R}$, and an error $\epsilon_P \geq 0$ such that for all action profiles $a_{1:n}$ and states y:*

$$\left| u_j^P\left(a_{1:n}, y\right) - \left(\sum_{i \in N} \lambda_{j,i} F_{j,i}(a_i, y) + c_j \right) \right| \leq \epsilon_P \tag{1}$$

   where $F_{j,i} : \mathcal{A}_i \times \mathcal{Y} \to \mathbb{R}^{\geq 0}$ depends only on user i's action and the state.
2. *(User alignment) For every user $i \in N$, there exist non-negative weights $(w_{j,i})_{j \in T}$, a constant $c_i \in \mathbb{R}$, and an error $\epsilon_U \geq 0$ such that for all actions a_i and states y:*

$$\left| u_i^U(a_i, y) - \left(\sum_{j \in T} w_{j,i} F_{j,i}(a_i, y) + c_i \right) \right| \leq \epsilon_U \tag{2}$$

We assume that for every user i, there is a provider j^* where $\lambda_{j^*,i} > 0$ and $w_{j^*,i} > 0$. Additionally, if $w_{j,i} > 0$, then $\lambda_{j,i} > 0$.

Our Weak Market Alignment condition can, for example, naturally model settings where AI providers care about the number of users who take a particular downstream action (e.g. purchase a product, prescribe a drug, or vote for a political candidate) with different importance weights placed on different users or user groups (e.g. demographic groups of consumers or voters). For every provider j, $F_{j,i}(a_i, y) \in [0, 1]$ can measure how desirable user i's action a_i is in a state y for that provider (e.g. $F_{j,i} = \mathbf{1}[a_i = a^*(j; y)]$ where $a^*(y, j)$ is provider j's preferred downstream action

[2] Note that if C is a K-shared conversation rule, it is also a T-shared conversation rule for any $T \subseteq K$. Thus, $\arg\max_{C_S \in \mathcal{C}_{S,R}(x_T^P)} u_i^U(C_S) \geq \arg\max_{C_S \in \mathcal{C}_{S,R}(x_{1:k}^P)} u_i^U(C_S)$.

in state y). The weights $\lambda_{j,i}$ then capture how much provider j values the resulting action taken by a particular user (or group), while $w_{j,i}$ capture how much relative utility the user receives from provider j's preferred outcome. We also define a Strong Market Alignment condition:

Definition 2.10 ((T, ϵ)-Strong Market Alignment)**.** *We say a set $T \subseteq K$ of providers satisfy (T, ϵ)-Strong Market Alignment if every provider $j \in T$ has a utility that can be approximately written as a non-negative linear combination of user utilities. That is, for every provider $j \in T$, there exist non-negative weights $(\lambda_{j,i})_{i \in N}$, a constant $c_j \in \mathbb{R}$, and an error $\epsilon \geq 0$ such that for all action profiles $a_{1:n}$ and states y:*

$$\left| u_j^P\left(a_{1:n}, y\right) - \left(\sum_{i \in N} \lambda_{j,i}\, u_i(a_i, y) + c_j \right) \right| \leq \epsilon \tag{3}$$

We assume coverage: for every user i, there exists a provider $j^* \in T$ with $\lambda_{j^*,i} > 0$.

This condition can be viewed as the dual of the "Weighted Alignment Assumption" in Collina et al. (2025b), which requires that, in a single-user setting, the user's utility can be written as a non-negative linear combination of provider utilities. We show that Strong Market Alignment is an instantiation of Weak Market Alignment, where $F_{j,i} = u_i^U(a_i, y)$ (Proposition C.4). In fact, Weak Market Alignment is strictly weaker (Proposition C.5).

3 The Personalized Game

We begin by studying the personalized game, where each provider can tailor his conversation rule to the identity of the user he is interacting with. We show that under Weak Market Alignment, every user achieves near-benchmark in *every* Nash equilibrium of the personalized game. In particular, she gets utility close to what she would get by interacting with a perfectly aligned provider restricted to using common information. The intuition is simple. With personalization, providers can decouple their choices across users: a provider can strategically influence the action of a single user by changing what she observes, without committing to the same behavior for every other user. This property, together with the alignment condition, implies every user must be getting high utility: if any user were doing poorly, some provider could profit by offering her a better personalized rule, contradicting equilibrium.

Importantly, our result does not require *every* provider to satisfy the alignment condition — an assumption that can be unrealistic in markets of many provider and many users. Instead, we only assume that there exists *some* subset of providers that does. This substantially weaker guarantee suffices to guarantee that competition among all providers yields aligned outcomes for users. In particular, our results will continue to hold if additional providers who do not satisfy the alignment condititon enter the market.

Theorem 3.1. *Suppose there is a set of providers $T \subseteq K$ that satisfy the $(T, \epsilon_U, \epsilon_P)$-Weak Market Alignment with weights $(\lambda_{j,i})_{i \in N}$ for each provider $j \in T$ and $(w_{j,i})_{j \in T}$ for each user $i \in N$. In the personalized game among all providers K, for each user i, her expected utility in any Nash equilibrium is at least $u_i^U(C^*_{S,T}(i)) - 2\epsilon_U - 2\epsilon_P \mu_i$, where $\mu_i = \sum_{j \in T} \frac{w_{j,i}}{\lambda_{j,i}}$.*

4 The Anonymous Game

In the previous section we established that every user obtains her benchmark utility in the personalized setting when Weak Market Alignment holds. Here, we turn attention to the anonymous version of the game and ask: what can we guarantee users when personalization is not available?

Our first result shows a separation: without personalization, users can suffer arbitrarily poor outcomes in Nash equilibrium—even when the Weak Market Alignment condition holds. The intuition is that since providers must commit to a single rule across users, strategic choices are no longer decomposable among users. A disclosure to one user is shown to everyone, and so providers must trade off how different users will respond. We show that this tension can push the market toward policies that reveal no useful information to any user. In particular, we construct a class of anonymous game instances where providers have conflicting preferences over which user should be informed. The users are symmetric (they share the same utility and private information), so they will best respond

in the same way. If each provider desire to hide information from one user is stronger than his desire to share information with the other, then each provider will prefer to withhold his information.

Theorem 4.1. *For any $\epsilon > 0$ and $c \in (\epsilon, 1]$, there exists an anonymous game exactly satisfying the Weak Market Alignment condition (Definition 2.9) that has a Nash equilibrium where every user i's expected utility is at most ϵ, while the optimal shared conversation rule benchmark $u_i^U(C^*_{S,K}(i))$ is at least c.*

Motivated by this separation, we then ask when strong user guarantees can be recovered in the anonymous setting. We give a simple setting where Strong Market Alignment suffices to guarantee that every user obtains her benchmark utility — namely, the setting where a *user-dominant* conversation rule (one that is simultaneously optimal for *every* user) is available.

Definition 4.2 (User-Dominant Conversation Rule)**.** *Given a provider j, we say that a conversation rule $C^*_{P_j}$ is user-dominant if for all users i and all alternative conversation rules C_{P_j} Bob j could deploy, $u_i^U(C^*_{P_j}; j) \geq u_i^U(C_{P_j}; j)$.*

Our next result shows that whenever a user-dominant rule exists[3], every user obtains utility close to her utility if she were to interact with this user-dominant rule, under the Strong Market Alignment condition. Intuitively, since provider utilities are non-negative linear combinations of user utilities, and user utilities are maximized by the user-dominant rule, providers will themselves prefer to deploy the user-dominant rule. As before, our results here do not require every provider in the marketplace to satisfy the alignment condition; it will suffice that there exist providers that do.

Theorem 4.3. *Suppose there exists a set of providers $T \subseteq K$ such that (T, ϵ)-Strong Market Alignment is satisfied with weights $(\lambda_{j,i})_{i \in N}$ for every provider $j \in T$. Furthermore, suppose for every provider $j \in T$, there exists a user-dominant conversation rule $C^*_{P_j}$ for all ysers $i \in N$. Then, for each user $i \in N$, her expected utility in any Nash equilibrium is at least $u_i^U(C^*_{S,T}(i)) - \frac{2\epsilon}{\lambda_i^*}$, where $\lambda_i^* = \max_{j \in T} \lambda_{j,i}$.*

More generally, even when no such user-dominant rule exists, Strong Market Alignment still ensures that every Nash equilibrium in the anonymous game gives users high utility, albeit relative to a slightly weaker benchmark. We defer discussion of these results to Appendix E.1.

5 EXPERIMENTS

We empirically test our alignment conditions using the OpinionQA dataset (Santurkar et al., 2023), which provides answer distributions for both demographic groups and LLMs on survey questions. This setting maps naturally to our framework: demographic groups are *users* (users with heterogeneous preferences), LLMs are *providers* (competing providers), and survey questions are states.

Setup. For each demographic group i and question y, we define user's utility as $u_i^U(a, y) = p_i(a \mid y)$, the probability that group i selects answer a. Similarly, for each model j, we define provider's utility from user i's action as $v_j(a, y) = p_j^B(a \mid y)$ and $u_j^P(a_{1:n}, y) = \frac{1}{n}\sum_{i=1}^n v_j(a_i, y)$, the model's predicted probability. We test on Political Ideology demographic partitions across two survey waves from the Pew American Trends Panel: W29 (gender and health) and W82 (economy and democracy). See Section G for additional demographic partitions, waves, and scoring rules.

Testing Weak Market Alignment. Given the structure of u_j^P, $F_{j,i} = v_j/n$ gives us $\epsilon_P = 0$. Therefore, checking the condition reduces to fitting $u_i^U \approx \sum_{j \in K} w_{j,i} v_j + c_i, \quad w_{j,i} \geq 0$. We do this using nonnegative least squares with cross-validation (5-fold over questions).

More providers ⇒ better fit. Nonnegative combinations of model utilities approximate group preferences substantially better than either the best single model or a simple average, and fit improves as K grows (Figure 1, top).

Heterogeneous difficulty across groups. Within a demographic partition, some groups are consistently harder to approximate than others (Figure 1, bottom).

[3]Observe that a user-dominant conversation rule exists whenever the conversation is permissive enough for a provider to reveal his private information — this "full-revelation" rule is user-dominant.

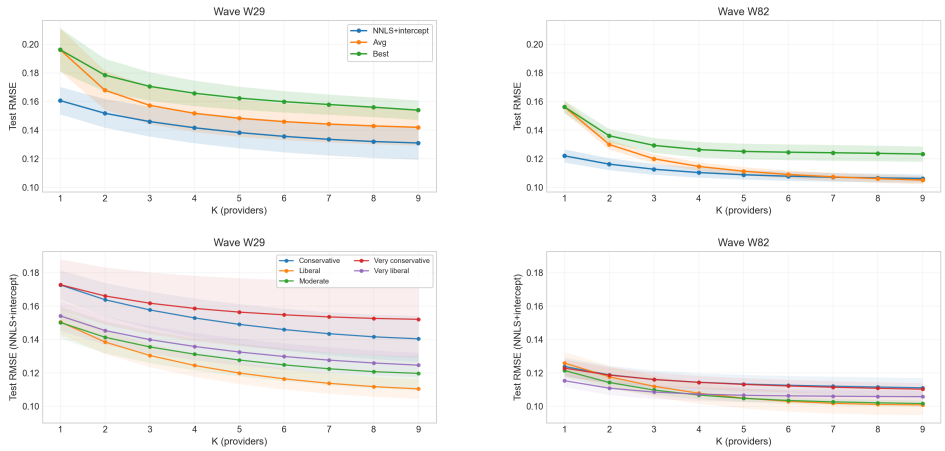

Figure 1: (Top) **Weak Market Alignment improves with more providers.** Mean RMSE (across groups) vs. number of available providers K. NNLS+intercept outperforms baselines; error decreases with K. (Bottom) **Some groups are harder to align.** Per-group RMSE vs. K. Within each partition, alignment quality varies across demographic groups.

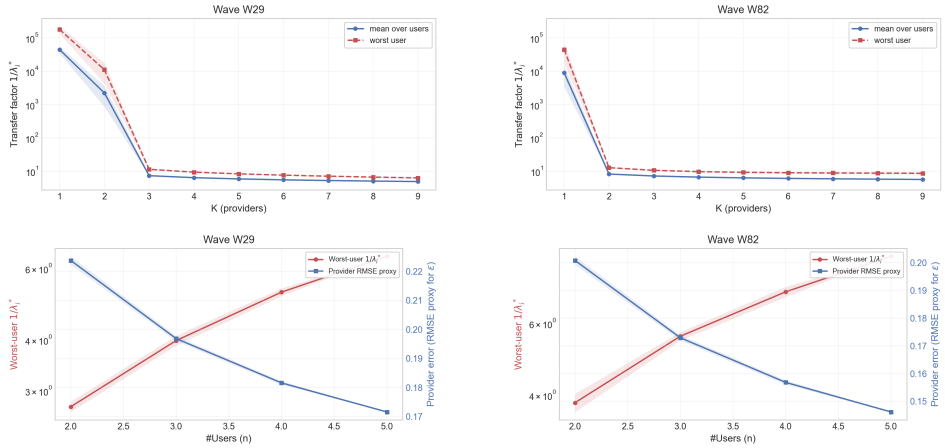

Figure 2: (Top) **More providers reduce transfer factors.** Transfer factor $1/\lambda_i^*$ vs. number of providers K: mean over users (solid) and worst user (dashed). The dramatic drop shows how provider diversity improves coverage. (Bottom) **More users: lower error but harder coverage.** Worst-user transfer factor $1/\lambda_i^*$ (red, left axis) vs. provider fitting error ϵ (blue, right axis) as number of users increases.

Testing Strong Market Alignment. We fit $u_j^P \approx \sum_i \lambda_{j,i}\, u_i^U + c_j$ with $\lambda_{j,i} \geq 0$ using simulated action distributions. Beyond the approximation error ϵ, the key quantity for our theorems is the *transfer factor* $1/\lambda_i^*$, where $\lambda_i^* = \max_j \lambda_{j,i}$. This measures how well some provider "covers" user i: small means strong coverage (good); large means weak coverage (bad).

More providers dramatically improve coverage. At $K{=}1$, transfer factors are enormous (10–100 or higher). Adding just one or two more provider ($K{=}2, 3$) drops transfer factors by orders of magnitude (Figure 2, top). The appendix visualizes the learned weights $\lambda_{j,i}$: different models specialize in covering different groups.

Strong alignment is harder than Weak. The Strong RMSE is $\sim 50\%$ larger than Weak RMSE. Comparing the theorem bounds: personalized game has slack $2\epsilon_U$ ($\epsilon_P = 0$) (Theorem 3.1), while anonymous game has slack $2\epsilon/\lambda_i^*$ (Theorem 4.3). Combined with transfer amplification at small K, the effective anonymous-game bound can be orders of magnitude looser than the personalized bound.

More users create a coverage–error trade-off. Increasing the number of demographic groups (users) within a partition has opposing effects (Figure 2, bottom): fitting error ϵ *decreases* (more groups means more signal), but the worst-case transfer factor $1/\lambda_i^*$ from Theorem 4.3 *increases* (harder to cover everyone).

REFERENCES

Scott Aaronson. The complexity of agreement. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pp. 634–643, 2005.

Ricardo Alonso and Odilon Câmara. Persuading voters. *American Economic Review*, 106(11): 3590–3605, 2016.

Itai Arieli and Yakov Babichenko. Private bayesian persuasion. *Journal of Economic Theory*, 182: 185–217, 2019.

Pak Hung Au and Keiichi Kawai. Competitive information disclosure by multiple senders. *Games and Economic Behavior*, 119:56–78, 2020.

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.

Natalie Collina, Surbhi Goel, Varun Gupta, and Aaron Roth. Tractable agreement protocols. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 1532–1543, 2025a.

Natalie Collina, Surbhi Goel, Aaron Roth, Emily Ryu, and Mirah Shi. Emergent alignment via competition, 2025b. URL `https://arxiv.org/abs/2509.15090`.

Natalie Collina, Ira Globus-Harris, Surbhi Goel, Varun Gupta, Aaron Roth, and Mirah Shi. Collaborative prediction: Tractable information aggregation via agreement. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 4712–4798. SIAM, 2026.

Rafael Frongillo, Eric Neyman, and Bo Waggoner. Agreement implies accuracy for substitutable signals, 2021. URL `https://arxiv.org/abs/2111.03278`.

Matthew Gentzkow and Emir Kamenica. Competition in persuasion. *The Review of Economic Studies*, 84(1):300–322, 2016.

Matthew Gentzkow and Emir Kamenica. Bayesian persuasion with multiple senders and rich signal spaces. *Games and Economic Behavior*, 104:411–429, 2017.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional ai: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1395–1417, 2024.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101 (6):2590–2615, 2011.

Michael Kearns, Aaron Roth, and Emily Ryu. Networked information aggregation via machine learning. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 4799–4845. SIAM, 2026.

Toryn Q Klassen, Parand A Alamdari, and Sheila A McIlraith. Pluralistic alignment over time. *arXiv preprint arXiv:2411.10654*, 2024.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Fei Li and Peter Norman. On bayesian persuasion with multiple senders. *Economics Letters*, 170: 66–70, 2018.

Aran Nayebi. Intrinsic barriers and practical pathways for human–ai alignment: An agreement-based complexity analysis. *arXiv preprint arXiv:2502.05934*, 2025.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Ali Shirali, Arash Nasr-Esfahany, Abdullah Alomar, Parsa Mirtaheri, Rediet Abebe, and Ariel Procaccia. Direct alignment with heterogeneous preferences. *arXiv preprint arXiv:2502.16320*, 2025.

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*, pp. 46280–46302, 2024.

Jacques-Francois Thisse and Xavier Vives. On the strategic choice of spatial price policy. *The American Economic Review*, pp. 122–137, 1988.

Wenhao Wu. Sequential bayesian persuasion. *Journal of Economic Theory*, 214:105763, 2023.

A Robustness to Adding Alices

So far, we have analyzed when pluralistically aligned outcomes emerge for a fixed population of users. In reality, new users — with different preferences and information — may regularly enter the ecosystem, possibly breaking the conditions that guarantee alignment. What happens to the guarantees enjoyed by pre-existing users when this occurs?

We study this question in the two regimes where we have identified sufficient conditions for aligned outcomes: the personalized setting under Weak Market Alignment, and the anonymous setting under Strong Market Alignment. In each regime, we analyze the effect of a single user entering the market. More specifically, given an original game instance G, we define the *augmented game* G^+ by adding user $n+1$ with utility $u^U_{n+1} : \mathcal{A}_{n+1} \times \mathcal{Y} \to [0,1]$ and feature distribution $\phi(x^U_{n+1})$ that is the extension of ϕ to a joint distribution given the addition of user $n+1$'s private features x^U_{n+1}. We allow each provider j's utility to change via the addition of an arbitrary function of the new user's action: $u^P_j(a_{1:n+1}, y) = (1-\beta_j)\cdot u^P_j(a_{1:n}, y) + \beta_j \cdot f(a_{n+1}, y)$ where $\beta_j \in (0,1)$ and $f : \mathcal{A}_{n+1} \times \mathcal{Y} \to [0,1]$ is an arbitrary function. We assume the original users' utilities stay the same.

Importantly, this new user may violate our alignment conditions, and we do not hope to prove utility guarantees for her; instead, we study the effect of her entrance on the utility guarantees of the other users. We show that adding a user to the personalized setting with (previously satisfied) Weak Market Alignment does *not* hurt the other users; they still attain the guarantees they had before her entrance. By contrast, adding a user to the anonymous setting with (previously satisfied) Strong Market Alignment can not only result in poor outcomes for the new user, but also break the utility guarantees that the other users previously had.

Theorem A.1. *Let $T \subseteq K$ be a set of providers that satisfy $(T, \epsilon_U, \epsilon_P)$-approximate Weak Market Alignment with respect to users N. In every Nash equilibrium of an augmented personalized game given by G^+, every original user $i \in N$ obtains expected utility at least $u^U_i(C^*_{S,T}(i)) - 2\epsilon_U - 2\epsilon_P \mu_i$, where $\mu_i = \sum_{j \in T} \frac{w_{j,i}}{\lambda_{j,i}}$.*

Theorem A.2. *Consider a game instance G where the $(K, 0)$-approximate Strong Market Alignment condition is satisfied and full revelation is possible. Then, there exists $(f_{1:K}, \beta_{1:K}, u^U_{n+1}, P(x^A_{n+1}))$ such that, in the induced augmented anonymous game G^+, there is a Nash equilibrium such that* all *users $i \in N \cup \{n+1\}$ get utility $< u^U_i(C^*_{S,K})$.*

B ADDITIONAL RELATED WORK

Pluralistic Alignment A recent literature on pluralistic alignment (see e.g. (Sorensen et al., 2024; Shirali et al., 2025; Huang et al., 2024; Klassen et al., 2024)) focuses on the technical alignment problem of a single model provider when they have many downstream users with different preferences. These approaches pre-suppose that the model designer wants to do this. We instead study conditions under which market competition between many providers, none of which may have incentives to align their models with their users (pluralistically or otherwise) will nevertheless lead to pluralistic alignment emerging from equilibrium.

C ADDITIONAL CONTENT FROM SECTION 2

C.1 GAME DETAILS

Definition C.1 (The Personalized Game)**.** *The personalized game proceeds as follows:*

1. *Every provider j simultaneously commits to a separate conversation rule $C_{P_j}(i)$ for each user i. Let $C_{\vec{P}_j} = (C_{P_j}(1), ..., C_{P_j}(n))$ be the vector of rules chosen by provider j and let $C_{\vec{P}} = (C_{\vec{P}_1}, \ldots, C_{\vec{P}_k})$ be the entire collection of chosen rules. Let $C_{\vec{P}}(i) = (C_{P_1}(i), \ldots, C_{P_k}(i))$ be the collection of rules chosen for user i.*
2. *Every user i observes $C_{\vec{P}}(i)$. She chooses a provider j to interact with, as well as a conversation rule C_{U_i} and a decision rule D_{U_i}.*
3. *The state y and players' private information $(x_1^U, \ldots, x_n^U, x_1^P, \ldots, x_k^P)$ are drawn from ϕ. Each user i observes x_i^U, and each provider j observes x_j^P.*
4. *Every user i engages in the conversation defined by C_{U_i} and $C_{P_j}(i)$ of her chosen provider and selects an action a_i according to D_{U_i}.*
5. *Each user i receives her utility $u_i^U(a_i, y)$. Each provider j receives his utility $u_j^P(a_{1:n}, y)$.*

Definition C.2 (The Anonymous Game)**.** *The anonymous game proceeds as follows:*

1. *Every provider j simultaneously commits to a conversation rule C_{P_j}. Let $C_{\vec{P}} = (C_{P_1}, \ldots, C_{P_k})$ be the collection of chosen rules.*
2. *Every user i observes $C_{\vec{P}}$. She chooses a provider j to interact with, as well as a conversation rule C_{U_i} and a decision rule D_{U_i}.*
3. *The state y and players' private information $(x_1^U, \ldots, x_n^U, x_1^P, \ldots, x_k^P)$ are drawn from ϕ. Each user i observes x_i^U, and each provider j observes x_j^P.*
4. *Every user i engages in the conversation defined by C_{U_i} and C_{P_j} of her chosen provider and selects an action a_i according to D_{U_i}.*
5. *Each user i receives her utility $u_i^U(a_i, y)$. Each provider j receives his utility $u_j^P(a_{1:n}, y)$.*

Definition C.3 (User's Best Response)**.** *User i's best response consists of the following:*

- *User i's best response decision rule $D_{U_i}^*$ is the decision rule that maximizes her utility under the posterior distribution $\pi(y|x_i^U, M^R)$ given her observed features $x_i^U \in \mathcal{X}_{U_i}$ and a conversation transcript $M^R \in \mathcal{M}^R$ with provider j:*

$$D_{U_i}^* = \arg\max_{a_i \in \mathcal{A}_i} \mathbb{E}_\pi[u_i^U(a_i, y)]$$

- *User i's best response conversation rule $C_{U_i}^*$ given a conversation rule C_{P_j} is the conversation rule that maximizes her expected utility given that she uses her best response decision rule:*

$$C_{U_i}^* = \arg\max_{C_{U_i}} \mathbb{E}_{\mathcal{I}_i(C_{P_j}, C_{U_i}, D_{U_i}^*; j)}[u_i(a_i, y)]$$

- *User i's optimal choice of a provider to interact with is the provider j^* that maximizes her expected utility given that she uses her best response conversation and decision rules:*

$$j^* = \arg\max_{j \in K} \mathbb{E}_{\mathcal{I}_i(C_{P_j}, C^*_{U_i}, D^*_{U_i}; j)}[u_i(a_i, y)]$$

We assume a fixed and common tie-breaking rule across all users.

C.2 Details about Alignment Conditions

Proposition C.4 (Strong Market Alignment Implies Weak Market Alignment)**.** *Fix a set $T \subset K$ of providers. Suppose the provider utilities satisfy the (T, ϵ)-approximate Strong Market Alignment condition. Then, the provider utilities satisfy the $(T, \epsilon, 0)$-approximate Weak Market Alignment condition. When $\epsilon = 0$, Strong Market Alignment implies Weak Market Alignment.*

Proof. Take $F_{j,i}(a_i, y) = u_i^U(a_i, y)$ for all $j \in T$ and $i \in N$. Then the provider separability condition is equivalent to the Strong Market Alignment condition:

$$\left| u_j^P\big(a_{1:n}, y\big) \; - \left(\sum_{i \in N} \lambda_{j,i} \, F_{j,i}(a_i, y) + c_j \right) \right| = \left| u_j^P\big(a_{1:n}, y\big) \; - \left(\sum_{i \in N} \lambda_{j,i} \, u_i(a_i, y) + c_j \right) \right| \leq \epsilon$$

By assumption there exists at least one provider j with $\lambda_{j,i} > 0$. Let n^* be $|\{j : \lambda_{j,i} > 0\}|$. Then, the Weak Market Alignment condition is exactly satisfied with $\epsilon_U = 0$ by taking weights $w_{j,i} = 1/n^*$ for all j with $\lambda_{j,i} > 0$, $w_{j,i} = 0$ for all other j, and $c_i = 0$. □

Proposition C.5 (Weak Market Alignment is Strictly Weaker)**.** *There exist provider and user utilities that exactly satisfy the Weak Market Alignment condition, but for every $\epsilon < 1/2$, the (T, ϵ)-Strong Market Alignment condition fails.*

Proof. Consider an instance with $n = k = 2$ (i.e. 2 providers and 2 users). Let $\mathcal{Y} = \{(y_1, y_2) \in \mathbb{R}^2 : y_1, y_2 \geq 0, \; y_1 + y_2 = 1\}$ and $\mathcal{A}_1 = \mathcal{A}_2 = [0, 1]$. Define:

$$F_{1,1}(a_1, y) = \left\langle \begin{bmatrix} a_1/2 \\ 0 \end{bmatrix}, y \right\rangle, \;\; F_{1,2}(a_2, y) = \left\langle \begin{bmatrix} a_2/2 \\ 0 \end{bmatrix}, y \right\rangle$$
$$F_{2,1}(a_1, y) = \left\langle \begin{bmatrix} 0 \\ a_1/2 \end{bmatrix}, y \right\rangle, \;\; F_{2,2}(a_2, y) = \left\langle \begin{bmatrix} 0 \\ a_2/2 \end{bmatrix}, y \right\rangle$$

Suppose $u_j^P(a_1, a_2, y) = F_{j,1}(a_1, y) + F_{j,2}(a_2, y)$ for each provider j and

$$u_i^U(a_i, y) = \frac{1}{2} F_{1,i}(a_i, y) + \frac{1}{2} F_{2,i}(a_i, y) = \left\langle \begin{bmatrix} \frac{a_i}{4} \\ \frac{a_i}{4} \end{bmatrix}, y \right\rangle$$

for each user i. Since $a_i \in [0, 1]$ and $y_1, y_2 \in [0, 1]$ with $y_1 + y_2 = 1$, we have $F_{j,i}(a_i, y) \in [0, 1/2]$, hence $u_j^P(a_1, a_2, y) \in [0, 1]$ and $u_i^U(a_i, y) \in [0, 1]$. Thus Weak Market Alignment (Definition 2.9) is satisfied.

Now, for provider 1, we have that

$$u_1^P(a_1, a_2, y) = \left\langle \begin{bmatrix} \frac{a_1 + a_2}{2} \\ 0 \end{bmatrix}, y \right\rangle.$$

We show that (T, ϵ)-Strong Market Alignment fails for every $\epsilon < \frac{1}{2}$. Fix $\epsilon < \frac{1}{2}$ and suppose for contradiction that there exist $\lambda_1, \lambda_2 \geq 0$ and $c \in \mathbb{R}$ such that for all $a_1, a_2 \in [0, 1]$ and $y \in \mathcal{Y}$,

$$\left| u_1^P(a_1, a_2, y) - \big(\lambda_1 u_1^U(a_1, y) + \lambda_2 u_2^U(a_2, y) + c\big) \right| \leq \epsilon.$$

Take $a_1 = a_2 = 1$, and let $y^{(0)} = (0, 1)$ and $y^{(1)} = (1, 0)$. Then $u_1^U(1, y^{(0)}) = u_1^U(1, y^{(1)}) = \frac{1}{4}$ and $u_2^U(1, y^{(0)}) = u_2^U(1, y^{(1)}) = \frac{1}{4}$, while $u_1^P(1, 1, y^{(0)}) = 0$ and $u_1^P(1, 1, y^{(1)}) = 1$. Writing $v = \frac{\lambda_1}{4} + \frac{\lambda_2}{4} + c$, the inequality implies both $|v| \leq \epsilon$ and $|1 - v| \leq \epsilon$, hence $\epsilon \geq \frac{1}{2}$, a contradiction.

The same argument applies for u_2^P (swap the roles of the two coordinates of y), so no provider satisfies (T, ϵ)-Strong Market Alignment for any $\epsilon < \frac{1}{2}$. Therefore Strong Market Alignment fails. □

D PROOFS FROM SECTION 3

Proof of Theorem 3.1. Let $C_{\vec{P}}$ be any Nash equilibrium strategy profile of the personalized game, and let $\mathcal{I}_{NE} = \mathcal{I}(C_{\vec{P}})$ be the induced distribution.

Now, assume for contradiction that there exists a user $i \in N$ such that her utility is strictly lower than $u_i^U(C^*_{S,T}(i)) - 2\epsilon_U - 2\epsilon_P\mu_i$. Consider a deviation by any provider $j \in T$ in his choice of conversation rule for user i from $C_{P_j}(i)$ to $C^*_{S,T}(i)$, the optimal T-shared conversation rule for user i. Note that by definition, any provider in T could implement $C^*_{S,T}(i)$.

Let $\mathcal{I}_{dev} = \mathcal{I}(C^*_{S,T}(i), C_{P_j}(-i), C_{P_{-j}})$ be the distribution induced by the deviation of provider j in his chosen conversation rule for user i (the conversation rules chosen by provider j for all other users, $C_{P_j}(-i)$, remain unchanged). As user i's utility was previously worse than $u_i^U(C^*_{S,T}(i))$, she will now choose to interact with the provider who implements $C^*_{S,T}(i)$, leading to the outcome under $\mathcal{I}_{dev}$.

Now, by the Nash equilibrium condition, no provider j has an incentive to deviate. Thus, for all providers $j \in T$:

$$\mathbb{E}_{\mathcal{I}_{dev}}[u_j^P(a_{1:n}, y)] \leq \mathbb{E}_{\mathcal{I}_{NE}}[u_j^P(a_{1:n}, y)].$$

Since every user observes only the conversation rules chosen for her, this deviation can only change the induced distribution for user i (in particular the distribution over her actions). Thus, applying Weak Market Alignment, we have that for any provider $j \in T$:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{I}_{dev}}\left[\sum_{i\in N}\lambda_{j,i}F_{j,i}(a_i, y) + c_j - \epsilon_P\right] \leq \mathbb{E}_{\mathcal{I}_{NE}}\left[\sum_{i\in N}\lambda_{j,i}F_{j,i}(a_i, y) + c_j + \epsilon_P\right] \\
\iff &\mathbb{E}_{\mathcal{I}_{dev}}\left[\lambda_{j,i}F_{j,i}(a_i, y) + \sum_{i'\neq i}\lambda_{j,i'}F_{j,i'}(a_{i'}, y) + c_j - \epsilon_P\right] \\
&\leq \mathbb{E}_{\mathcal{I}_{NE}}\left[\lambda_{j,i}F_{j,i}(a_i, y) + \sum_{i'\neq i}\lambda_{j,i'}F_{j,i'}(a_{i'}, y) + c_j + \epsilon_P\right] \\
\iff &\mathbb{E}_{\mathcal{I}_{dev}}\left[\lambda_{j,i}F_{j,i}(a_i, y)\right] + \mathbb{E}_{\mathcal{I}_{NE}}\left[\sum_{i'\neq i}\lambda_{j,i'}F_{j,i'}(a_{i'}, y)\right] + c_j - \epsilon_P \\
&\leq \mathbb{E}_{\mathcal{I}_{NE}}\left[\lambda_{j,i}F_{j,i}(a_i, y)\right] + \mathbb{E}_{\mathcal{I}_{NE}}\left[\sum_{i'\neq i}\lambda_{j,i'}F_{j,i'}(a_{i'}, y)\right] + c_j + \epsilon_P \\
\iff &\lambda_{j,i}\mathbb{E}_{\mathcal{I}_{dev}}\left[F_{j,i}(a_i, y)\right] \leq \lambda_{j,i}\mathbb{E}_{\mathcal{I}_{NE}}\left[F_{j,i}(a_i, y)\right] + 2\epsilon_P
\end{aligned}$$

where the third step follows from the fact that only the induced distribution over user i's actions changes under the deviation. For any j with $\lambda_{j,i} > 0$, we thus have:

$$\mathbb{E}_{\mathcal{I}_{dev}}\left[F_{j,i}(a_i, y)\right] \leq \mathbb{E}_{\mathcal{I}_{NE}}\left[F_{j,i}(a_i, y)\right] + 2\frac{\epsilon_P}{\lambda_{j,i}}$$

Taking a weighted sum over all providers using the non-negative weights $(w_{j,i})_{j\in T}$ from the Weak Market Alignment condition:

$$\sum_{j\in T} w_{j,i}\mathbb{E}_{\mathcal{I}_{dev}}[F_{j,i}(a_i, y)] \leq \sum_{j\in T} w_{j,i}\mathbb{E}_{\mathcal{I}_{NE}}[F_{j,i}(a_i, y)] + 2\epsilon_P\sum_{j\in T}\frac{w_{j,i}}{\lambda_{j,i}}.$$

Note that by assumption, if $w_{j,i} > 0$, then $\lambda_{j,i} > 0$, and so we can apply the inequality here. By linearity of expectation, this is equivalent to:

$$\mathbb{E}_{\mathcal{I}_{dev}}\left[\sum_{j\in T} w_{j,i}F_{j,i}(a_i, y)\right] \leq \mathbb{E}_{\mathcal{I}_{NE}}\left[\sum_{j\in T} w_{j,i}F_{j,i}(a_i, y)\right] + 2\epsilon_P\sum_{j\in T}\frac{w_{j,i}}{\lambda_{j,i}}.$$

Now we use the market alignment assumption, which states that $\sum_j w_{j,i} F_{j,i}(a_i, y)$ can be approximated by $u_i^U(a_i, y) + c_i$. We have that:

$$\mathbb{E}_{\mathcal{I}_{dev}}\left[\sum_{j\in T} w_{j,i} F_{j,i}(a_i, y)\right] \geq \mathbb{E}_{\mathcal{I}_{dev}}[u_i^U(a_i, y) + c_i - \epsilon_U] = \mathbb{E}_{\mathcal{I}_{dev}}[u_i^U(a_i, y)] + c_i - \epsilon_U \geq u_i^U(C^*_{S,T}(i)) + c_i - \epsilon_U.$$

Additionally:

$$\mathbb{E}_{\mathcal{I}_{NE}}\left[\sum_{j\in T} w_{j,i} F_{j,i}(a_i, y)\right] \leq \mathbb{E}_{\mathcal{I}_{NE}}[u_i^U(a_i, y) + c_i + \epsilon_A] = \mathbb{E}_{\mathcal{I}_{NE}}[u_i^U(a_i, y)] + c_i + \epsilon_U.$$

Combining these inequalities, we get:

$$u_i^U(C^*_{S,T}(i)) + c_i - \epsilon_U \leq \mathbb{E}_{\mathcal{I}_{NE}}[u_i^U(a_i, y)] + c_i + \epsilon_U + 2\epsilon_B \sum_{j\in T} \frac{w_{j,i}}{\lambda_{j,i}}.$$

The constant offset c_i cancels, and we are left with:

$$u_i^U(C^*_{S,T}(i)) - \epsilon_U \leq \mathbb{E}_{\mathcal{I}_{NE}}[u_i^U(a_i, y)] + \epsilon_U + 2\epsilon_P \sum_{j\in T} \frac{w_{j,i}}{\lambda_{j,i}}.$$

Thus:

$$\mathbb{E}_{\mathcal{I}_{NE}}[u_i^U(a_i, y)] \geq u_i^U(C^*_{S,T}(i)) - 2\epsilon_U - 2\epsilon_P \sum_{j\in T} \frac{w_{j,i}}{\lambda_{j,i}}.$$

which completes the proof. □

E PROOFS FROM SECTION 4

Proof of Theorem 4.1. Given ϵ and c, we construct an instance as follows. Consider a game with two providers and two users. Choose $M \in \mathbb{N}$ with $M > \frac{c}{\epsilon}$. The state y is drawn uniformly from $\{1, ..., M\}$. Each provider observes the state, i.e. $x_1^P = x_2^P = y$, while users observe nothing. Users have the action space $\mathcal{A}_1 = \mathcal{A}_2 = \{1, ..., M, \perp\}$. Suppose $R = 1$ (one round of conversation), and provider conversation rules map the observed state to a (randomized) signal in the state space.

In this game, the providers' utilities are:

$$u_1^P(a_1, a_2, y) = \frac{1}{D}\mathbf{1}[a_1 = y] + \mathbf{1}[a_2 = \perp]$$
$$u_2^P(a_1, a_2, y) = \mathbf{1}[a_1 = \perp] + \frac{1}{D}\mathbf{1}[a_2 = y]$$

where $D > 1$ is any fixed constant. The users' utilities are:

$$u_1^U(a_1, y) = c \cdot \mathbf{1}[a_1 = y] + \epsilon \cdot \mathbf{1}[a_1 = \perp]$$
$$u_2^U(a_2, y) = c \cdot \mathbf{1}[a_2 = y] + \epsilon \cdot \mathbf{1}[a_2 = \perp]$$

Note that the users' and providers' utilities exactly satisfy the Weak Market Alignment assumption. Furthermore, all utilities are between 0 and 1.

We show that if both providers choose conversation rules that reveal no information (e.g. the rule is: always signal 1, regardless of the observed state), this forms a Nash equilibrium in the public commitment game. Notice that if both providers choose to reveal no information, each user receives utility ϵ in expectation by choosing the action $\perp$ and utility $\frac{c}{M} < \epsilon$ in expectation by choosing any other action. Thus each user's utility-maximizing action is $\perp$, and each provider receives utility 1.

Consider any deviation by provider 1 to another conversation rule, resulting in a vector of provider conversation rules $C_{\vec{P}}$. Since both users have the same utility and same private information, $\mathcal{I}_1(C_{\vec{P}}) = \mathcal{I}_2(C_{\vec{P}})$. Let $p = \Pr_{\mathcal{I}_1(C_{\vec{P}})}[a_1 \neq \perp] = \Pr_{\mathcal{I}_2(C_{\vec{P}})}[a_2 \neq \perp]$. Then, Bob 1's expected utility is: $\frac{1}{D}\Pr_{\mathcal{I}_1(C_{\vec{P}})}[a_1 = y] + \Pr_{\mathcal{I}_2(C_{\vec{P}})}[a_2 = \perp] \leq \frac{p}{D} + 1 - p \leq 1$ (since $D > 1$). Thus, no

deviation can give any provider utility strictly higher than 1, and so the no disclosure signal profile is a Nash equilibrium.

Note as both providers have the same private information, in this case the optimal shared conversation rule is the rule which reveals y. This induces the action $a_i = y$ and gives both users expected utility c. This completes our proof. □

Proof of Theorem 4.3. Fix any Nash equilibrium strategy profile $C_{\vec{P}}$ and let $\mathcal{I}_{NE} = \mathcal{I}(C_{\vec{P}})$. Suppose there is a user i such that her utility is strictly less than $u_i^U(C^*_{S,T}(i)) - \frac{2\epsilon}{\lambda_i^*}$. We show that this implies the existence of some Bob who has a profitable deviation, thus contradicting the premise that $C_{\vec{P}}$ is a Nash equilibrium.

Let $\tilde{j} = \arg\max_{j \in T} \lambda_{j,i}$. Thus $\lambda_i^* = \lambda_{\tilde{j},i}$. Consider a deviation by provider $\tilde{j}$ to the user-dominant conversation rule $C^*_{P_{\tilde{j}}}$. Let $\mathcal{I}_{dev} = \mathcal{I}(C^*_{P_{\tilde{j}}}, C_{P_{-\tilde{j}}})$ be the distribution induced by this deviation. By definition of $C^*_{P_{\tilde{j}}}$, $u_i^U(C^*_{P_{\tilde{j}}}; \tilde{j}) \geq u_i^U(C^*_S(i))$, so user i's utility in equilibrium must be strictly less than $u_i^U(C^*_{P_{\tilde{j}}}; \tilde{j}) - \frac{2\epsilon}{\lambda_{\tilde{j},i}}$. Thus, under this deviation, user i will now choose to interact with provider $\tilde{j}$ and receive utility $u_i^U(C^*_{P_{\tilde{j}}}; \tilde{j})$, increasing her utility by $> \frac{2\epsilon}{\lambda_{\tilde{j},i}}$.

Now, by the user-dominance of $C^*_{P_{\tilde{j}}}$, we have that for all other users i', $u_{i'}^U(C^*_{P_{\tilde{j}}}; \tilde{j}) \geq u_{i'}^U(C_{P_{\tilde{j}}}; \tilde{j})$. That is, users' utilities can only (weakly) increase under this deviation. Then, applying approximate Strong Market Alignment, we have:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{I}_{dev}}[u^P_{\tilde{j}}(a_{1:n}, y)] &\geq \mathbb{E}_{\mathcal{I}_{dev}}\left[\sum_{i \in N} \lambda_{\tilde{j},i} u_i^U(a_i, y) + c_{\tilde{j}} - \epsilon\right] \\
&= \sum_{i \in N} \lambda_{\tilde{j},i} \mathbb{E}_{\mathcal{I}_{dev}}\left[u_i^U(a_i, y)\right] + c_{\tilde{j}} - \epsilon \\
&> \lambda_{\tilde{j},i}\left(\mathbb{E}_{\mathcal{I}_{NE}}\left[u_i^U(a_i, y)\right] + \frac{2\epsilon}{\lambda_{\tilde{j},i}}\right) + \sum_{i' \neq i} \lambda_{\tilde{j},i'} \mathbb{E}_{\mathcal{I}_{NE}}\left[u_{i'}^U(a_{i'}, y)\right] + c_{\tilde{j}} - \epsilon \\
&= \sum_{i \in N} \lambda_{\tilde{j},i} \mathbb{E}_{\mathcal{I}_{NE}}\left[u_i^U(a_i, y)\right] + c_{\tilde{j}} + \epsilon \\
&= \mathbb{E}_{\mathcal{I}_{NE}}\left[\sum_{i \in N} \lambda_{\tilde{j},i} u_i^U(a_i, y) + c_{\tilde{j}}\right] + \epsilon \\
&\geq \mathbb{E}_{\mathcal{I}_{NE}}\left[u^P_{\tilde{j}}(a_{1:n}, y) - \epsilon\right] + \epsilon \\
&= \mathbb{E}_{\mathcal{I}_{NE}}\left[u^P_{\tilde{j}}(a_{1:n}, y)\right]
\end{aligned}
$$

where the first and final inequalities follow from approximate Strong Market Alignment, and the third inequality follows from the users' increases in utility under the deviation. Therefore, provider $\tilde{j}$ can strictly increase his utility by deviating to $C^*_{P_{\tilde{j}}}$, and so $C_{\vec{P}}$ cannot be a Nash equilibrium. This completes the proof.

□

Proof of Theorem E.1. Consider any Nash equilibrium strategy profile $C_{\vec{P}}$. Suppose that there is a user i^* such that her utility is strictly less than $u_{i^*}^U(C^*_{S,T,R-1}(i^*)) - \delta_{i^*}$. We show that this implies the existence of some provider who has a profitable deviation, thus contradicting the premise that $C_{\vec{P}}$ is a Nash equilibrium.

Recall that under the Strong Market Alignment condition, for every provider j's utility:

$$
\left| u^P_j(a_{1:n}, y) \; - \left(\sum_{i \in N} \lambda_{j,i} \, u_i^U(a_i, y) + c_j\right)\right| \leq \epsilon
$$

where $\lambda_{j,i} \geq 0$ and $c_j \in \mathbb{R}$ for all i. Furthermore, for all users i, there exists at least one provider j with $\lambda_{j,i} > 0$. Let $j^* = \arg\min_{j:\lambda_{j,i^*}>0} \left(\Delta_R(j) \frac{\sum_{i' \neq i^*} \lambda_{j,i'}}{\lambda_{j,i^*}} + \frac{2\epsilon}{\lambda_{j,i^*}}\right)$.

Suppose provider j^* deviates to the following conversation rule $C'_{P_{j^*}}$: on the first round, every user i self-selects into one of $|N|$ "branches," using the message space of size $\geq |N|$. In the remaining $R-1$ rounds, provider j^* implements $C^*_{S,T,R-1}(i)$ in each branch i. Note that this is a valid R-round conversation rule. Let $C^*_{\vec{P}} = (C'_{P_{j^*}}, C_{P_{-j^*}})$ be the profile of provider strategies under this unilateral deviation.

As all users are strategic, any user i who does interact with provider j^* will either declare her own type and get utility $u_i^U(C^*_{S,T,R-1}(i))$, or declare another type and get this same optimal $(R-1)$-round utility $u_i^U(C^*_{S,T,R-1}(i))$.

Observe that if user i^* interacts with provider j^* under $C^*_{\vec{P}}$ and self-selects into branch i^*, her expected utility will be $u_{i^*}^U(C^*_{S,T,R-1}(i^*))$, and so it will have increased by at least δ_{i^*}.

Furthermore, every user $i \neq i^*$ has her utility decreased by at most $\Delta_R(j^*)$. To see this, note that there are three cases for each user i:

- In $C_{\vec{P}}$, user i is interacting with provider j^*. Then her utility under $C_{\vec{P}}$ is at most $u_i^U(C^*_R(i); j)$, and her utility under $C^*_{\vec{P}}$ is lower by at most $\Delta_R(j^*)$.

- In $C_{\vec{P}}$, user i is interacting with some provider $j \neq j^*$, and in $C^*_{\vec{P}}$ she is also interacting with some provider $j \neq j^*$. Then, as the other providers have not changed their conversation rules, user i's utility does not decrease at all.

- In $C_{\vec{P}}$, user i is interacting with some provider $j \neq j^*$, and in $C^*_{\vec{P}}$ she is interacting with provider j^*. She could still have gotten the same expected utility from provider j under $C^*_{\vec{P}}$, so her utility must have weakly *increased* under the deviation to $C^*_{\vec{P}}$.

Thus, every other user i's expected utility can only decrease by at most $\Delta_R(j^*)$ by assumption. So, we can compute the change in provider j^*'s utility:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{I}(C'_{P_{j^*}}, C_{P_{-j^*}})}[u^P_{j^*}(a_{1:n}, y)] - \mathbb{E}_{\mathcal{I}(C_{\vec{p}})}[u^P_{j^*}(a_{1:n}, y)] \\
&\geq \mathbb{E}_{\mathcal{I}(C'_{P_{j^*}}, C_{P_{-j^*}})}\left[\lambda_{j^*,i^*} u^U_{i^*}(a_{i^*}, y) + \sum_{i \in N \neq i^*} \lambda_{j^*,i} u^U_i(a_i, y) + c_j - \epsilon\right] \\
&\quad - \mathbb{E}_{\mathcal{I}(C_{\vec{P}})}\left[\lambda_{j^*,i^*} u^U_{i^*}(a_{i^*}, y) + \sum_{i \in N \neq i^*} \lambda_{j^*,i} u^U_i(a_i, y) + c_j + \epsilon\right] \\
&= \mathbb{E}_{\mathcal{I}(C'_{B_{j^*}}, C_{P_{-j^*}})}\left[\lambda_{j^*,i^*} u^U_{i^*}(a_{i^*}, y) + \sum_{i \in N \neq i^*} \lambda_{j^*,i} u^U_i(a_i, y) + c_j\right] \\
&\quad - \mathbb{E}_{\mathcal{I}(C_{\vec{P}})}\left[\lambda_{j^*,i^*} u^U_{i^*}(a_{i^*}, y) + \sum_{i \in N \neq i^*} \lambda_{j^*,i} u^U_i(a_i, y) + c_j\right] - 2\epsilon \\
&= \lambda_{j^*,i^*}\left(\mathbb{E}_{\mathcal{I}(C'_{P_{j^*}}, C_{P_{-j^*}})}\left[u^U_{i^*}(a_{i^*}, y)\right] - \mathbb{E}_{\mathcal{I}(C_{\vec{P}})}\left[u^U_{i^*}(a_{i^*}, y)\right]\right) \\
&\quad + \sum_{i \in N \neq i^*} \lambda_{j^*,i}\left(\mathbb{E}_{\mathcal{I}(C'_{P_{j^*}}, C_{P_{-j^*}})}\left[u^U_i(a_i, y)\right] - \mathbb{E}_{\mathcal{I}(C_{\vec{B}})}\left[u^U_i(a_i, y)\right]\right) - 2\epsilon \\
&> \lambda_{j^*,i^*}\delta - \Delta_R(j^*)\sum_{i \neq i^*} \lambda_{j^*,i} - 2\epsilon \\
&= \lambda_{j^*,i^*}\left(\Delta_R(j^*)\frac{\sum_{i \neq i^*} \lambda_{j^*,i}}{\lambda_{j^*,i^*}} + \frac{2\epsilon}{\lambda_{j^*,i^*}}\right) - \Delta_R(j^*)\sum_{i \neq i^*} \lambda_{j^*,i} - 2\epsilon \\
&= 0
\end{aligned}
$$

Therefore, provider j^* can increase his expected utility by deviating to $C'_{P^*_j}$, and the strategy profile $C_{\vec{P}}$ cannot be a Nash equilibrium. This completes the proof. □

E.1 Beyond User-Dominant conversation rules

We now extend positive results in the anonymous game with Strong Market Alignment beyond settings where a user-dominant conversation rule is available. The key observation is: any provider can always use the first round to elicit a user's identity (which is feasible whenever the message space has size at least N), and use the remaining $R-1$ rounds to implement that user's optimal shared $(R-1)$-conversation rule. If Strong Market Alignment holds, and sacrificing one round of conversation is not too harmful, then providers have incentive to adopt this rule. In effect, this strategy recovers personalization — with the caveat that users need not report their identities truthfully. We show that under this rule, users have no incentive to lie.

To prove this result, then, we need to quantify the marginal value of a single additional round for a user. For any provider j, let $u^U_i(C^*_R(i); j) = \max_{C \in \mathcal{C}_R}(u^U_i(C; j))$ denote user i's utility under her favorite R-round conversation rule offered by provider j. Next, define $\Delta_R(j) = \max_{i \in N}\left(u^U_i(C^*_R(i); j) - u^U_i(C^*_{S,T,R-1}(i))\right)$ to be the largest gain (across all users) from interacting with the best R-round conversation rule deployed by provider j versus the best $(R-1)$-round T-shared conversation rule. Intuitively, if R is large, then the marginal value of a single round — and hence $\Delta_R(j)$ — is small.

Theorem E.1. *Suppose there exists a set of providers $T \subseteq K$ such that (T, ϵ)-Strong Market Alignment is satisfied with weights $(\lambda_{j,i})_{i \in N}$ for every provider $j \in T$. Assume $|\mathcal{M}| \geq N$. Then, for each user i, her expected utility in any Nash equilibrium is at least $u^U_i(C^*_{S,T,R-1}(i)) - \delta_i$, where $\delta_i = \min_{j:\lambda_{j,i}>0}\left(\Delta_R(j)\frac{\sum_{i' \neq i}\lambda_{j,i'}}{\lambda_{j,i}} + \frac{2\epsilon}{\lambda_{j,i}}\right)$.*

F PROOFS FROM SECTION A

Proof of Theorem A.1. Let $C_{\vec{P}}$ be any Nash equilibrium strategy profile of the augmented personalized game, and let $\mathcal{I}_{NE} = \mathcal{I}(C_{\vec{P}})$ be the induced distribution. Now, assume for contradiction that there exists a user $i \in N$ such that her utility is strictly lower than $u_i^U(C^*_{S,T}(i)) - 2\epsilon_U - 2\epsilon_P\mu_i$. Consider a deviation by any provider $j \in T$ in his choice of conversation rule for user i from $C_{P_j}(i)$ to $C^*_{S,T}(i)$, the optimal shared conversation rule for user i. Note that by definition, any provider could implement $C^*_S(i)$.

Let $\mathcal{I}_{dev} = \mathcal{I}(C^*_{S,T}(i), C_{P_j}(-i), C_{P_{-j}})$ be the distribution induced by the deviation of provider j in his chosen conversation rule for user i (the conversation rules chosen by provider j for all other users, $C_{P_j}(-i)$, remain unchanged). As user i's utility was previously worse than $u_i^U(C^*_S(i))$, she will now choose to interact with the provider who implements $C^*_{S,T}(i)$, leading to the outcome under $\mathcal{I}_{dev}$.

Now, by the Nash equilibrium condition, no provider j has an incentive to deviate. Thus, for all providers $j \in T$:

$$\mathbb{E}_{\mathcal{I}_{dev}}[u_j^P(a_{1:n+1}, y)] \leq \mathbb{E}_{\mathcal{I}_{NE}}[u_j^P(a_{1:n+1}, y)].$$

Since every user observes only the conversation rules chosen for them, this deviation can only change the induced distribution for user i (in particular the distribution over her actions). Thus, using our assumption on how provider utilities change under an additional user and applying approximate Weak Market Alignment, we have that for any provider $j \in T$:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{I}_{dev}}[(1-\beta_j)\cdot u_j^P(a_{1:n}, y) + \beta_j \cdot f(a_{n+1}, y)] \leq \mathbb{E}_{\mathcal{I}_{NE}}[(1-\beta_j)\cdot u_j^P(a_{1:n}, y) + \beta_j \cdot f(a_{n+1}, y)] \\
\iff &\mathbb{E}_{\mathcal{I}_{dev}}\left[(1-\beta_j)\cdot\left(\sum_{i\in N}\lambda_{j,i}F_{j,i}(a_i, y) + c_j - \epsilon_P\right) + \beta_j \cdot f(a_{n+1}, y)\right] \\
&\leq \mathbb{E}_{\mathcal{I}_{NE}}\left[(1-\beta_j)\cdot\left(\sum_{i\in N}\lambda_{j,i}F_{j,i}(a_i, y) + c_j + \epsilon_P\right) + \beta_j \cdot f(a_{n+1}, y)\right] \\
\iff &\mathbb{E}_{\mathcal{I}_{dev}}\left[(1-\beta_j)\cdot\left(\lambda_{j,i}F_{j,i}(a_i, y) + \sum_{i'\in N: i'\neq i}\lambda_{j,i'}F_{j,i'}(a_{i'}, y) + c_j - \epsilon_P\right) + \beta_j \cdot f(a_{n+1}, y)\right] \\
&\leq \mathbb{E}_{\mathcal{I}_{NE}}\left[(1-\beta_j)\cdot\left(\lambda_{j,i}F_{j,i}(a_i, y) + \sum_{i'\in N: i'\neq i}\lambda_{j,i'}F_{j,i'}(a_{i'}, y) + c_j + \epsilon_P\right) + \beta_j \cdot f(a_{n+1}, y)\right] \\
\iff &(1-\beta_j)\cdot\left(\mathbb{E}_{\mathcal{I}_{dev}}\left[\lambda_{j,i}F_{j,i}(a_i, y)\right] + \mathbb{E}_{\mathcal{I}_{NE}}\left[\sum_{i'\in N: i'\neq i}\lambda_{j,i'}F_{j,i'}(a_{i'}, y)\right] + c_j - \epsilon_P\right) + \beta_j \cdot \mathbb{E}_{\mathcal{I}_{NE}}[f(a_{n+1}, y)] \\
&\leq (1-\beta_j)\cdot\left(\mathbb{E}_{\mathcal{I}_{NE}}\left[\lambda_{j,i}F_{j,i}(a_i, y)\right] + \mathbb{E}_{\mathcal{I}_{NE}}\left[\sum_{i'\in N: i'\neq i}\lambda_{j,i'}F_{j,i'}(a_{i'}, y)\right] + c_j + \epsilon_P\right) + \beta_j \cdot \mathbb{E}_{\mathcal{I}_{NE}}[f(a_{n+1}, y)] \\
\iff &(1-\beta_j)\cdot\lambda_{j,i}\mathbb{E}_{\mathcal{I}_{dev}}\left[F_{j,i}(a_i, y)\right] \leq (1-\beta_j)\cdot\lambda_{j,i}\mathbb{E}_{\mathcal{I}_{NE}}\left[F_{j,i}(a_i, y)\right] + 2(1-\beta_j)\cdot\epsilon_P
\end{aligned}$$

where the third step follows from the fact that only the induced distribution over user i's actions changes under the deviation. Since $\beta_j \in (0, 1)$, for any j with $\lambda_{j,i} > 0$, we thus have:

$$\mathbb{E}_{\mathcal{I}_{dev}}\left[F_{j,i}(a_i, y)\right] \leq \mathbb{E}_{\mathcal{I}_{NE}}\left[F_{j,i}(a_i, y)\right] + 2\frac{\epsilon_P}{\lambda_{j,i}}$$

Taking a weighted sum over all providers using the non-negative weights $(w_{j,i})_{j\in T}$ from the Weak Market Alignment condition:

$$\sum_{j\in T} w_{j,i}\mathbb{E}_{\mathcal{I}_{dev}}[F_{j,i}(a_i, y)] \leq \sum_{j\in T} w_{j,i}\mathbb{E}_{\mathcal{I}_{NE}}[F_{j,i}(a_i, y)] + 2\epsilon_P\sum_{j\in T}\frac{w_{j,i}}{\lambda_{j,i}}.$$

Note that by assumption, if $w_{j,i} > 0$, then $\lambda_{j,i} > 0$, and so we can apply the inequality here. By linearity of expectation, this is equivalent to:

$$\mathbb{E}_{\mathcal{I}_{dev}}\left[\sum_{j\in T} w_{j,i}F_{j,i}(a_i, y)\right] \leq \mathbb{E}_{\mathcal{I}_{NE}}\left[\sum_{j\in T} w_{j,i}F_{j,i}(a_i, y)\right] + 2\epsilon_P \sum_{j\in T}\frac{w_{j,i}}{\lambda_{j,i}}.$$

Now we use the market alignment assumption, which states that $\sum_j w_{j,i}F_{j,i}(a_i, y)$ can be approximated by $u_i^U(a_i, y) + c_i$. We have that:

$$\mathbb{E}_{\mathcal{I}_{dev}}\left[\sum_{j\in T} w_{j,i}F_{j,i}(a_i, y)\right] \geq \mathbb{E}_{\mathcal{I}_{dev}}[u_i^U(a_i, y)+c_i-\epsilon_U] = \mathbb{E}_{\mathcal{I}_{dev}}[u_i^U(a_i, y)]+c_i-\epsilon_U \geq u_i^U(C^*_{S,T}(i))+c_i-\epsilon_U.$$

Additionally:

$$\mathbb{E}_{\mathcal{I}_{NE}}\left[\sum_{j\in T} w_{j,i}F_{j,i}(a_i, y)\right] \leq \mathbb{E}_{\mathcal{I}_{NE}}[u_i^U(a_i, y) + c_i + \epsilon_U] = \mathbb{E}_{\mathcal{I}_{NE}}[u_i^U(a_i, y)] + c_i + \epsilon_U.$$

Combining these inequalities, we get:

$$u_i^U(C^*_{S,T}(i)) + c_i - \epsilon_U \leq \mathbb{E}_{\mathcal{I}_{NE}}[u_i^U(a_i, y)] + c_i + \epsilon_U + 2\epsilon_P \sum_{j\in T}\frac{w_{j,i}}{\lambda_{j,i}}.$$

The constant offset c_i cancels, and we are left with:

$$u_i^U(C^*_{S,T}(i)) - \epsilon_U \leq \mathbb{E}_{\mathcal{I}_{NE}}[u_i^U(a_i, y)] + \epsilon_U + 2\epsilon_P \sum_{j\in T}\frac{w_{j,i}}{\lambda_{j,i}}.$$

Thus:

$$\mathbb{E}_{\mathcal{I}_{NE}}[u_i^U(a_i, y)] \geq u_i^U(C^*_{S,T}(i)) - 2\epsilon_U - 2\epsilon_P \sum_{j\in T}\frac{w_{j,i}}{\lambda_{j,i}}.$$

which completes the proof. □

Proof of Theorem A.2. Consider a game with $R = 1$ rounds with one user and one provider. Note that as we have only a single provider, computing an equilibrium between providers reduces to computing a best action for that provider. Let $x_1^U = \emptyset$, and let $x_1^P = y$. Furthermore, assume that the message space is large enough to encode y. Thus, full disclosure is possible. Let there be two states y_1 and y_2 which each occur with probability $\frac{1}{2}$, and let the user have action set $\{y_1, y_2, \perp\}$. Her utility is

$$u_1^U = \mathbb{I}[a_1 = y] + \frac{2}{3}\mathbb{I}[a_1 = \perp]$$

Meanwhile, the provider's utility is

$$u_1^P = \frac{1}{2}\mathbb{I}[a_1 = y] + \frac{1}{3}\mathbb{I}[a_1 = \perp]$$

Note that this instance satisfies exact Strong Market Alignment: the provider will disclose y in his conversation rule, and the user will always match the state and get expected utility $u_i^U(C^*_{S,K}) = 1$. Now, consider the instance $A(G, f_1, \frac{1}{2}, \{1, 2, \perp\}, u_2^U, \emptyset, P(\emptyset))$. Here $f_1 = \mathbb{I}[a_2 = \perp]$, and thus the provider's utility $\hat{u}_1^P$ in this augmented game is

$$\hat{u}_1^P = \frac{1}{4}\mathbb{I}[a_1 = y] + \frac{1}{6}\mathbb{I}[a_1 = \perp] + \frac{1}{2}\mathbb{I}[a_2 = \perp]$$

Additionally, $u_2^A = \mathbb{I}[a_2 = y] + \frac{2}{3}\mathbb{I}[a_2 = \perp]$.

Now, we will show that under any best action from the provider, the user 1 has expected utility $\frac{2}{3} < u_i^U(C^*_{S,K})$. To see this, note that if the provider does not provide any information to the users, both users will select action $\perp$. In this case, the provider's utility is $\frac{1}{6} + \frac{1}{2} = \frac{2}{3}$. Furthermore, note that as both users have the same hidden information and utilities, they will have the same

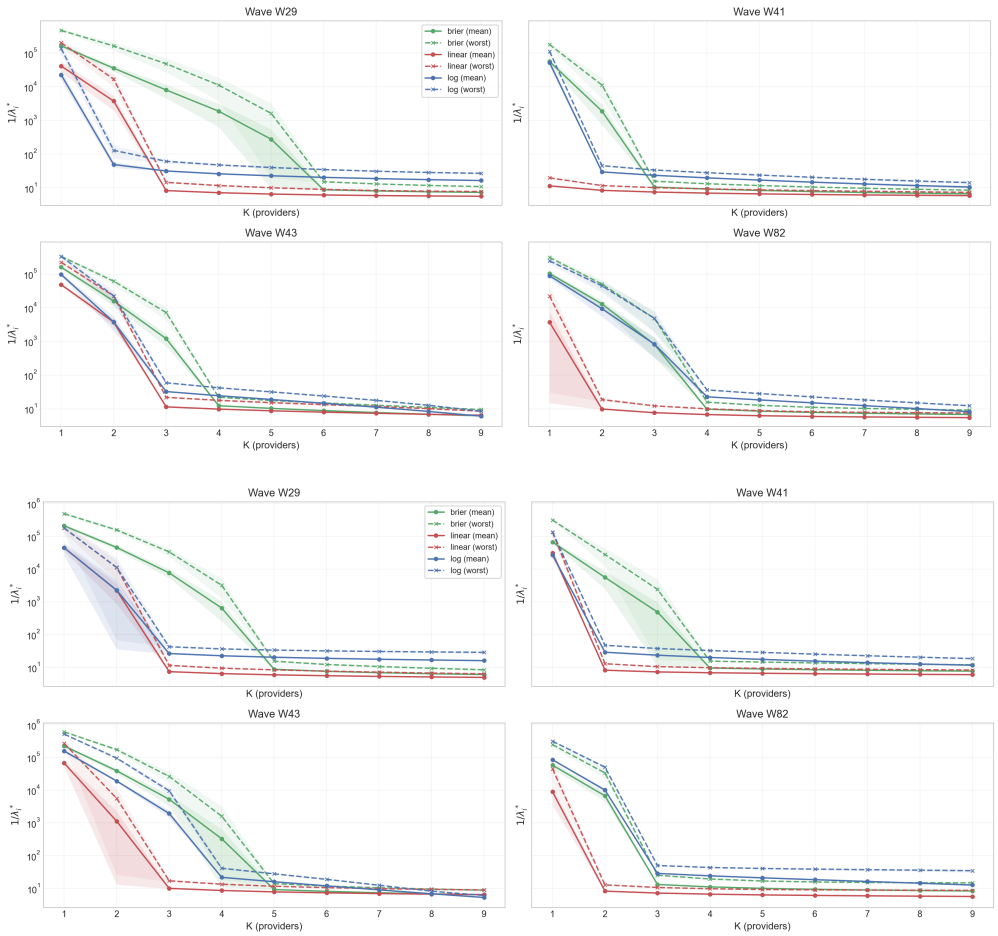

Figure 3: **Transfer factor vs. market size (log/Brier/linear; Education and Political ideology).** Panels: W29 (gender/health, 2017), W41 (future outlook, 2018), W43 (personal well-being, 2019), W82 (economy/democracy, 2021). Curves show different scoring rules (mean and worst across users). Shaded bands: CV standard error.

distribution over actions against any provider conversation rule. Now, consider the case where the provider commits to a conversation rule where, if the users best-respond, each user has probability $p > 0$ of playing something other than $\perp$. If both users plays something other than $\perp$, the provider's utility is at most $\frac{1}{4}$. So his expected utility is at most

$$\frac{1}{4}p + (1-p)\frac{2}{3} < \frac{2}{3}$$

This is strictly less than that of providing no information, so the best response to the provider's conversation rule for user 1 will always be to play $\perp$. Thus, her expected utility is at most $\frac{2}{3}$, completing the proof. □

G Additional Experiments

Additional scoring rules. The main text uses the *linear* score $s(p, a) = p(a)$. Here we replicate all experiments with two alternative scoring rules: the *log* score, which emphasizes assigning high probability mass to the realized option (likelihood), and the *Brier* (quadratic) score, which penalizes miscalibration and overconfidence. All plots use the same cross-validation protocol and shaded standard-error bands as in the main text.

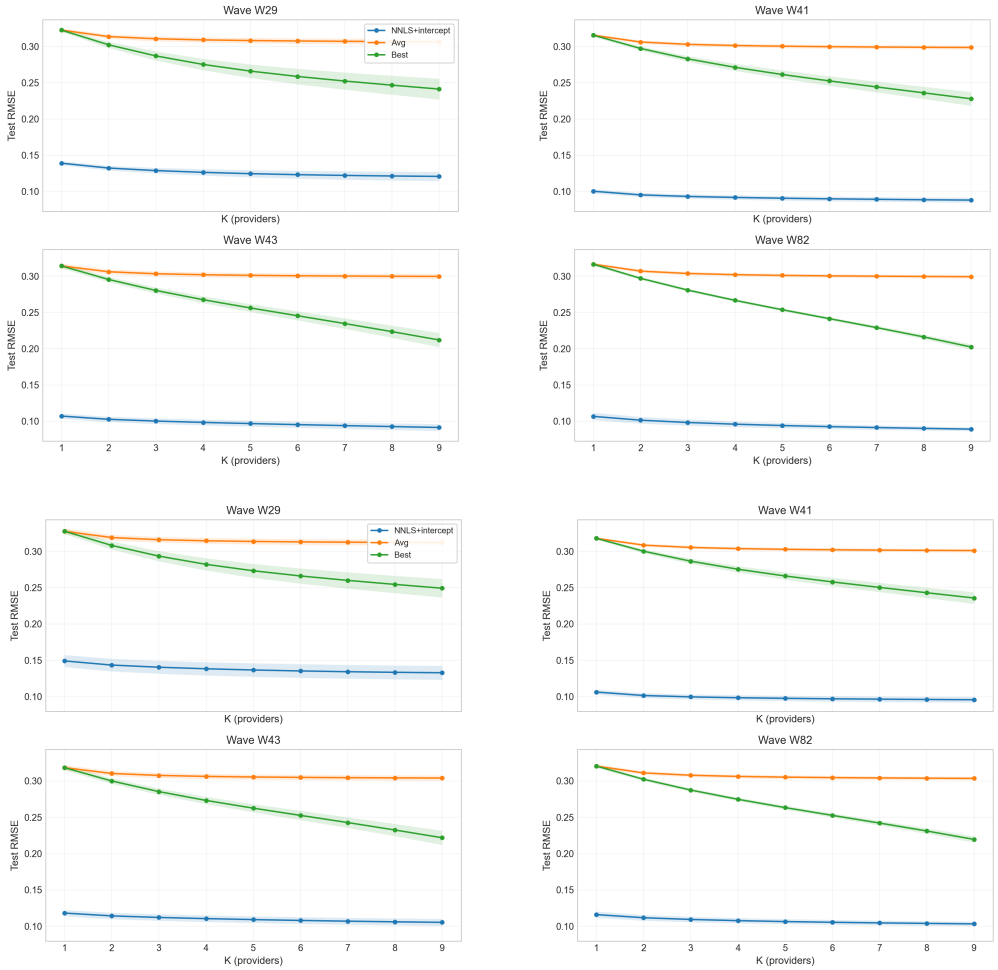

Figure 4: **Weak Market Alignment (user alignment; log score; Education and Political ideology).** Panels: W29 (gender/health, 2017), W41 (future outlook, 2018), W43 (personal well-being, 2019), W82 (economy/democracy, 2021). Shaded bands: CV standard error.

Additional takeaways. Across scoring rules (log, Brier, linear), the qualitative picture is stable: increasing K improves user-alignment fits and rapidly reduces worst-case transfer at small K, with diminishing returns thereafter. Across demographic partitions (Race/Education/Political ideology/Sex), the same patterns hold, but the identity of the hardest-to-cover user and the magnitude of transfer can vary by partition and wave.

Equal-weight averaging. In some demographic partitions (e.g., Race, Sex), equal-weight averaging approaches the NNLS solution at large K, suggesting more homogeneous model behavior for these groups. For more politically polarized partitions (Political ideology), optimized weighting provides larger improvements over simple averaging, reflecting greater model differentiation.

Interpreting learned weights. To understand which models contribute most to each group's approximation (Weak alignment) and which groups each model "covers" (Strong alignment), we visualize the learned weights. For each wave we fit on each CV train fold (5-fold over questions) and report the median weight across folds. For weak alignment, the weights $w_{j,i}$ indicate how much model j contributes to approximating user i's utility. For strong alignment, the weights $\lambda_{j,i}$ indicate how much model j's utility depends on user i's utility (i.e., "coverage").

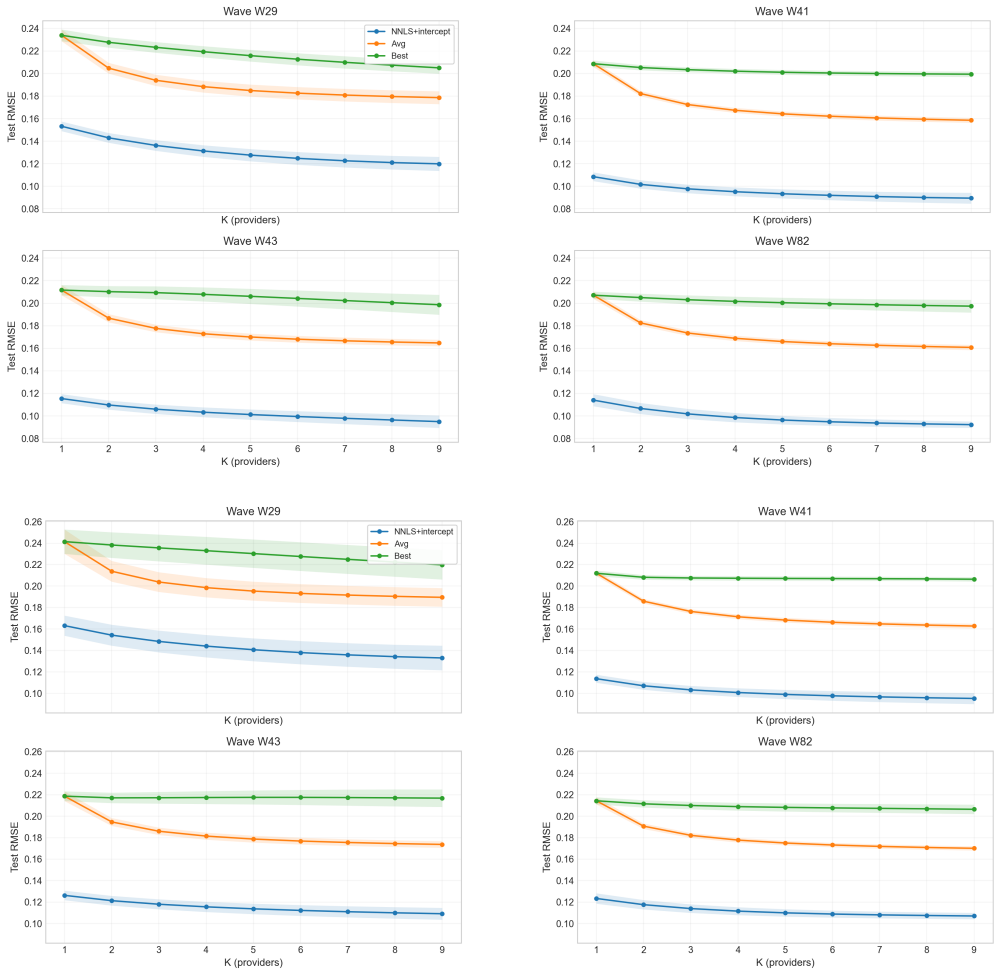

Figure 5: **Weak Market Alignment (user alignment; Brier score; Education and Political ideology).** Panels: W29 (gender/health, 2017), W41 (future outlook, 2018), W43 (personal well-being, 2019), W82 (economy/democracy, 2021). Shaded bands: CV standard error.

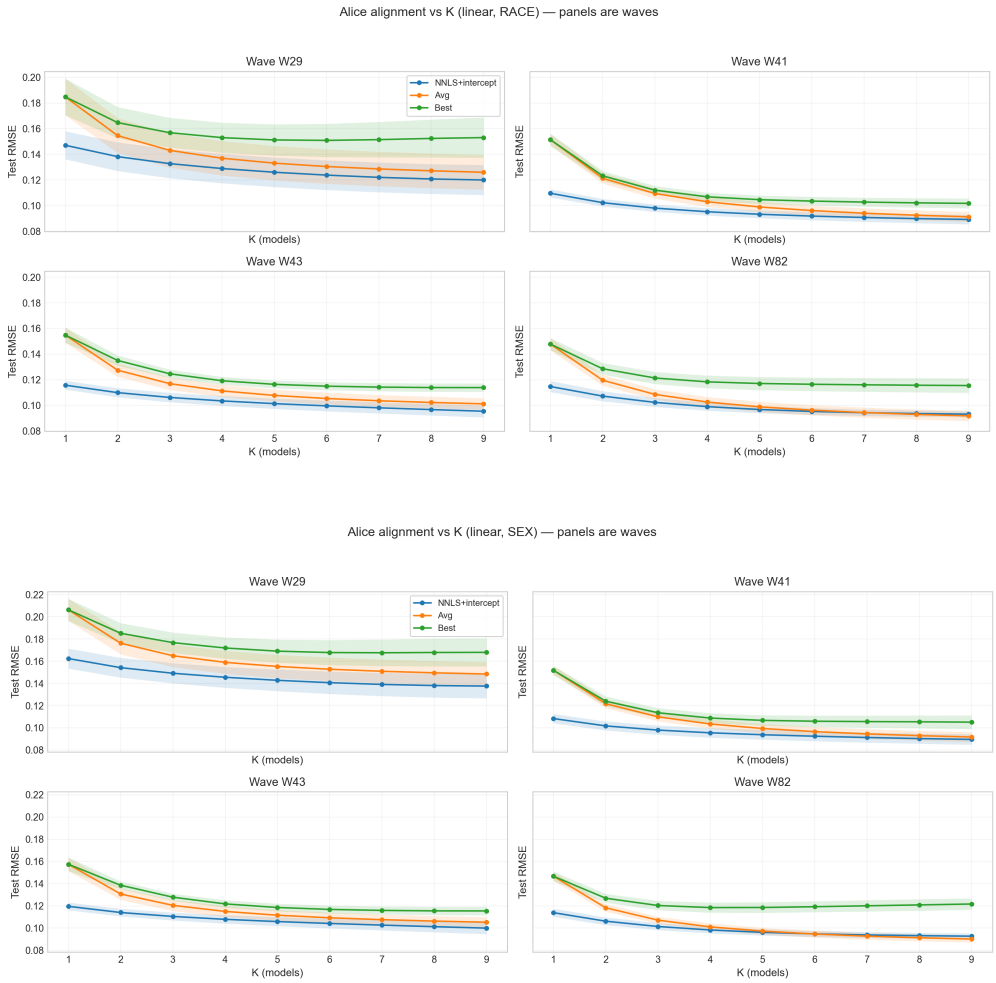

Figure 6: **Weak Market Alignment (user alignment; linear; Race and Sex).** Panels: W29 (gender/health, 2017), W41 (future outlook, 2018), W43 (personal well-being, 2019), W82 (economy/democracy, 2021). Shaded bands: CV standard error.

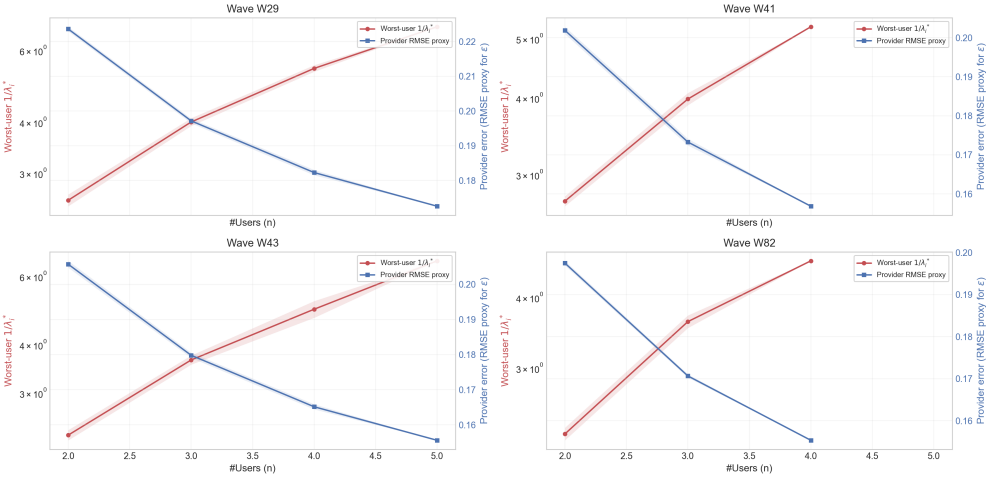

Figure 7: **Error–coverage trade-off (Race).** Worst-user transfer factor (red) vs. fitting error (blue) as number of users increases.

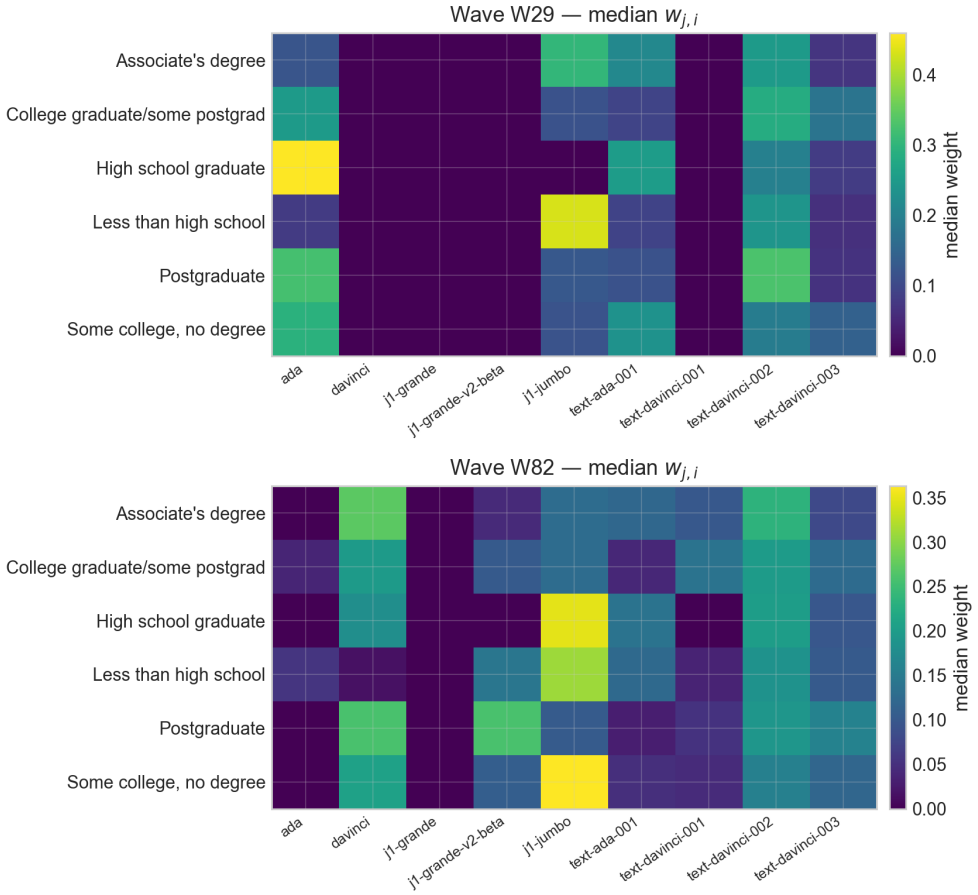

Figure 8: **Weak alignment weights $w_{j,i}$ (linear score; Education).** Rows are users (demographic groups) and columns are models. Panels: W29 (gender/health, 2017) and W82 (economy/democracy, 2021); heatmaps show the median weight across CV folds.

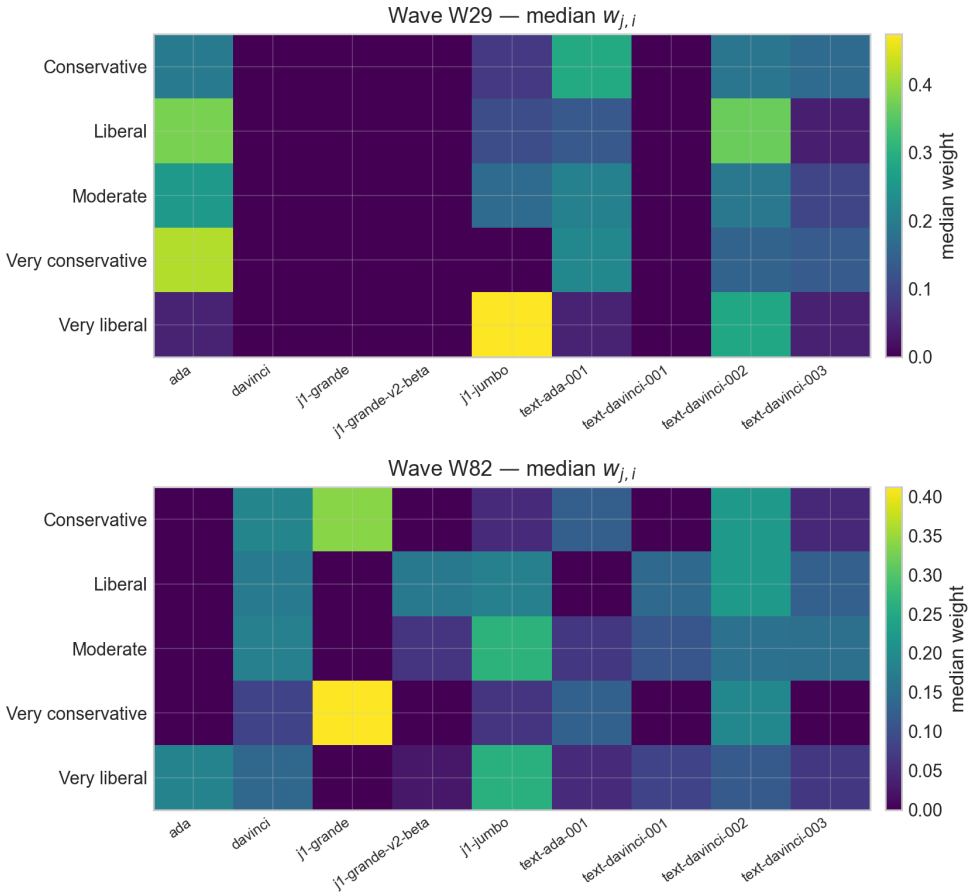

Figure 9: **Weak alignment weights $w_{j,i}$ (linear score; Political ideology).** Rows are users (demographic groups) and columns are models. Panels: W29 (gender/health, 2017) and W82 (economy/democracy, 2021); heatmaps show the median weight across CV folds.

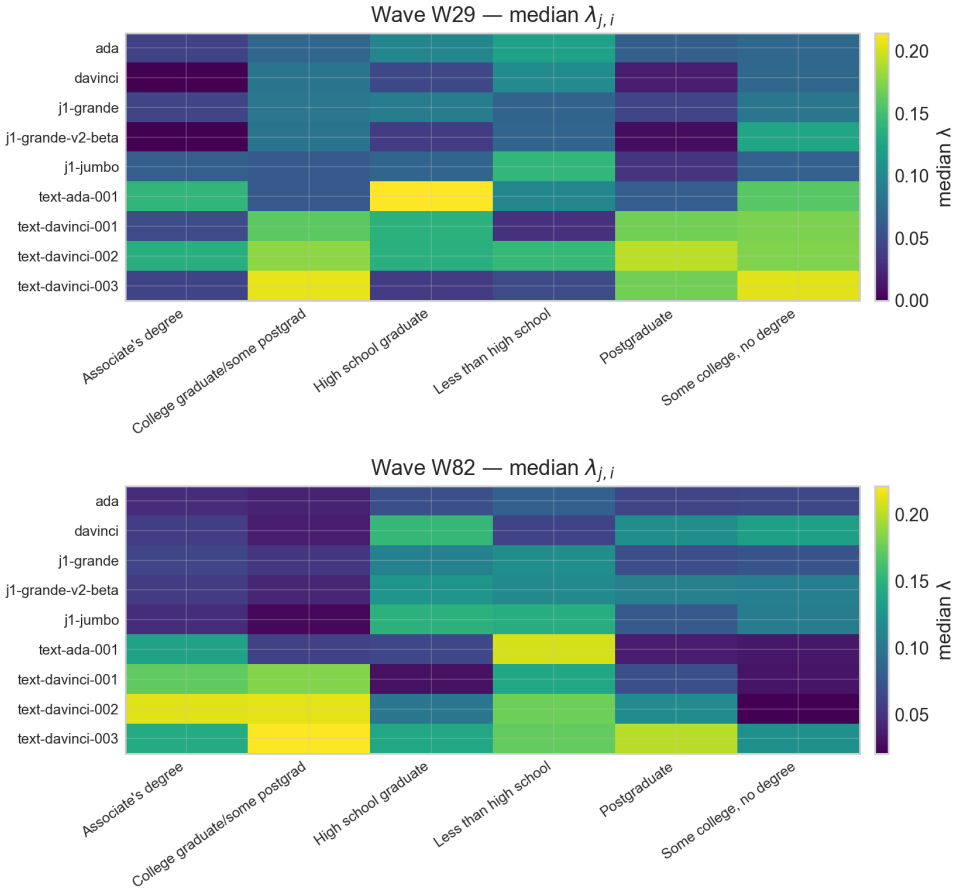

Figure 10: **Strong alignment weights $\lambda_{j,i}$ (linear score; Education).** Rows are models and columns are users. Panels: W29 (gender/health, 2017) and W82 (economy/democracy, 2021); heatmaps show the median $\lambda_{j,i}$ across CV folds.

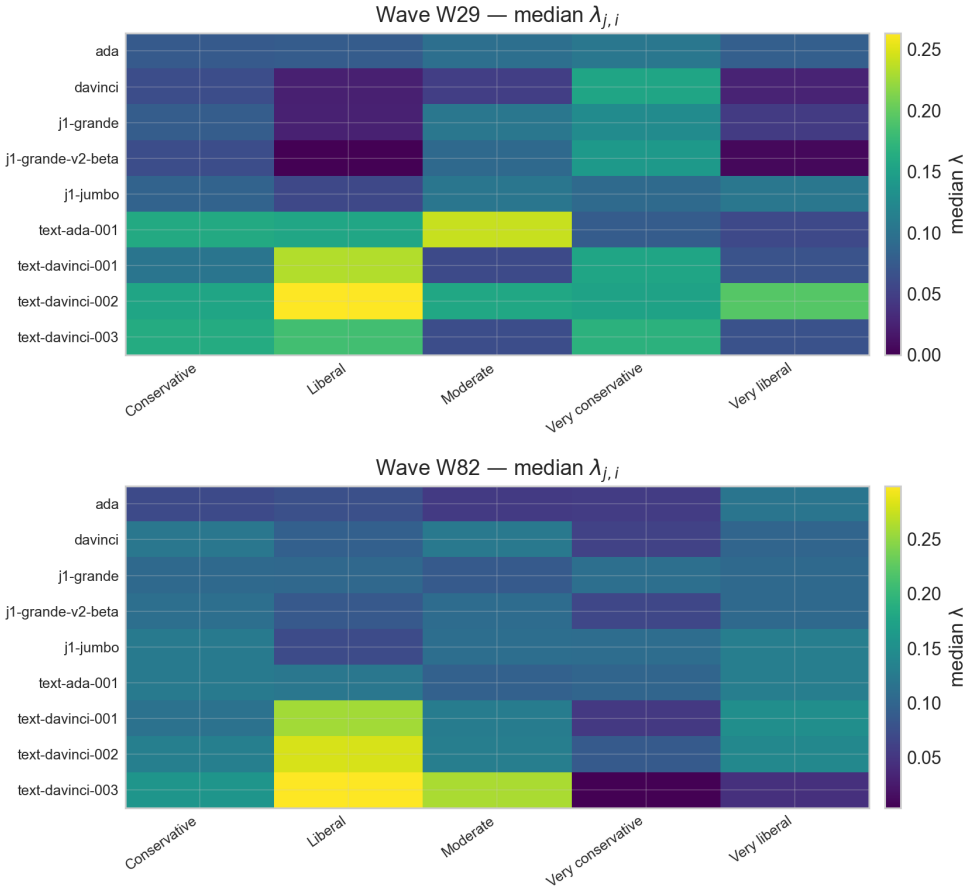

Figure 11: **Strong alignment weights $\lambda_{j,i}$ (linear score; Political ideology).** Rows are models and columns are users. Panels: W29 (gender/health, 2017) and W82 (economy/democracy, 2021); heatmaps show the median $\lambda_{j,i}$ across CV folds.