Monoculture or Multiplicity: Which Is It?

Mila Gorecki **Moritz Hardt**

Max Planck Institute for Intelligent Systems, Tübingen
Tübingen AI Center
{mgorecki, hardt}@tuebingen.mpg.de

Abstract

Two narratives about machine learning ecosystems grew out of the recent algorithmic fairness discourse. In one, dubbed *monoculture*, algorithmic ecosystems tend toward homogeneity akin to a single model making all decisions. Individuals then face the risk of systematic exclusion with no recourse. In the other, *model multiplicity*, many models solve the same task with similar accuracy, causing excessive variation in individual outcomes. Both narratives are compelling, yet, seemingly at odds: model multiplicity can't materialize in a strict monoculture. In this work, we conduct a comprehensive empirical evaluation to test both claims. We work from the premise that increasingly decision makers will use large language models for consequential prediction tasks. We therefore examine 50 language models, open source models ranging in size from 1B to 141B parameters and state-of-the-art commercial models, under 4 different prompt variations, and across 6 different prediction tasks. Evaluating both new and old quantitative measures of monoculture and multiplicity, we find the empirical landscape sits between the two extremes. Each narrative finds some empirical support, but neither tightly fits the observations. Systematic exclusion with no recourse is rare, but model similarity is real. Even when starting from a single model, prompt variation induces some diversity in predictions. Our results contribute empirical insights that illuminate the underappreciated middle ground between monoculture and multiplicity.

1 Introduction

Bureaucratic decisions have always provoked two distinct anxieties. One is the fear of systematic exclusion with no recourse, discretion, or alternatives [Weber, 2019, Merton, 1940, Eubanks, 2019]. The other is a fear of haphazard, inconsistent, and arbitrary decisions [Lipsky, 1980, Pasquale, 2016]. In one case decision making is too rigid, in the other it is too loose.

These concerns extend to algorithmic ecosystems—networks of multiple institutions that each use algorithms for consequential decisions about individuals. In this context, the two fears map onto a growing academic discourse about algorithmic monoculture and model multiplicity. In a monoculture, the algorithmic ecosystem collapses to a single rule leading to homogeneous outcomes and reduced welfare [Creel and Hellman, 2021, Kleinberg and Raghavan, 2021, Bommasani et al., 2022]. In contrast, model multiplicity proliferates the algorithmic ecosystem with inconsistent outcomes [Breiman, 2001, Marx et al., 2020]. Both narratives issue compelling warnings of what society might face, but there is apparent tension between the two. One problem, if severe enough, is an antidote to the other [Black et al., 2022, Ganesh et al., 2025, Gur-Arieh and Lee, 2025].

In this work, we systematically study the degree to which either narrative fits the empirical landscape. In doing so, we anticipate a near future where decision makers prompt large language models for risk assessments and predictions about individuals. Traditionally, institutions have procured special-

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

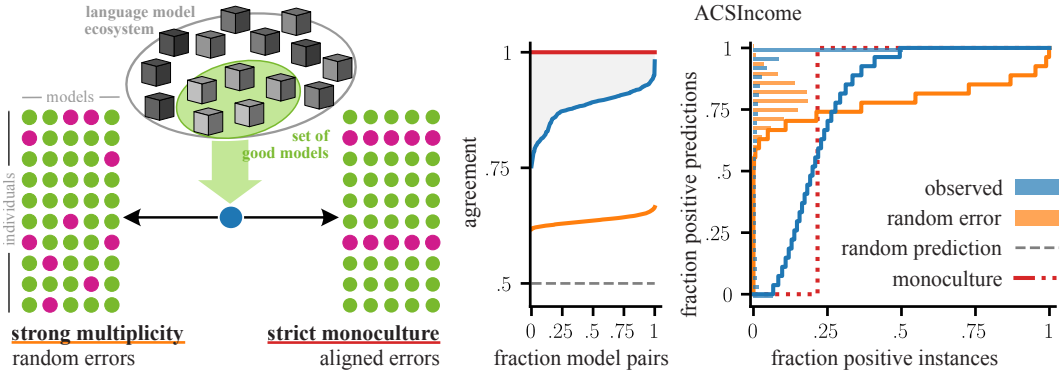

Figure 1: **Left.** Schematic of our setup. We empirically measure the degree of multiplicity and monoculture present in recent large language models. We zero-shot models and select those with accuracy within $\epsilon = 0.05$ of the best. **Middle.** Agreement curve for ACSIncome: x% of model pairs agree on at most y% of the positive instances. Observed agreement (blue) is higher than under random errors (orange) and random predictions (dashed gray), but well below strict monoculture (red). **Right.** Recourse curve for ACSIncome: x% of positive instances are accepted by no more than y% of models; e.g., 25% of instances are accepted by no more than 66% of the models. Under strict monoculture (red, dotted) all models collapse to one, so individuals are either accepted or rejected by all; the accepted fraction equals to the model's true positive rate (here: mean TPR). The y-axis bar plot shows the distribution of recourse levels.

purpose predictive models for consequential decisions—primarily, regression models trained on task-specific tabular records. Pre-trained language models promise a tempting alternative: these models work with any data encoding and flexibly solve numerous tasks, thus mitigating inefficiencies in the old machine learning pipeline.

How the use of large language models influences the situation, however, is neither obvious nor has it been studied systematically to date. Large language models share overlapping training data—much of the accessible internet—and there are relatively few model providers due to excessive training costs. These factors seem to promote monoculture. On the other hand, much research on evaluation points to the inconsistencies and instabilities of answers generated from large language models.

We broadly evaluate 50 large language models across seven prediction tasks and under different prompting strategies. What emerges is a rich empirical landscape that defies either of the two narratives. Language models show significant variation in their predictions, largely contradicting the idea of a monoculture. When rejected by one model, individuals consistently find recourse in other models. At the same time, model similarity is real and model errors are far from independent. To the extent that model multiplicity occurs, however, its presence isn't strong enough to fully mitigate the problem of systematic exclusion for a non-negligible fraction of the population. Our findings point at the unglamorous middle ground between monoculture and multiplicity as the most likely place for algorithmic ecosystems to end up in.

1.1 Our contributions

We empirically measure the degree of multiplicity and monoculture present in recent large language models under zero- and few-shot prompting. We evaluate 50 large language models, both open source models ranging in size from 1B to 141B parameters and state-of-the-art commercial model, across 7 standard prediction tasks, six based on U.S. Census Data and one from the medical domain.

Models share inductive biases. We evaluate the extent to which different models agree in their predictions under identical zero-shot prompts, focusing on the *Rashomon set*—models achieving accuracy near the best (Figure 1, left panel). We consider two extremes: Under strong multiplicity, all models err randomly. For instance, two models with 80% accuracy would agree in 68% of their predictions if errors were randomly located. Higher agreement rates indicate model similarity. Under strict monoculture, agreement rates approach 100%. Figure 1 (middle panel) shows that the

empirically observed agreement lies solidly between the two extremes, far from monoculture, but also far from strong multiplicity.

Individuals generally have high recourse. We define the *recourse level* of an individual with respect to a model set as the fraction of models that accept that individual, capturing the number of ways they have in finding a positive prediction within the model ecosystem. We focus on positive instances, i.e., those individuals that should be accepted. Figure 1 (right panel) shows the *recourse curve*, corresponding to the recourse level for each population quantile. For example, 7% are wrongly rejected by all models, thus having no recourse, while 51% are correctly accepted by all models. Recourse is generally high with 79% of positive instances accepted by at least half the models. The observed recourse curve is far from what it would look like under either monoculture or multiplicity.

Prompt variation provides recourse even in a single model. Previously, we gave different models identical prompts. In reality, it is extremely unlikely that two institutions would prompt models in exactly the same way. Different institutions likely use different data encodings or follow different practices. We consider only minor prompt variations including changes in feature order, granularity, and formatting. Even these minor variations provide recourse levels that rule out monoculture.

All findings are robust to few-shot prompting. We repeat our analysis under 10-shot prompting. Few-shot prompting generally improves accuracy. However, the results about model agreement and individual recourse are consistent and suggest the same conclusions.

Summary. Our evaluation suggests that neither monoculture nor multiplicity tightly fits the empirical observations. Recourse levels are far from perfect, but sufficiently high to rule out monoculture. This is true even though our setup favors monoculture: we evaluate similar language models under identical use or minor prompt variations. Model multiplicity is a real phenomenon, but models nevertheless share strong inductive biases. Their predictions exhibit high agreement rates, well above those we'd see under randomly located errors.

1.2 Related Work

The study of model multiplicity and the notion of the Rashomon set go back to Breiman [2001]. In recent years, this phenomenon has again gained traction [Black et al., 2022, Creel and Hellman, 2021, D'Amour et al., 2020, Jain et al., 2024]. Much prior research examined multiplicity in traditional models such as of regression, linear classification, tree-based models, and small neural networks in controlled settings [Marx et al., 2020, Bommasani et al., 2022, Semenova et al., 2022, Castillo et al., 2008, Hsu and Calmon, 2022, Hsu et al., 2024]. By systematically varying either training data or hyperparameters, these studies examine the effects that different training practices have on the Rashomon set. To address the concerns of arbitrariness and justifiability, other lines of work either mitigate multiplicity by aggregating models from the Rashomon set [Black et al., 2021, 2022, Cooper et al., 2024, Roth et al., 2023, Behzad et al., 2025] or leverage it to incorporate additional selection criteria [Semenova et al., 2022, Coston et al., 2021, Black et al., 2024].

Creel and Hellman [2021] argue that monopolistic power of few companies may lead to *algorithmic leviathans*, resulting in arbitrary decisions deployed at scale. They propose to intentionally introduce variance to the model, that is, to counteract monoculture by intentionally introducing multiplicity. Kleinberg and Raghavan [2021] coin the term *algorithmic monoculture* and provide a theoretical model where it can occur. Other theoretical models explore the relationship of competition and monoculture [Jagadeesan et al., 2023, Raghavan, 2024]. Black et al. [2022] argue that multiplicity can be a bulwark against monoculture, providing opportunities to choose the model that best promotes fairness or welfare among equivalent ones. Ganesh et al. [2025] give a comprehensive discussion of the moral concern of arbitrariness and its relationship to multiplicity and monoculture.

Closely related to monoculture, Bommasani et al. [2022] formalize *outcome homogenization* as the amplification of systemic denial relative to what would be expected under independent model predictions. The authors argue that we may see growing homogenization due to the rise of foundation models and conduct experiments on a vision-text model (CLIP) and an early language model (RoBERTa-base). Toups et al. [2024] analyze systemic failure and outcome homogenization for three commercial APIs used for emotion recognition and sentiment analysis.

There is a broader debate about the diversity of content generated by large language models that's beyond the scope of our work, as we focus on predictions and risk assessment.

2 Measuring monoculture and multiplicity

Preliminaries. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be a dataset of N i.i.d. samples from a joint distribution $P(\mathcal{X}, \mathcal{Y})$, where each $x_i \in \mathcal{X}$ is a feature vector and $y_i \in \{0, 1\}$ is the corresponding label. A binary predictor is a function $h\colon \mathcal{X} \to \{0, 1\}$. As algorithmic monoculture and model multiplicity arise in algorithmic ecosystems, we consider a model set $\mathcal{M} = \{h_1, h_2, \ldots, h_M\}$, which consists of M binary predictors.

Prior work on multiplicity focuses on a set of good models that achieve accuracy close to that of a reference model h_0, typically the best model found by empirical risk minimization. In reference to the Rashomon effect [Breiman, 2001], this model set is called Rashomon set.

Definition 1 (Rashomon set, ϵ-level set)**.** *Given a baseline model h_0, performance metric θ, and error tolerance ϵ, the* Rashomon set *is given by $\mathcal{R}_\epsilon(h_0) = \{h \in \mathcal{H} : \theta(h) \geq \theta(h_0) - \epsilon\}$.*

Throughout the paper we use accuracy as performance metric θ and refer to the parameter ϵ as *accuracy deficit*. We call the set of available models that deviate in performance at most by ϵ from the best model found the *empirical Rashomon set* to indicate that we don't have the full Rashomon set $\mathcal{R}_\epsilon$. Note that characterizing the full Rashomon set becomes computationally infeasible for large hypothesis classes $\mathcal{H}$, rendering efficient search a core issue of predictive multiplicity [Marx et al., 2020, Hsu and Calmon, 2022, Hsu et al., 2023].

Definition 2. *Given two classifiers h, h', their* pairwise agreement *over dataset $\mathcal{D}$ is given by the fraction of individuals their predictions agree on*

$$\text{agree}(h, h') = \tfrac{1}{N} \textstyle\sum_{i=1}^N \mathbb{1}[h(x_i) = h'(x_i)] \tag{1}$$

Pairwise agreement quantifies model similarity based on their predictions. Note that models can agree on both correct and incorrect predictions. We take strict monoculture to correspond to agreement 1[2]. Strong multiplicity corresponds to the agreement rate we'd expect if errors were randomly located. For two models h, h' with accuracies a and a', respectively, this corresponds to the expression

$$\text{agree}_{rand}(h, h') = aa' + (1 - a)(1 - a').$$

In this work, we focus on predictive multiplicity—situations where competing models assign conflicting predictions to the same individual. For our analysis, we also adopt two commonly used measures, ambiguity and discrepancy. Note that several other forms of multiplicity have been studied, see Ganesh et al. [2025] for a comprehensive overview.

Definition 3 (Ambiguity, [Marx et al., 2020])**.** *The* ambiguity *of a prediction problem over the model set $\mathcal{M}$ is the proportion of points x_i in a dataset that can be assigned a conflicting prediction by a competing classifier h compared to a baseline classifier $h_0 \in \mathcal{M}$:*

$$\alpha(h_0) = \tfrac{1}{N} \textstyle\sum_{i=1}^N \max_{h \in \mathcal{M}} \mathbb{1}[h(x_i) \neq h_0(x_i)] \tag{2}$$

Ambiguity captures the extent to which individuals are affected by the choice of model, because they would receive a conflicting prediction from another model in the Rashomon set. When switching the baseline model for a competing model, the conflicting predictions must all be realized by that particular model. Discrepancy captures the maximum change that can be realized by any other model in the Rashomon set relative to the baseline classifier.

Definition 4 (Discrepancy, [Marx et al., 2020])**.** *The* discrepancy *of a prediction problem over the model set $\mathcal{M}$ is the maximum proportion of conflicting predictions between the baseline classifier h_0 and any competing classifier h:*

$$\delta(h_0) = \max_{h \in \mathcal{M}} \tfrac{1}{N} \textstyle\sum_{i=1}^N \mathbb{1}[h(x_i) \neq h_0(x_i)] \tag{3}$$

Note that ambiguity and discrepancy are monotonic, meaning they can only increase as additional models are considered. Given that our empirical Rashomon set may omit some models achieving accuracy similar to the best model, monotonicity ensures that our measures provide lower bounds on those computed from the full Rashomon set.

[2]Accuracy differences within the Rashomon set can oppose strict monoculture: two models with accuracies $a \neq a'$ can agree on at most $1 - |a - a'|$ of the total population. This bound, however, applies only at the population level. Our analysis focuses on positive instances. Thus, the bound does not directly apply without additional assumptions about conditional agreement. If the positive class constitutes at least an ϵ-fraction of the data, perfect agreement within this subset remains possible, even if the models' overall accuracies differ slightly.

Recourse with respect to a model set. The concept of monoculture implies homogenization of predictions, wherein, in the extreme case, all decision-makers rely on a single model. To measure its impact on individuals, Bommasani et al. [2022] consider the fraction of individuals receiving a negative prediction from all models. We generalize this idea by introducing the notion of *recourse with respect to a model set*. It captures the ability of an individual to obtain a favorable outcome by switching to a different model within the model set.

Definition 5 (Recourse with respect to a model set)**.** *For an instance $x \in \mathcal{X}$, we define* recourse level *with respect to a model set $\mathcal{M}$ as* $\mathrm{rec}(x) = \frac{1}{|\mathcal{M}|} \sum_{h \in \mathcal{M}} \mathbb{1}[h(x) = y^*]$*, where y^* denotes the favorable outcome.*

Note that this notion of recourse of an individual is a property of a model set. In contrast to the single-model notion of *algorithmic recourse* [Ustun et al., 2019], we consider favorable predictions obtained by changing the model, not the individual features. For our analysis, we distinguish:

- **Full recourse**: $rec(x) = 1$
- **Substantial recourse**: $rec(x) > 0.5$, the majority of models yield a favorable outcome.
- **Limited recourse**: $rec(x) \leq 0.5$, the majority of models yield a unfavorable outcome.
- **No recourse**: $rec(x) = 0$, the individual faces systematic exclusion.

The fraction of individuals having no recourse corresponds to the observed rate of systemic failure defined by Bommasani et al. [2022]. They further define the notion of *outcome homogenization* as the observed ratio of individuals with no recourse relative to the fraction of individuals that would have no recourse, if model errors were random.

The notion of a favorable outcome depends on the prediction task and context. It may correspond to predicting 1 (e.g. loan approval), 0 (e.g. fraud detection), or to a correct prediction irrespective of the label. In this work, we focus on the recourse level of *positive instances* – individuals who, according to their ground-truth label, should be granted an opportunity. We treat $y^* = 1$ as the favorable outcome, allowing us to analyze cases where such opportunities are denied. Under this framing, the average recourse level among positive instances is a natural generalization of the *true positive rate* from a single model to a model family.

We generalize the random error baseline used by Bommasani et al. [2022] to varying recourse levels. Recall, strict multiplicity corresponds to the case where errors are randomly located. Assuming that model h_m has true positive rate p_m, the fraction of models accepting a fixed positive instance x follows a Poisson-Binomial distribution so that

$$\mathbb{E}[\mathrm{rec}(x)] = \frac{1}{M} \sum_{\pi \sim \Pi_k} \prod_{m=1}^{M} p_m^{\pi^{(m)}} (1 - p_m)^{1 - \pi^{(m)}}, \tag{4}$$

where Π_k describes the set of distinct permutations of k successes and $M - k$ failures. Further let $\pi^{(m)} \in \{0, 1\}$ indicate whether for permutation π model m provides a success (positive prediction) or a failure (negative prediction). Note that the success probability in the baseline should reflect the outcome deemed favorable. In our case, since we are focusing on positive instances achieving positive outcomes, p_m corresponds to the true positive rate. More generally, when considering correct classifications, the appropriate metric would be accuracy. Conversely, if a negative outcome (i.e., class 0) is considered favorable, the corresponding success probability is the true negative rate.

3 Experimental Design

Working from the premise that decision makers will increasingly use language models for decision making, we evaluate the consistency of model predictions across a diverse set of language models on multiple binary classification tasks. In each task, models are prompted with a natural-text representation of the features x using a standard multiple-choice format. We then compare similarity of model outputs and their impact on individuals varying either the models or the prompt.

Prompting language models. We evaluate 50 language models, including both open-source models—ranging from 1B to 141B parameters—and state-of-the-art commercial models. A complete

Table 1: Recourse levels and multiplicity measures across tasks using identical prompts.

task	$\lvert\mathcal{R}_\epsilon\rvert$	no recourse	substantial recourse	full recourse	ambiguity	discrepancy
ACSIncome	27	0.07	0.80	0.51	0.43	0.19
ACSEmployment	8	0.06	0.82	0.64	0.30	0.15
BRFSS Blood Pressure	23	0.03	0.59	0.14	0.83	0.38
SIPP	16	0.05	0.77	0.48	0.47	0.18
ACSTravelTime	12	0.04	0.72	0.24	0.72	0.35
ACSPublicCoverage	21	0.14	0.38	0.00	0.86	0.47
ACSMobility	5	0.68	0.00	0.00	0.32	0.20

list is provided in Appendix A.1. For open-source models, we include both base pretrained variants and their instruction-tuned counterparts if available. We use the `folktexts` package [Cruz et al., 2024] to zero- and few-shot the language models in a standardized multiple-choice format and to extract predicted risk scores. Since language models are widely miscalibrated on non-realizable tasks, we adopt the approach from Cruz et al. [2024] to calibrate the model predictions. For each model and task, we fit a decision threshold t on $n = 2000$ samples from a validation set to maximize balanced accuracy. The threshold is then applied to turn the risk scores into class predictions.

Prediction tasks. We evaluate model predictions on seven binary classification tasks derived from three data sources. Five tasks are based on the American Community Survey (ACS) Public Use Microdata Sample (PUMS), a high-quality dataset from the U.S. Census Bureau [Flood et al., 2018]. Ding et al. [2022] construct several predictions tasks on this dataset, for which `folktexts` [Cruz et al., 2024] provide natural-text mappings for prompting. We use ACSIncome (individual's income is above $\$50,000$), ACSEmployment (individual is employed), ACSPublicCoverage (individual is covered by public health insurance) and ACSTravelTime (commute time to work is greater than 20 minutes) and ACSMobility (individual moved in the last year). To complement these, we extend `folktexts` by two additional tasks. BRFSS Blood Pressure is a health-related prediction task introduced by Gardner et al. [2024] on large-scale surveys from the Behavioral Risk Factors Surveillance System [BRFSS, Centers for Disease Control and Prevention (CDC), 2021]. The task is to predict if an individual has been diagnosed with hypertension. SIPP is defined on the longitudinal Survey of Income and Program Participation [SIPP, U.S. Census Bureau, 2014]. Here, the goal is to predict whether a person's income is significantly above the Official Poverty Measure (OPM). See Appendix A.2 for a detailed description of all tasks. We randomly sample 10% of each dataset for evaluation, yielding test sets ranging in size from approximately 60,000 to 320,000 instances for most tasks. The SIPP task has a smaller test set of almost 4,000 instances.

Our analysis focuses on the impact on positive instances, that is individuals with a positive label $y = 1$. We refer the reader to Appendix D.4 for results for negatives instances. In the next two sections we examine similarity of model predictions for identical prompts across different models (Section 4) and, starting from a single model, explore the impact of prompt variations on monoculture and multiplicity (Section 5).

4 Models disagree on identical prompts

A central concern in increasingly interconnected algorithmic ecosystems is the potential convergence toward homogeneity—where different models yield highly similar outputs that, if flawed, could systematically limit individuals' opportunities. To explore the degree of homogeneity and its consequences, we evaluate our base set of 50 language models using identical zero-shot prompts. Aggregate performance varies considerably (Appendix B). We thus restrict our analysis to the Rashomon set—models whose accuracy falls within $\epsilon = 0.05$ of the best-performing model. Across tasks, the empirical Rashomon set comprises models from at least four providers, encompassing different sizes as well as both base and instruction-tuned variants.

Model predictions are highly similar. As shown in Figure 1 (left), pairwise comparisons on the ACSIncome task reveal that all model pairs exhibit substantially higher agreement (blue line) than would be expected if errors were randomly distributed (orange line). Specifically, model pairs

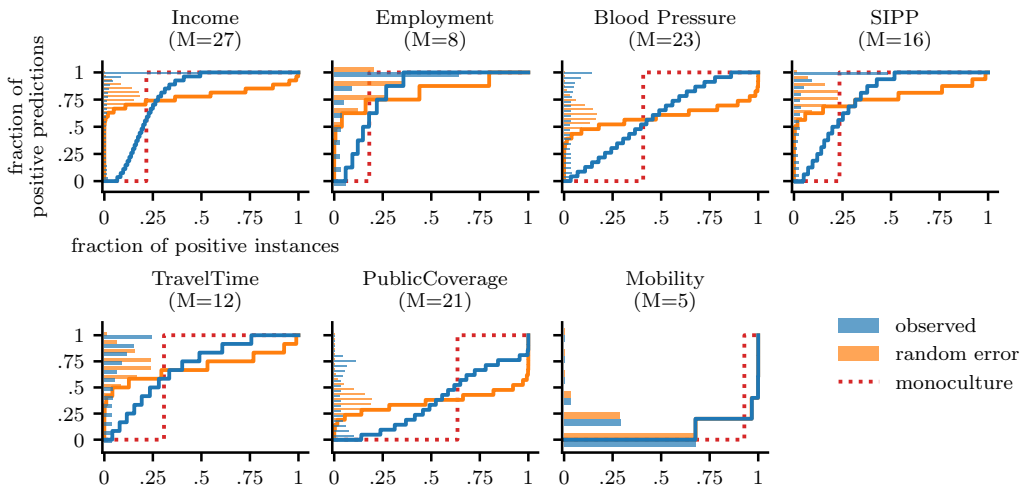

Figure 2: Recourse curves across tasks: x% of positive instances are accepted by no more than y% of models. We zero-shot models and select those achieving accuracy within $\epsilon = 0.05$ from the best. For example, on ACSIncome we observe (blue) 20% of positive instances being accepted by at most 50% of the models. Under random errors (orange) this would rarely happen. Under strict monoculture (red, dotted) individuals only experience no or full recourse. Here, the mean TPR is used for illustration. The bar plot shows the distribution of recourse levels.

agree on at least 75.34% of positive instances, with a mean agreement of 88.37%. In contrast, the random-error baseline yields agreement rates between 61.69% and 66.49% (mean 63.66%). These results demonstrate that models share considerable inductive biases. However, agreement remains below perfect consensus, which would be observed in a strict monoculture. This suggests that models are highly similar, but not perfectly aligned in their predictions. This pattern holds when disaggregating by model provider, model variant, or demographic groups (Appendix C). Extending this analysis across tasks (Figure 19, top panel) shows a consistent pattern of high predictive similarity, with empirically observed agreement rates consistently falling between the two extremes of strong multiplicity and strict monoculture.

Individuals generally have high recourse. Pairwise agreement rates – unless exactly 0 or 1 – do not directly capture the individual-level impact of predictions within a model ecosystem. In particular, they do not reveal the extent to which entire model sets concur on an individual's outcome. To address this, we examine recourse: the fraction of models that accept an individual, reflecting the individual's ability to obtain a favorable outcome by switching to another model. The recourse curve in Figure 1 (right) shows these values across population quantiles for ACSIncome. We find that 79% of positive instances are accepted by at least half of the models. The median individual is accepted by 96% of the models, compared to approximately 78% under the random-error baseline. These results suggest that recourse levels are generally high. Notably, 7% of positive instances are rejected by all models, meaning they receive no recourse. Conversely, 51% are correctly accepted by all models. Relative to a baseline of random errors, we observe a higher likelihood of both extreme outcomes, no recourse and full recourse, reflecting the influence of shared inductive biases. At the same time, the distribution of recourse levels differs considerably from a strict monoculture, wherein the recourse curve would collapse into a step function determined by the false negative rate of the sole remaining model. Thus, despite substantial predictive similarity, individuals largely retain the opportunity to obtain favorable outcomes by switching models.

Extending this analysis across tasks (Figure 2, top panel, and Table 1) reveals a generally consistent pattern of high recourse levels, with a small fraction of individuals experiencing no recourse. Across tasks, the distribution of recourse levels largely falls between the extremes of strong multiplicity and strict monoculture. Two tasks, ACSMobility and ACSPublicCoverage, deviate from this pattern, exhibiting higher rates of no recourse and smaller fractions of individuals with substantial recourse. Both tasks are characterized by low predictive signal and high class imbalance, which also affect the

Figure 3: Severity of monoculture and predictive multiplicity as a function of accuracy deficit ϵ from the best model. Each column corresponds to one task. Predictions are obtained via 0-shot prompting. **Top.** The fraction of positive instances that experience no recourse (blue) is consistently higher than what would be expected under random errors (orange). While stable across different Rashomon sets, the gap between the observed and expected fraction of no recourse decreases with increasing ϵ. **Middle.** Discrepancy increases with ϵ, potentially affording opportunities for recourse for some individuals. Due to monotonicity, their values are likely higher for the full Rashomon set. **Bottom.** Number of models in the empirical Rashomon set.

Rashomon set. See Appendix D.1 for a detailed discussion. We further assess recourse levels for different demographic groups defined by sex, race, and age, to examine whether predictive similarity affects these groups differently. While levels of agreement and recourse are comparable between female and male individuals, we observe notable differences in multiple tasks when disaggregating by race or age. These observations provide preliminary evidence that certain groups may face disproportionate barriers to finding recourse. Differences between tasks suggest that disparities may also be task-dependent. A more comprehensive understanding of these dynamics remains an important direction for future work. Further details and analysis are provided in Appendix C.2.

Multiplicity is high, but we might still underestimate it. As our analysis is limited to available pre-trained models, we cannot influence training choices or exhaustively explore the parameter space. Nonetheless, our empirical Rashomon set likely includes models representative of those actually deployed by decision makers. Among those models, we find substantial discrepancies: on ACSIncome, 19% of positive instances receive a different classification when switching to another model (Table 1); across tasks, at least 15% do. Both ambiguity and discrepancy are monotonic with respect to the size of the Rashomon set [Marx et al., 2020]. Therefore, an exhaustive search over the parameter space would likely yield even higher values for both measures.

Impact of accuracy deficit ϵ. The accuracy deficit ϵ specifies the maximum performance gap among models in the Rashomon set. To assess its impact on our findings, we compute discrepancy and the prevalence of no recourse across varying values of ϵ (Figure 3). Even small increases in ϵ rapidly expand the empirical Rashomon set and observed discrepancy δ_ϵ, indicating that many models achieve comparable accuracy while generating divergent individual-level predictions. This diversity creates potential avenues for recourse and reduces the likelihood that decision makers converge on a single model. Nevertheless, whenever the Rashomon set includes more than one model, a non-negligible fraction of individuals experiences no recourse—substantially higher than expected under random errors. Full recourse curves for varying values of ϵ are provided in Appendix D.2.

Consistent under few-shot prompting We repeat our analysis under 10-shot prompting, providing each model with 10 class-balanced examples. We find that aggregate performance increases slightly, resulting in overall larger empirical Rashomon sets. Observed recourse levels remain at a similar order of magnitude with slightly decreased discrepancy, indicating a mild trend towards monoculture. See Appendix E for detailed results.

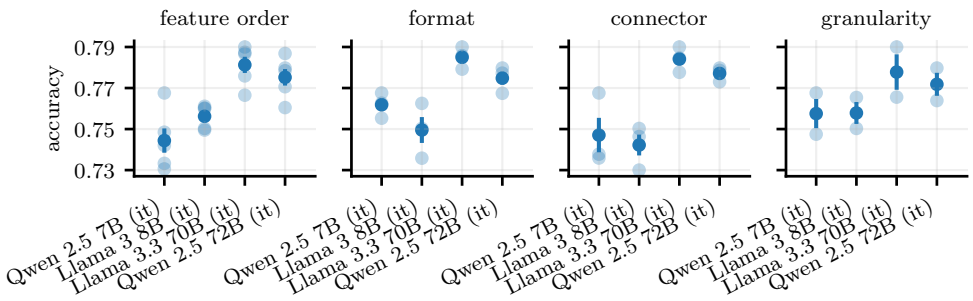

Figure 4: Minor prompt variations induce changes in accuracy of up to 3 percentage points, consistently across models on ACSIncome. Each subplot varies a single aspect of the prompt, keeping the others fixed to default. Light blue dots show accuracy for individual variations, dark blue dots indicate the mean accuracy with error bars.

5 Recourse through minor prompt variations

Under monoculture, algorithmic ecosystems tend toward homogeneity resembling a single model making all decisions. To simulate this in the context of LLMs used for consequential decision-making, we select the two best-performing models on ACSIncome – Llama 3.3 70B (it) and Qwen 2.5 72B (it) – along with smaller variants, Llama 3 8B (it) and Qwen 2.5 7B (it). Even after training, LLMs accept a wide range of prompt formats, styles or even modalities. This flexibility means that decision-makers may naturally vary in how they compose prompts. In line with prior work on prompt sensitivity of language models [Petroni et al., 2019, Shin et al., 2020, Sclar et al., 2023, Voronov et al., 2024], we find that minor prompt variations cause significant changes in model predictions.

Prompt variations. We vary four minor aspects of prompt construction: ***feature order***, testing five orders in which features of an individual are presented - the default order given by `folktexts`, its reverse, and three random permutations; ***format***, presenting features as *bullet* list, a *comma*-separated list or plain *text* in the form '<feature name> <connector> <feature value>'; ***connector***, choosing the symbol linking <feature name> and <feature value> among '*is*', '=' and ':'; and ***granularity***, using either the original feature mapping provided by `folktexts` or a lower-resolution version such as age groups instead of the exact age. Examples are provided in Appendix G.

We find that varying a single aspect of the prompt construction can shift accuracy up to 3 percentage points (Figure 4) and with it, induce predictive multiplicity. The effect on accuracy is consistent across models, with no clear trend by size or model family. While granularity changes alter the information content, we find their impact on accuracy to be comparable to that of other variations.

Prompt variations still induce substantial multiplicity. Since decision-makers may vary in more than one aspect of how they construct prompts, we evaluate all four models on the cross-product of these variations, resulting in $V = 90$ distinct prompting styles per model. To compare effects of prompt variations and model changes directly, we fix the number of prompts and models to be the same: we randomly sample M prompt styles ($M = 27$ for ACSIncome) and evaluate agreement and recourse across 100 independent repetitions (Figure 5 top panel). Pairwise agreement across models increases under prompt variations (blue), though it remains similar to the agreement observed when varying the model under identical prompting (gray line), suggesting that substantial disagreement persists in both settings. Comparing recourse curves (bottom panel), we find that prompt variation leads to a higher fraction of positive instances with full recourse, while the rate of instances with no recourse remains similar. This suggests a mild trend toward monoculture; nevertheless, prompt variations alone still enable recourse for a considerable fraction of individuals, with the majority experiencing substantial recourse.

Prompt variations and few-shot prompting Few-shot prompting introduces additional sources of variation, such as the choice and ordering of examples. Consistent with prior work [Zhao et al., 2021, Lu et al., 2022, Gao et al., 2021, Schick and Schütze, 2021], we find that these factors impact model performance and add variability in model predictions, impacting recourse we observe (Appendix F).

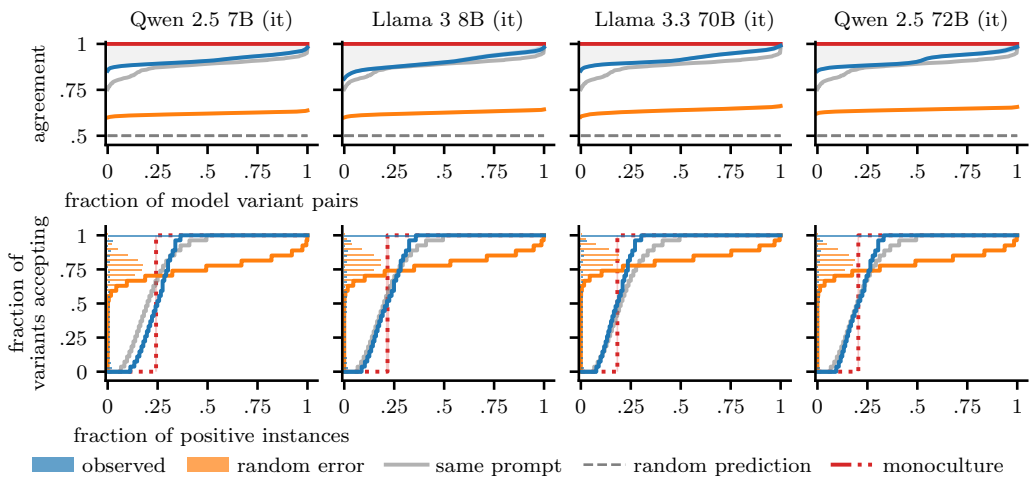

Figure 5: Agreement and recourse across prompt variations on ACSIncome. We zero-shot models with varying prompting styles, subsampling $V = 27$ prompt variations per model to match the number of different models in Figure 1. **Top.** Agreement curve: x% of prompt variation pairs agree on up to y% of the positive instances. Observed agreement rates (blue) surpass those under random errors (orange), but remain similar to agreement rates observed with identical prompting across different models (gray, solid). **Bottom.** Recourse curve: x% of the positive instances are accepted by at most y% of the prompt variations. For example, 25% of the positive instances are accepted by at most 50% of the variations. Observed recourse (blue), random errors (orange). Under strict monoculture (red, dotted) individual only experience no or full recourse. Here, the mean TPR is used for illustration. Bar plot on the y-axis shows density function of recourse level in the population.

6 Discussion

Our evaluation suggests that neither monoculture nor multiplicity fits the empirical landscape. Every way we look at it, the observations fall strictly between the two extremes of strict monoculture and strong multiplicity. Still, predictive similarity among the evaluated models is high: across tasks, a notable fraction of individuals face no recourse, yet disagreement between models provides opportunities for the majority to find recourse. Disaggregating by demographics reveals that these effects are unevenly distributed, with certain groups disproportionately affected. These findings underscore the need for a deeper understanding of the system-levels dynamics and the individual harms arising from them. We believe that conducting such system-level analyses can serve as an important initial step in this direction. Some might contend that our experimental setup is biased toward monoculture. After all, we evaluate all models on exactly the same dataset. In reality, two different institutions almost certainly collect different datasets. The records that two different institutions have about the same individual are likely quite different in most cases. This introduces an additional source of variation that we don't capture. As a result, our empirical findings likely understate how far reality is from monoculture. We study monoculture in the sense of systematic exclusion from opportunity in the context of consequential decision making using predictive risk assessment. We do not address broader sociological questions about how the use of language models might homogenize culture and expression.

In our study, we work from the assumption that increasingly decision makers will prompt language models for consequential decisions. Admittedly, this is a look into the future rather than the present. But it's a highly plausible near future given current trends of adoption of language models. Language models must strike institutions like a perfect fit for the mundane bureaucratic processes in which consequential decisions take place. These models can sift through volumes of messily encoded data and come up with some answer to any question. It's urgent, then, to ask: What harms might individuals face in a language model ecosystem that governs consequential decisions? Our work points to a troubling conclusion: the harms individuals face resist capture by either the lens of monoculture or that of multiplicity.

Acknowledgments and Disclosure of Funding

We thank Yatong Chen, Ricardo Dominguez-Olmedo, Karolin Frohnapfel, Mina Remeli, and Mara Seyfert for invaluable feedback on earlier versions of this paper. We also thank André Cruz and Vivian Nastl for their continued support with implementation and data-related aspects. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Mila Gorecki.

References

01.AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open Foundation Models by 01.AI, January 2025.

Barry Becker and Ronny Kohavi. Adult Dataset, 1996.

Tina Behzad, Sílvia Casacuberta, Emily Ruth Diana, and Alexander Williams Tolbert. Reconciling Predictive Multiplicity in Practice - ICML Workshop Version, January 2025.

Emily Black, Klas Leino, and Matt Fredrikson. Selective Ensembles for Consistent Predictions. In *International Conference on Learning Representations*, October 2021.

Emily Black, Manish Raghavan, and Solon Barocas. Model Multiplicity: Opportunities, Concerns, and Solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 850–863, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533149.

Emily Black, Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. The Legal Duty to Search for Less Discriminatory Algorithms, June 2024.

Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?, November 2022.

Leo Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–215, 2001. ISSN 0883-4237.

Enrique Castillo, Roberto Mínguez, Carmen Castillo, and Antonio S. Cofiño. Dealing with the multiplicity of solutions of the $\ell 1$ and $\ell\infty$ regression models. *European Journal of Operational Research*, 188(2):460–484, July 2008. ISSN 0377-2217. doi: 10.1016/j.ejor.2007.04.020.

Centers for Disease Control and Prevention (CDC). BRFSS survey data (2015, 2017, 2019, 2021), 2021.

A. Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification, March 2024.

Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing Fairness Over the Set of Good Models Under Selective Labels. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2144–2155. PMLR, July 2021.

Kathleen Creel and Deborah Hellman. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 816, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445942.

André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores, July 2024.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification Presents Challenges for Credibility in Modern Machine Learning, November 2020.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New Datasets for Fair Machine Learning. *Advances in Neural Information Processing Systems*, 34, January 2022.

Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. Picador St. Martin's Press, New York, first picador edition edition, 2019. ISBN 978-1-250-07431-7 978-1-250-21578-9.

Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, and J Robert Warren. *Integrated Public Use Microdata Series, Current Population Survey: Version 6.0*. Minneapolis, MN: IPUMS, 2018.

Prakhar Ganesh, Afaf Taik, and Golnoosh Farnadi. The Curious Case of Arbitrariness in Machine Learning, January 2025.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making Pre-trained Language Models Better Few-shot Learners. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.295.

Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking Distribution Shift in Tabular Data with TableShift, February 2024.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the Science of Language Models, June 2024.

Shira Gur-Arieh and Christina Lee. Consistently Arbitrary or Arbitrarily Consistent: Navigating the Tensions Between Homogenization and Multiplicity in Algorithmic Decision-Making. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 3336–3349, New York, NY, USA, June 2025. Association for Computing Machinery. ISBN 979-8-4007-1482-5. doi: 10.1145/3715275.3732215.

Moritz Hardt and Michael P. Kim. Backward baselines: Is your model predicting the past?, June 2022.

Hsiang Hsu and Flavio du Pin Calmon. Rashomon Capacity: A Metric for Predictive Multiplicity in Classification. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*. arXiv, October 2022.

Hsiang Hsu, Guihong Li, Shaohan Hu, and Chun-Fu Chen. Dropout-Based Rashomon Set Exploration for Efficient Predictive Multiplicity Estimation. In *The Twelfth International Conference on Learning Representations*, October 2023.

Hsiang Hsu, Ivan Brugere, Shubham Sharma, Freddy Lecue, and Chun-Fu Chen. RashomonGB: Analyzing the Rashomon Effect and Mitigating Predictive Multiplicity in Gradient Boosting. 2024.

Meena Jagadeesan, Michael Jordan, Jacob Steinhardt, and Nika Haghtalab. Improved Bayes Risk Can Yield Reduced Social Welfare Under Competition. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023.

Shomik Jain, Vinith Suriyakumar, Kathleen Creel, and Ashia Wilson. Algorithmic Pluralism: A Structural Approach To Equal Opportunity. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 197–206, June 2024. doi: 10.1145/3630106.3658899.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of Experts, January 2024.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, June 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2018340118.

Michael Lipsky. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. Publications of Russell Sage Foundation. Russell Sage Foundation, New York, 1980. ISBN 978-0-87154-544-2 978-1-61044-663-1.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive Multiplicity in Classification. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6765–6774. PMLR, November 2020.

Robert K. Merton. Bureaucratic structure and personality. In *Social Theory and Social Structure*, pages 195–206. Free Press, 1940.

MetaAI. Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/, 2024.

Mistral AI. Mistral small 3, 2025.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious, January 2025.

OpenAI. GPT-3.5 turbo, 2024.

OpenAI. Introducing GPT-4.1 in the API, April 2025.

Frank Pasquale. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, Cambridge, Massachusetts London, England, first harvard university press paperback edition edition, 2016. ISBN 978-0-674-97084-7 978-0-674-36827-9.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language Models as Knowledge Bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, January 2025.

Manish Raghavan. Competition and Diversity in Generative AI, December 2024.

Aaron Roth, Alexander Tolbert, and Scott Weinstein. Reconciling Individual Probability Forecasts. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 101–110, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 979-8-4007-0192-4. doi: 10.1145/3593013.3593980.

Timo Schick and Hinrich Schütze. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.20.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, October 2023.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the Existence of Simpler Machine Learning Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1827–1858, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533232.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.346.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open Models Based on Gemini Research and Technology, April 2024a.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving Open Language Models at a Practical Size, October 2024b.

Connor Toups, Rishi Bommasani, Kathleen A. Creel, Sarah H. Bana, Dan Jurafsky, and Percy Liang. Ecosystem-level analysis of deployed machine learning reveals homogeneous outcomes. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 51178–51201, Red Hook, NY, USA, May 2024. Curran Associates Inc.

U.S. Census Bureau. Survey of Income and Program Participation (SIPP), 2014.

U.S. Census Bureau. American Community Survey (ACS) Public Use Microdata Sample (PUMS), 2018.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 10–19, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287566.

Anton Voronov, Lena Wolf, and Max Ryabinin. Mind Your Format: Towards Consistent Evaluation of In-Context Learning Improvements. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6287–6310, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.375.

Jamelle Watson-Daniels, David C. Parkes, and Berk Ustun. Predictive Multiplicity in Probabilistic Classification, June 2023.

Max Weber. *Economy and Society: A New Translation*. Harvard University Press, 2019.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 Technical Report, September 2024.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706. PMLR, July 2021.

Technical Appendices and Supplementary Material

A Additional experimental details

A.1 Models

We evaluate 50 language models, including both open-source models—ranging from 1B to 141B parameters—and state-of-the-art commercial models. A complete list of the models is provided in Table 2. For open-source models, we include both base pretrained variants and their instruction-tuned counterparts if available. Instruction-tuned models may be appealing to decision makers as they are trained specific instruction-response pairs to enhance their ability to follow natural language instructions. As a result, they often show better alignment with human preferences on realizable tasks. However, on unrealizable tasks such as those studied in our work, they have also been observed to produce overconfident risk scores [Cruz et al., 2024]. Accordingly, to facilitate a more comprehensive characterization of model behavior, we also consider the corresponding base model variants. Base models can be naturally attractive to decision makers for several reasons: Unlike instruction-tuned models, they do not impose an instruction-response format, offering greater flexibility to be adapted in custom workflows or in deployment settings where a particular alignment may be undesirable. As detailed in Appendix B, the two model variants exhibit comparable performance, resulting in models from both variants being included in the Rashomon set of each task.

Table 2: Models evaluated for this work. Model size in billions of parameters is indicated by N. Model weights for open-source models were retrieved from the corresponding HuggingFace (HF) repositories.

Language Model	N	HF Repository	Citation
Gemma 2B	2.51	google/gemma-2b	Team et al. [2024a]
Gemma 2B (it)	2.51	google/gemma-1.1-2b-it	Team et al. [2024a]
Gemma 7B	8.54	google/gemma-7b	Team et al. [2024a]
Gemma 7B (it)	8.54	google/gemma-1.1-7b-it	Team et al. [2024a]
Gemma 2 9B	9.24	google/gemma-2-9b	Team et al. [2024b]
Gemma 2 9B (it)	9.24	google/gemma-2-9b-it	Team et al. [2024b]
Gemma 2 27B	27.2	google/gemma-2-27b	Team et al. [2024b]
Gemma 2 27B (it)	27.2	google/gemma-2-27b-it	Team et al. [2024b]
Llama 3 8B	8.03	meta-llama/Meta-Llama-3-8B	MetaAI [2024]
Llama 3 8B (it)	8.03	meta-llama/Meta-Llama-3-8B-Instruct	MetaAI [2024]
Llama 3 70B	70.6	meta-llama/Meta-Llama-3-70B	MetaAI [2024]
Llama 3 70B (it)	70.6	meta-llama/Meta-Llama-3-70B-Instruct	MetaAI [2024]
Llama 3.1 8B	8.03	meta-llama/Meta-Llama-3.1-8B	MetaAI [2024]
Llama 3.1 8B (it)	8.03	meta-llama/Meta-Llama-3.1-8B-Instruct	MetaAI [2024]
Llama 3.1 70B	70.6	meta-llama/Meta-Llama-3.1-70B	MetaAI [2024]
Llama 3.1 70B (it)	70.6	meta-llama/Meta-Llama-3.1-70B-Instruct	MetaAI [2024]
Llama 3.2 1B	1.24	meta-llama/Meta-Llama-3.2-1B	MetaAI [2024]
Llama 3.2 1B (it)	1.24	meta-llama/Meta-Llama-3.2-1B-Instruct	MetaAI [2024]
Llama 3.2 3B	3.21	meta-llama/Meta-Llama-3.2-3B	MetaAI [2024]
Llama 3.2 3B (it)	3.21	meta-llama/Meta-Llama-3.2-3B-Instruct	MetaAI [2024]
Llama 3.3 70B (it)	70.6	meta-llama/Meta-Llama-3.3-70B-Instruct	MetaAI [2024]
Mistral 7B	7.24	mistralai/Mistral-7B-v0.1	Jiang et al. [2023]
Mistral 7B (it)	7.24	mistralai/Mistral-7B-Instruct-v0.2	Jiang et al. [2023]
Mixtral 8x7B	46.7	mistralai/Mixtral-8x7B-v0.1	Jiang et al. [2024]
Mixtral 8x7B (it)	46.7	mistralai/Mixtral-8x7B-Instruct-v0.1	Jiang et al. [2024]
Mixtral 8x22B	141	mistralai/Mixtral-8x22B-v0.1	Jiang et al. [2024]
Mixtral 8x22B (it)	141	mistralai/Mixtral-8x22B-Instruct-v0.1	Jiang et al. [2024]
Mistral Small 24B	23.6	mistralai/Mistral-Small-24B-Base-2501	Mistral AI [2025]
Mistral Small 24B (it)	23.6	mistralai/Mistral-Small-24B-Instruct-2501	Mistral AI [2025]
Yi 6B	6.06	01-ai/Yi-6B	01.AI et al. [2025]
Yi 6B (chat)	6.06	01-ai/Yi-6B-Chat	01.AI et al. [2025]
Yi 34B	34.4	01-ai/Yi-34B	01.AI et al. [2025]
Yi 34B (chat)	34.4	01-ai/Yi-34B-Chat	01.AI et al. [2025]
Qwen 2 7B	7.62	Qwen/Qwen2-7B	Yang et al. [2024]

Continued on next page

Table 2 – *Continued from previous page*

Language Model	N	HF Repository	Citation
Qwen 2 7B (it)	7.62	Qwen/Qwen2-7B-Instruct	Yang et al. [2024]
Qwen 2 72B	72.7	Qwen/Qwen2-72B	Yang et al. [2024]
Qwen 2 72B (it)	72.7	Qwen/Qwen2-72B-Instruct	Yang et al. [2024]
Qwen 2.5 7B	7.62	Qwen/Qwen2.5-7B	Qwen et al. [2025]
Qwen 2.5 7B (it)	7.62	Qwen/Qwen2.5-7B-Instruct	Qwen et al. [2025]
Qwen 2.5 72B	72.7	Qwen/Qwen2.5-72B	Qwen et al. [2025]
Qwen 2.5 72B (it)	72.7	Qwen/Qwen2.5-72B-Instruct	Qwen et al. [2025]
OLMo 1B 0724	1.28	allenai/OLMo-1B-0724-hf	Groeneveld et al. [2024]
OLMo 1B	1.18	allenai/OLMo-1B-hf	Groeneveld et al. [2024]
OLMo 7B 0724	6.89	allenai/OLMo-7B-0724-hf	Groeneveld et al. [2024]
OLMo 7B	6.89	allenai/OLMo-7B-hf	Groeneveld et al. [2024]
OLMo 7B (it)	6.89	allenai/OLMo-7B-Instruct-hf	Groeneveld et al. [2024]
OLMo 2 7B	7.3	allenai/OLMo-2-1124-7B	OLMo et al. [2025]
OLMo 2 7B (it)	7.3	allenai/OLMo-2-1124-7B-Instruct	OLMo et al. [2025]
gpt-3.5-turbo-0125	*unknown*	-	OpenAI [2024]
gpt-4.1	*unknown*	-	OpenAI [2025]

A.2 Prediction Tasks

We evaluate model predictions on seven binary classification tasks derived from three data sources, Census Bureau's American Community Survey (ACS) Public Use Microdata Sample (PUMS) [U.S. Census Bureau, 2018], the Behavioral Risk Factors Surveillance System [BRFSS, Centers for Disease Control and Prevention (CDC), 2021] and the Survey on Income and Program Participation [SIPP, U.S. Census Bureau, 2014]. In this section we will provide more details about the data sources and each of the the prediction tasks used.

ACS prediction tasks. Five of the prediction tasks used for this work are based on ACS PUMS data, which is derived from US Census data and provides a rich, diverse, and high-quality representation of the US population. While a wide range of prediction tasks could be defined on this data source, we adopt five prediction tasks predefined in the popular `folktables` package [Ding et al., 2022]. These tasks span a broad range of prediction challenges – from high predictive signal to more difficult low-signal settings (such as ACS Mobility) – and collectively make up a diverse benchmark suite. While some features (e.g., age, race, and sex) appear across tasks due to their relevance for fairness analysis, their overall feature sets differ, and none is a strict subset of another. Further, each task is constructed on a distinct subpopulation (e.g., adults, employed individuals). As a result, individuals do not consistently appear across tasks: The maximum pairwise overlap corresponds to at most 10.1% relative to each task's size. No individual appears in all five tasks.

The `folktexts` package by Cruz et al. [2024] complements each task with a natural-text mapping for every feature, ready be used for language model prompting. To enable comparison with existing benchmarks and prior work on multiplicity, we use the task definitions including feature sets and population filters as provided. Following Ding et al. [2022] and Cruz et al. [2024] we analyze data from the 2018 1-year-horizon person-level survey, although any ACS survey year could be used. The following paragraphs detail each ACS prediction task. Table 3 provides a description and exemplary natural language encoding for all features.

ACSIncome The goal of the ACSIncome task is to predict whether a person's yearly income is above $50,000, given by the PINCP column. The ACS columns used as features are AGEP, COW, SCHL, MAR, OCCP, POBP, RELP, WKHP, SEX, and RAC1P. The column PINCP is binarized and used as target. The sub-population over which the task is conducted is employed US residents with age greater than 16 years. The ACSIncome prediction task was put-forth as the successor to the popular UCI Adult dataset [Becker and Kohavi, 1996], used extensively in the algorithmic fairness literature.

ACSEmployment The goal of the ACSEmployment is to predict whether an individual is employed, given by the ESR column. The ACS columns used as features are AGEP, SCHL, MAR, RELP, DIS, ESP, CIT, MIG, MIL, ANC, NATIVITY, DEAR, DEYE, DREM, SEX and RAC1P. The sub-population over which the task is conducted is US residents.

ACSTravelTime The goal of the ACSTravelTime task is to predict whether a person's commute time to work is greater than 20 minutes, given by the JWMNP column. The ACS columns used as features are: AGEP, SCHL, MAR, SEX, DIS, ESP, MIG, RELP, RAC1P, PUMA, ST, CIT, OCCP, JWTR, POWPUMA, and POVPIP. The sub-population over which the task is conducted is employed US residents with age greater than 16 years.

ACSPublicCoverage The goal of the ACSPublicCoverage task is to predict whether an individual is covered by public health insurance, given by the PUBCOV column. The ACS columns used as features are: AGEP, SCHL, MAR, SEX, DIS, ESP, CIT, MIG, MIL, ANC, NATIVITY, DEAR, DEYE, DREM, PINCP, ESR, ST, FER, and RAC1P. The sub-population over which the task is conducted is US residents with age below 65 years old, and with personal income below $30,000.

ACSMobility The goal of the ACSMobility task is to predict whether an individual has changed their home address in the last year, given by the MIG column. The ACS columns used as features are: AGEP, SCHL, MAR, SEX, DIS, ESP, CIT, MIL, ANC, NATIVITY, RELP, DEAR, DEYE, DREM, RAC1P, GCL, COW, ESR, WKHP, JWMNP, and PINCP. The sub-population over which the task is conducted is US residents with age between 18 and 35.

BRFSS prediction task. The `tableshift` package [Gardner et al., 2024] provides a unified API to 15 prediction tasks, including some of the above mentioned ACS prediction tasks. For this work, we complement the ACS prediction tasks by one additional health-related prediction task, BRFSS Blood Pressure. The data comes from the Behavioral Risk Factor Surveillance System [BRFSS, Centers for Disease Control and Prevention (CDC), 2021]. BRFSS is a US-wide system of telephone surveys that assess health-related risk behaviors, chronic health conditions, and the use of preventive services by US residents. BRFSS collects data in all 50 states as well as the District of Columbia and three US territories. Following the implementation in `tableshift`, we use the default year range (2015–2021) of biannual BRFSS survey data. Further, we use the dataset as preprocessed in the `tableshift` package, with the exception of feature normalization, which is omitted to preserve the raw data entries used for mapping to natural text. Following the codebook available for BRFFS, we define a natural-text mapping for every feature. Table 4 provides a description and exemplary natural language encoding for all features.

BRFSS Blood Pressure The goal of the BRFSS Blood Pressure task is to predict whether an individual has been told by a health professional that they have high blood pressure (hypertension). The features used include several risk factor like age, family history, other medical conditions, race, sex and social and economic factors. The BRFSS columns used as features are: BMI5CAT, AGEG5YR, FRUIT_ONCE_PER_DAY, VEG_ONCE_PER_DAY, DRNK_PER_WEEK, RFBING5, TOTINDA, SMOKE100, SMOKDAY2, CHCSCNCR, CHCOCNCR, DIABETES, POVERTY, EMPLOY1, IYEAR, STATE, MEDCOST, PRACE1 and SEX. The subpopulation over which the task is defined includes all US residents, with no additional filters.

SIPP prediction task. Following Hardt and Kim [2022], we define one task based on the Survey of Income and Program Participation (SIPP) data [U.S. Census Bureau, 2014]. SIPP is an important longitudinal survey administered by the US Census Bureau that provides information on the dynamics of income, employment, household composition, and government program participation. The survey is considered nationally representative and interviews individuals for several years, providing monthly data about changes in household and family composition and economic circumstances over time. As Hardt and Kim [2022], we consider Wave 1 and Wave 2 of the SIPP 2014 panel data for the prediction task and adopt their data cleaning and pre-processing steps. Following the codebook available for SIPP, we define a natural-text mapping for every feature. Table 5 provides a description and exemplary natural language encoding for all features.

SIPP The goal of the SIPP task is to predict whether an individual's income is well above the Official Poverty Measure (OPM), a cash-income-based measure of poverty. The target variable is

calculated using the OPM ratio from Wave 2 data. In total, a set of 50 features is constructed from one or multiple variables appearing in the Wave 1 raw data are used for the SIPP task. These include features of the individual related to

- demographics (AGE, GENDER, RACE, EDUCATION MARITAL_STATUS, CITIZENSHIP,
  FAMILY_SIZE_AVG, ORIGIN),
- income (INCOME, HOUSEHOLD_INC, RECEIVED_WORK_COMP, INCOME_FROM_ASSISTANCE, SAVINGS_INV_AMOUNT),
- health (MEDICARE_ASSISTANCE, MEDICAID_ASSISTANCE, HEALTHDISAB, DAYS_SICK, HOSPITAL_NIGHTS, PRESCRIPTION_MEDS, HEALTH_INSURANCE_PREMIUMS,
  HEALTH_OVER_THE_COUNTER_PRODUCTS_PAY,
  HEALTH_MEDICAL_CARE_PAY,
  HEALTH_HEARING, HEALTH_SEEING, HEALTH_COGNITIVE, HEALTH_AMBULATORY,
  HEALTH_SELF_CARE, HEALTH_ERRANDS_DIFFICULTY, HEALTH_CORE_DISABILITY,
  HEALTH_SUPPLEMENTAL_DISABILITY, VISIT_DOCTOR_NUM, VISIT_DENTIST_NUM),
- hardship (LIVING_QUARTERS_TYPE, LIVING_OWNERSHIP, FOOD_ASSISTANCE, WIC_ASSISTANCE, SNAP_ASSISTANCE),
- and details on program participation (TANF_ASSISTANCE ,TANF ,TRANSPORTATION_ASSISTANCE, _COMP, UNEMPLOYMENT_COMP_AMOUNT, SOCIAL_SEC_BENEFITS, VA_BENEFITS_AMOUNT, RETIREMENT_INCOME_AMOUNT, SURVIVOR_INCOME_AMOUNT,
  DISABILITY_BENEFITS_AMOUNT, SEVERANCE_PAY_PENSION, FOSTER_CHILD_CARE_AMT, CHILD_SUPPORT_AMT, ALIMONY_AMT).

Table 3: Description of all column-to-text mappings implemented for ACS features. The variable part of each example is shown in typeset gray font. See the ACS PUMS data dictionary for the full list of available variables.[1]

Column	*Description*	*Example*
AGEP	Age	age is 29 years old
COW	Class of worker	class of worker is Working for a non-profit organization
SCHL	Educational attainment	highest educational attainment is Bachelor's degree
MAR	Marital status	marital status is Married
OCCP	Occupation	occupation is Human Resources Manager
POBP	Place of birth	place of birth is El Salvador
RELP	Relationship	relationship to the reference survey respondent in the survey is Brother or sister
WKHP	Work-hours per week	usual number of hours worked per week is 40 hours
SEX	Sex	sex is Female
RAC1P	Race	race is Black or African American
PINCP	Total yearly income	total yearly income is $75,000
DIS	Disability status	disability status is With a disability
ESP	Employment status of parents	employment status of parents is living with two parents, only Father is employed
CIT	Citizenship status	citizenship status is Born in the United States
MIG	Mobility (lived here 1 year ago)	mobility status over the last year is lived in the same house 1 year ago
MIL	Military service	military service status is Never served in the military
ANC	Ancestry	ancestry is Single ancestry
NATIVITY	Nativity	nativity is foreign born
DEAR	Hearing	hearing status is No hearing difficulty
DEYE	Vision	vision status is With vision difficulty.
DREM	Cognition	cognition status is No cognitive difficulty
ESR	Employment status	employment status is Civilian employed, at work.
ST	State	resident state is Colorado
FER	Person has given birth within the last year	person has given birth within the last year is Person has not given birth within the last year
PUBCOV	Public health coverage status	public health coverage status is Not covered by public health insurance
JWTR	Means of transportation to work	means of transportation to work is Bicycle
PUMA	Public Use Microdata Area (PUMA) code	Public Use Microdata Area (PUMA) code is Southeast Colorado
POWPUMA	Place of Work PUMA	Public Use Microdata Area (PUMA) code for the place of work is Southeast Colorado
POVPIP	Income-to-Poverty Ratio	income-to-poverty ratio is 40% of the poverty line income, which is below the poverty line
JWMNP	Commute time	commute time is 20 minutes
GCL	Household includes grandparents	grandparent living with grandchildren is Household does not include grandparents living with grandchildren

[1] https://www.census.gov/programs-surveys/acs/microdata/documentation.html

Table 4: Description of all column-to-text mappings implemented for BRFSS features. The variable part of each example is shown in `typeset gray font`. See the details in BRFSS Documentation for the full list of available variables.[2]

Column	***Description***	***Example***
BMI5CAT	Body Mass Index (kg/m^2) category	body mass index (kg/m^2) category is `normal weight`
AGEG5YR	Age group (in intervals of 5 years)	age group (in intervals of 5 years) is `62-64 years old`
FRUIT_ONCE_PER_DAY	Consumption of fruit one or more times per day	consumption of fruit one or more times per day is `Yes`
VEG_ONCE_PER_DAY	Consumption of vegetables one or more times per day	consumption of vegetables one or more times per day is `Yes`
DRNK_PER_WEEK	Total number of alcoholic beverages consumed per week	total number of alcoholic beverages consumed per week is `23 alcoholic beverages per week`
RFBING5	Binge drinking behavior (i.e. $\geq$ 5 drinks per occasion for males, $\geq$ 4 drinks per occasion for females)	binge drinking behavior is `Yes`
TOTINDA	Leisure-time physical activity in the past 30 days	leisure-time physical activity in the past 30 days is `Yes, had physical activity or exercise during the past 30 days other than regular job.`
SMOKE100	History of smoking at least 100 cigarettes in their lifetime	history of smoking at least 100 cigarettes in their lifetime is `No`
SMOKDAY2	Current frequency of cigarette smoking	current frequency of cigarette smoking is `some days`
CHCSCNCR	Prior diagnosis of skin cancer	prior diagnosis of skin cancer is `Yes`
CHCOCNCR	Prior diagnosis of other cancer than skin cancer	prior diagnosis of other cancer than skin cancer is `No`
DIABETES	Prior diagnosis of diabetes	prior diagnosis of diabetes is `No`
POVERTY	Binary Indicator: Individual's income falls below 2021 poverty guideline for a family of four	binary indicator of whether individual's income falls below 2021 poverty guideline for a family of four is `Yes`
EMPLOY1	Employment Status	current employment status is `Retired`
IYEAR	Survey year	year of survey is `2017`
STATE	State of residence	state of residence is `Georgia`
MEDCOST	Unmet medical need due to costs in the last 12 months	unmet medical need due to costs in the last 12 months is `No`
PRACE1	Preferred race category	preferred race category is `White`
SEX	Sex	sex is `Female`

[2] `https://www.cdc.gov/brfss/annual_data/annual_data.htm`

Table 5: Description of all column-to-text mappings implemented for SIPP features. The variable part of each example is shown in `typeset gray font`. See the SIPP codebook and the pre-processing script provided by Hardt and Kim [2022] for the full list of available variables.[1]

Column	***Description***	***Example***
demographics		
AGE	age	age is `67 years old`
GENDER	gender	gender of person is `Male`
RACE	race	races the person identifies with is `White only`
EDUCATION	highest educational attainment	highest level of education completed is `9th grade`
MARITAL_STATUS	marital status	marital status is `Divorced`
CITIZENSHIP_STATUS	US citizenship status	US citizenship status is `Yes`
FAMILY_SIZE_AVG	average family size in the reference year	average number of persons in family is `2 persons`
ORIGIN	of Spanish, Hispanic, or Latino origin	Spanish, Hispanic, or Latino origin is `No`
income		
INCOME	total individual income from earnings, investment and property in the reference year	total personal income is `$6288`
HOUSEHOLD_INC	total household income in the reference year	total monthly income of all household members is `$13500`
RECEIVED_WORK_COMP	receives worker's compensation payments	received worker's compensation payments is No is
INCOME_FROM_ASSISTANCE	monthly amount received from assistance	total income from public assistance, benefits or compensation is `$6288`
SAVINGS_INV_AMOUNT	total value of IRA, KEOGH, 401k, 403b, 503b, and Thrift Savings Plan accounts	total value of retirement accounts is `$0`
hardship		
LIVING_QUARTERS_TYPE	type of living quarters	type of living quarters is `house, apartment, flat`
LIVING_OWNERSHIP	living quarters are owned, rented or occupied without payment	ownership status of living quarters is `owned or being bought by someone in the household`
FOOD_ASSISTANCE	received some form of food assistance	received food assistance is `No`
WIC_ASSISTANCE	percentage of year in which individual received assistance from WIC (Women, Infants, and Children supplemental program)	percentage of the year the respondent received WIC assistance is `0.00%`
SNAP_ASSISTANCE	percentage of year in which individual received assistance from SNAP (Supplemental Nutrition Assistance Program)	percentage of the year the respondent received SNAP/food stamps assistance is `100.00%`
health		
MEDICARE_ASSISTANCE	percentage of year in which individual received assistance from MEDICARE	percentage of the year the respondent was covered by Medicare is `100.00%`
MEDICAID_ASSISTANCE	percentage of year in which individual received assistance from MEDICAID	percentage of the year the respondent was covered by Medicaid is `0%`

(Continued on next page)

(Continued from previous page)

Column	***Description***	***Example***
HEALTHDISAB	has a physical, mental or other health condition limiting the amount of work they can do	has a physical, mental or other health condition that limits the kind or amount of work they can do is `Yes`
DAYS_SICK	number of sick days	number of days sick in the last year is `10 days`
HOSPITAL_NIGHTS	number of nights in a hospital	number of nights in a hospital is `0 nights`
PRESCRIPTION_MEDS	did take any prescription medications	uses prescription medications is `Yes`
HEALTH_INSURANCE_PREMIUMS	expenditures for comprehensive health insurance premiums	amount paid for comprehensive health insurance premiums is `$0`
HEALTH_OVER_THE_COUNTER_PRODUCTS_PAY	out-of-pocket expenditures for over-the-counter health-related products	amount paid for over-the-counter health-related products is `$100`
HEALTH_MEDICAL_CARE_PAY	out-of-pocket expenditures for on medical care	amount paid for non-premium medical out-of-pocket expenditures on medical care is `$300`
HEALTH_HEARING	has serious difficulty hearing	is deaf of has hearing difficulties is `No`
HEALTH_SEEING	serious difficulty seeing	is blind or has seeing difficulties is `Yes`
HEALTH_COGNITIVE	has serious difficulty concentrating, remembering, or making decisions	has serious difficulty concentrating, remembering, or making decisions is `No`
HEALTH_AMBULATORY	has serious difficulty walking or climbing stairs	has serious difficulty walking or climbing stairs is `Yes`
HEALTH_SELF_CARE	has difficulty dressing or bathing	has difficulty with self-care such as dressing or bathing is `No`
HEALTH_ERRANDS_DIFFICULTY	has difficulty doing errands alone	has difficulty doing errands alone is `No`
HEALTH_CORE_DISABILITY	has at least one of six core disability measures	has a core disability is `Yes, with a core disability`
HEALTH_SUPPLEMENTAL_DISABILITY	answered positively to at least one core questions, three child disability questions, or two work disability questions	answered positively to at least one core questions, three child disability questions, or two work disability questions is `Yes, with a disability`
VISIT_DOCTOR_NUM	number of visits to a doctor, nurse, or any other type of medical provider	number of visits to a doctor, nurse, or any other type of medical provider is `25 visits`
VISIT_DENTIST_NUM	number of visits to a dentist or other dental professional	number of dentist visits is `1 visit`
program participation details		
TANF_ASSISTANCE	percentage of year in which individual received assistance from TANF	percentage of the year the respondent received TANF benefit is `0.00%`
TRANSPORTATION_ASSISTANCE	received some kind of transportation assistance	receives transportation assistance is `No`
UNEMPLOYMENT_COMP	received unemployment compensation payments at any time during the reference year	receives unemployment compensation payments is `No`
UNEMPLOYMENT_COMP_AMOUNT	total amount of unemployment compensation payments	amount of unemployment compensation per month is `$0`
SOCIAL_SEC_BENEFITS	received social security benefits for themselves or on behalf of a child	received social security benefits is `No`

(Continued on next page)

(Continued from previous page)

Column	***Description***	***Example***
VA_BENEFITS_AMOUNT	total monthly amount of VA benefits	total amount of VA benefits per month is $0
RETIREMENT_INCOME_AMOUNT	total monthly amount of retirement income	total amount of retirement income per month is $0
SURVIVOR_INCOME_AMOUNT	total monthly amount of survivor income	is
DISABILITY_BENEFITS_AMOUNT	total monthly amount of payments due to sickness, accident or disability	total amount of disability benefits or income per month is $0
SEVERANCE_PAY_PENSION	received any severance pay or lump sum payments from a pension or retirement plan during the reference period	receives any severance pay or lump sum payments from a pension or retirement plan is No
FOSTER_CHILD_CARE_AMT	amount of foster child care payments received in each month	amount of foster child care payments received per month is $0
CHILD_SUPPORT_AMT	amount of child support payments in each month	amount of child support payments received per month is $0
ALIMONY_AMT	amount of alimony payments received in each month	amount of alimony payments received per month is $0

[1] https://www2.census.gov/programs-surveys/sipp/data/datasets/2014/

Table 6: Dataset statistics. Test set size is reported as *samples* $\times$ *features*.

Task	Test set size (*samples* $\times$ *features*)	Positive instances ($y = 1$)	Negative instances ($y = 0$)
ACSIncome	$166,450 \times 10$	$61,233$	$105,217$
ACSEmployment	$323,611 \times 16$	$146,740$	$176,871$
ACSMobility	$62,094 \times 21$	$16,446$	$45,648$
ACSTravelTime	$146,665 \times 16$	$64,285$	$82,380$
ACSPublicCoverage	$113,829 \times 19$	$33,971$	$79,858$
BRFSS Blood Pressure	$84,676 \times 19$	$44,586$	$40,090$
SIPP Poverty	$3,972 \times 50$	$2,035$	$1,937$

A.3 Experimental Details

We follow the default configuration provided by `folktexts`, adopting a random 80/10/10 split for training, validation, and test sets. All evaluations are performed exclusively on the test set; no models are trained in this study. For few-shot prompting experiments, we randomly sample 10 examples from the training set to construct the prompt context. These same examples are reused across all few-shot prompts to ensure that each model receives an identical prompt.

To report performance of XGBoost, we use the implementation provided by `scikit-learn`, with default hyperparameters. No additional hyperparameter tuning was performed.

Table 6 summarizes the datasets used in this study. The main text focuses on results for positive instances (i.e., samples with true label $y = 1$). Results for negative instances ($y = 0$) are presented in Appendix D.4.

Code. We provide the code necessary to reproduce our analysis, along with a step-by-step guide for obtaining model predictions, available here: `https://github.com/socialfoundations/mono-multi`.

Resources used. We use an internal compute cluster with NVIDIA A100 and H100 GPUs. Zero-shot evaluation of all models required approximately 1000 GPU hours, while 10-shot prompting added an additional $2,500$ GPU hours.

B Aggregate Model Performance

Aggregate performance varies considerably across models and tasks. See Figure 6 for accuracy and Figure 7 for balanced accuracy. Thresholds for all models where chosen to maximize balanced accuracy. For task with imbalanced label distributions, such as ACSMobility and ACSPublicCoverage, only a few models outperform the constant majority-class baseline when selected for inclusion in the Rashomon set based on overall accuracy (Figure 6). Generally, we observe that models with more parameters tend to perform better and are frequently included in the Rashomon set (highlighted in black), a trend that is even more pronounced under balanced accuracy. Variation in performance across models is lower for balanced accuracy, likely because model predictions are explicitly optimized for this metric (Figure 7). Finally, we find that high performance on one task rarely translates to other tasks, indicating limited cross-task generalizability.

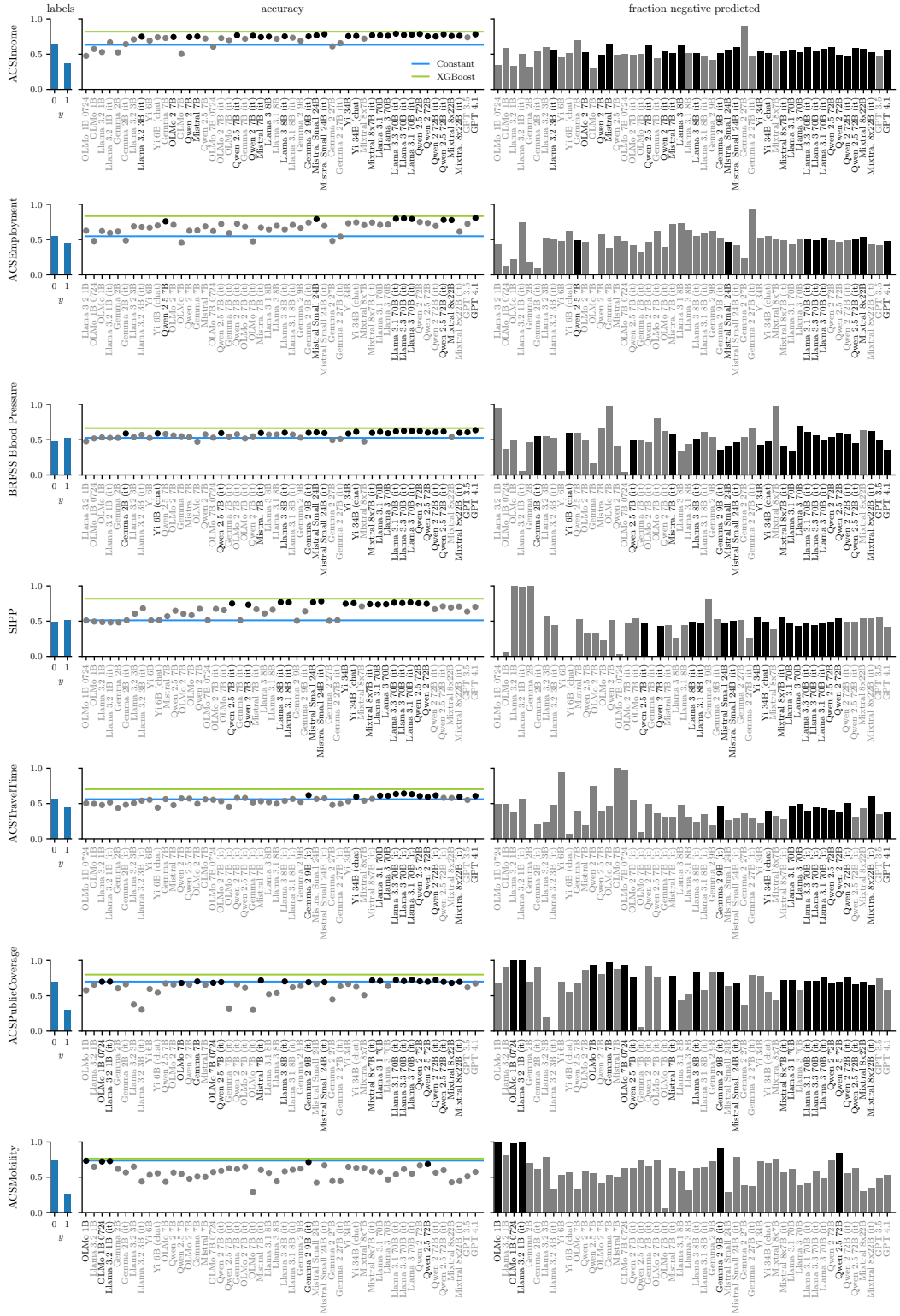

Figure 6: **Zero-shot performance** on the test set. Each row is a task; columns show the ground truth label distribution, **accuracy**, and the fraction of negative predictions. Models in the middle and right panel are ordered by parameter size. In the accuracy panel, the blue line marks the constant majority-class predictor, and the green line indicates XGBoost performance. Models included in the Rashomon set ($\epsilon = 0.05$) are shown in black; others in gray.

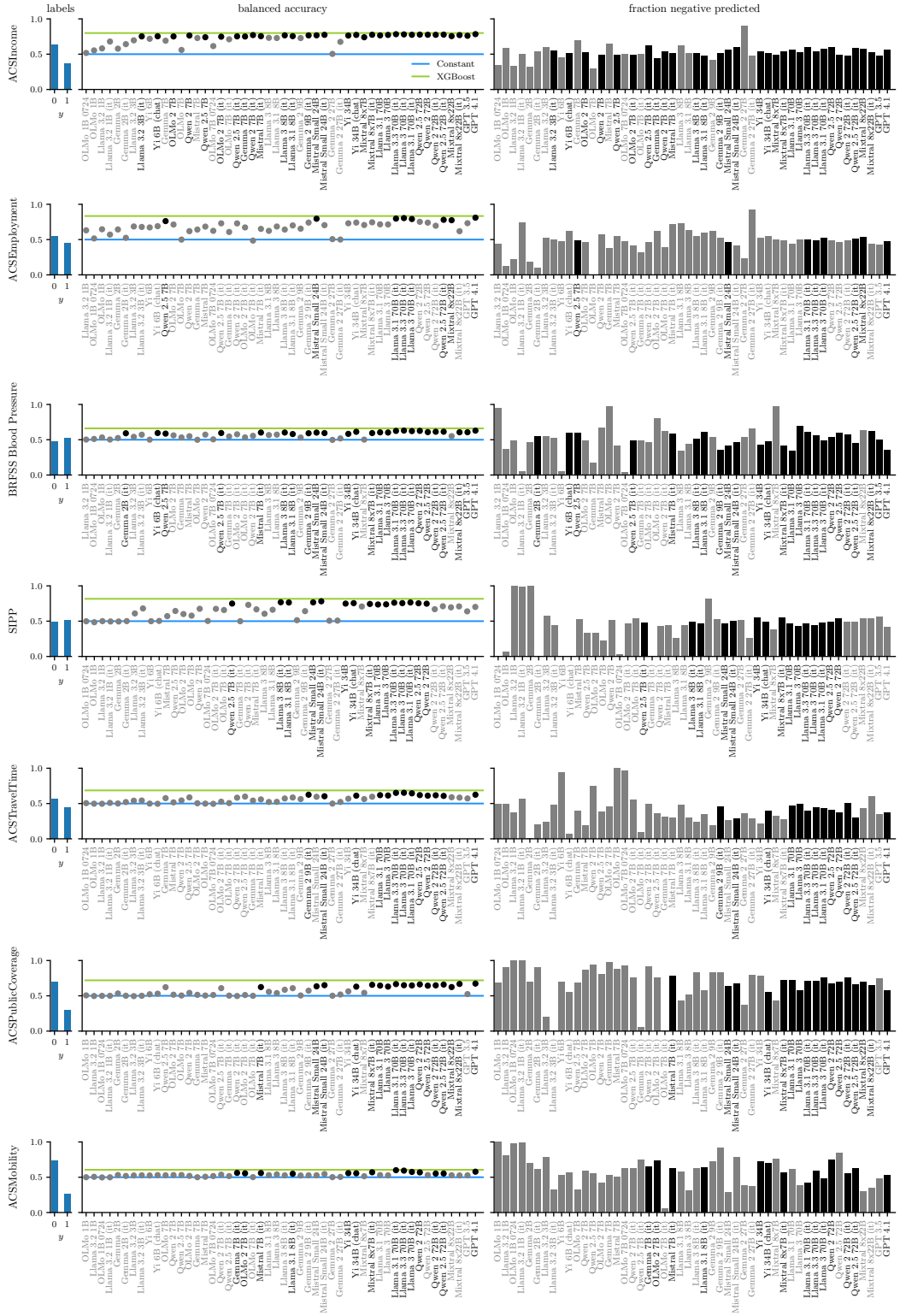

Figure 7: **Zero-shot performance** on the test set. Each row is a task; columns show the ground truth label distribution, **balanced accuracy**, and the fraction of negative predictions. Models in the middle and right panel are ordered by parameter size. In the accuracy panel, the blue line marks the constant majority-class predictor, and the green line indicates XGBoost performance. Models included in the Rashomon set ($\epsilon = 0.05$) based on balanced accuracy are shown in black; others in gray.

C Zero-Shot Prompting with Identical Prompts: Disaggregated Results

In the main paper, we report agreement and recourse levels aggregated across base and instruction-tuned models from different model providers, and across individuals in the test dataset to assess the model landscape. In this section, we break down our results along several axes. The first section focuses on model-specific characteristics: model provider and model variant. In the second section, we disaggregate results by demographic groups (sex, race and age) to examine how predictive similarity among language models affects these groups.

C.1 Results by Model Attributes

In this section, we present fine-grained results by model attributes, focusing on model provider and model variant. To enable fair comparisons of how agreement rates and recourse levels vary across different model groups, we subsample larger groups to match the size of the smaller group. Reported values then correspond to the mean and standard error computed over 1,000 independent repetitions, or over the maximum number of unique subsamples possible given the group sizes. Results on the ACSMobility task are omitted, as its Rashomon set is small to begin with, and further breakdowns would produce prohibitively small model groups.

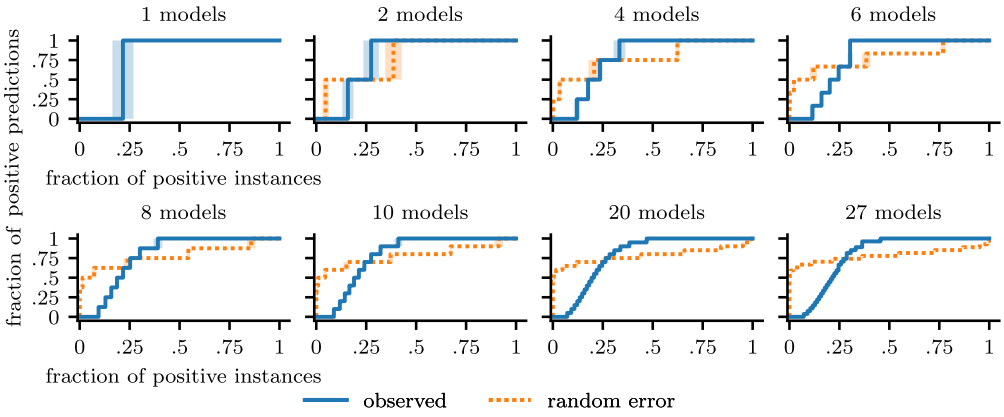

Figure 8: Recourse curves are sensitive to the model set size. Illustrative example based on the Rashomon set for ACSIncome ($\epsilon = 0.05$).

Recourse is sensitive to the model set size. To motivate the subsampling, we illustrate how recourse curves depend on the size of the model set. By expectation, full agreement is more likely in smaller sets, and the recourse curves are correspondingly coarser with fewer distinct values. In larger model sets, achieving full agreement requires a high degree of predictive similarity across all models. Even a single model deviating from the majority directly reduces the fraction of individuals experiencing either full or no recourse. At the same time, the impact on the individual is minimal: an individual who experiences nearly full recourse is still very likely to receive a favorable outcome in a large model ecosystem, without relying on recourse. We demonstrate this on the ACSIncome task by varying the number of models (not the restrictiveness of the Rashomon set!) and reporting the mean and standard error in Figure 8. Recall, that models in the Rashomon vary by up to 5 accuracy points.

To summarize, our metric *recourse with respect to a model set* is inherently sensitive to the set size, as is the random error baseline. To ensure comparability across model groups resulting from disaggregation, we subsample the larger groups without replacement to match the size of the smallest group.

C.1.1 Disaggregation by Model Provider

Results in the main paper are aggregated over models from different families and developers to assess the empirical landscape of the current language model ecosystem. Technical reports for various open-source models reveal that their pre-training data mixes frequently draw from overlapping

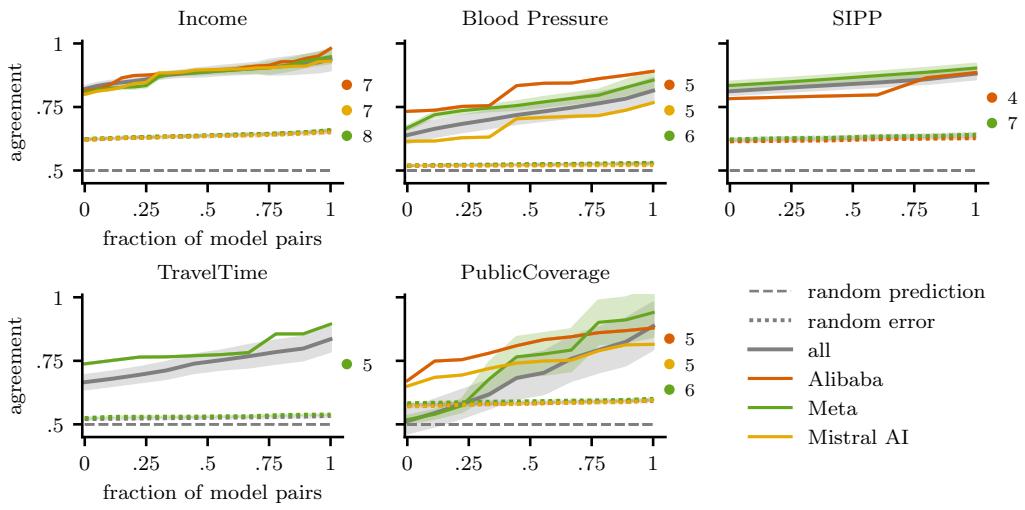

Figure 9: **Agreement rates by model provider:** x% of model pairs agree on at most y% of positive instances. Models are grouped by provider, showing only groups with at least four models, subsampled to match the smallest group. Mean and standard error are reported; group sizes are in the task legend. Across all groups, observed agreement (solid colored) exceeds random-error (dotted) and random-prediction (dashed gray) baselines, but remains below strict monoculture (agreement = 1). For most tasks, within-group agreement is slightly higher than across the full Rashomon set (solid gray).

sources such as CommonCrawl, Wikipedia and Wikibooks, GitHub, ArXiv, and Semantic Scholar, indicating considerable overlap in the training datasets even across developers. In this section, we break results down by model developer, yielding subgroups whose models are likely to share not only larger portions of training data, but also similar architecture choices and training pipelines. Under the component-sharing hypothesis proposed by Bommasani et al. [2022], such increased overlaps may lead to higher within-group agreement. Examining these patterns sheds light on the effects of increased component sharing, and also helps anticipate the potential outcomes when models from a single developer dominate a domain or when training pipelines converge across the ecosystem.

Note that for all tasks, the Rashomon set includes models from many different providers, often resulting in very small groups with only one or two models, which limits interpretability. To address this, we report results only for groups with at least four models; as a result, the disaggregated results presented here do not cover the full Rashomon set. For ACSEmployment, all groups fell below this threshold, so this task is excluded from the analysis in this section.

Agreement rates. When adjusting for the size of the model sets, we find that within-provider agreement is elevated for some groups but generally comparable to the overall level of agreement computed across all models (Figure 9). In particular, on ACSIncome, within-group agreement is nearly indistinguishable from the overall average. In contrast, within-group agreement varies notably on BRFSS Blood Pressure. Notably, agreement among Mistral models is slightly below the overall level and considerably lower than that of other providers, which may reflect the heterogeneity of the Mistral group, which includes both dense and mixture-of-experts (MoE) models. On ACSPublicCoverage, the Meta group exhibits high standard error, suggesting that some models provided by Meta are more similar to each other than others. Across tasks, we find that models from Meta and Alibaba are consistently contained in the Rashomon set and display slightly elevated within-group agreement. Nevertheless, there is no consistent pattern indicating that a single provider's models consistently show exceptionally high within-group agreement, and very few model pairs reach full agreement. Thus, even restricted to a single provider, agreement remains well below the level of strict monoculture.

Baseline agreement under random errors remains relatively stable across subgroups, suggesting that overall accuracy differences alone cannot explain the observed disparities. Instead, the slightly elevated agreement among models from the same model provider likely reflects shared training

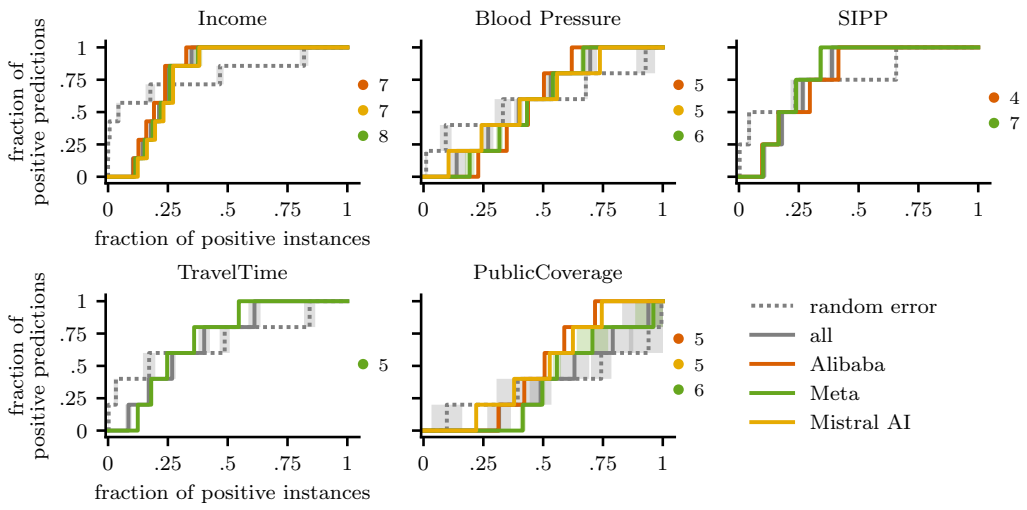

Figure 10: **Recourse curves by model provider:** x% of positive instances are accepted by at most y% of models. Only provider groups with ≥ 4 models are shown; all groups are subsampled to match the smallest group. Mean (solid colored) and standard error (shaded) are reported, with group sizes in the task legends.

data, architecture choices or training pipelines, in line with the shared component hypothesis of Bommasani et al. [2022].

Recourse curves. When disaggregating recourse levels by model provider, the overall picture remains stable: both no recourse and full recourse occur much more frequently than would be expected under random errors (Figure 10). Even when restricted to a single provider, disagreement between models still creates opportunities to find recourse in other models. Across all tasks, the majority of positive instances experiences substantive or full recourse, receiving a favorable outcome by the majority of models.

Comparing within-group recourse levels to those computed across all models in the Rashomon set shows a tendency toward more homogeneous outcomes, with larger fractions of positive instances experiencing either no or full recourse. On BRFSS Blood Pressure, for example, the higher agreement among models from Alibaba translates into increases in both extremes, with more individuals facing no recourse as well as more achieving full recourse compared to the fraction computed across all models in the Rashomon set. On ACSPublicCoverage the pattern differs: for models from Alibaba the prevalence of no recourse remain comparable to that across all models. Increased agreement levels mostly stem from true positives, visible from higher levels of full recourse. Similarly, full recourse is also increased among models from Mistral AI, but in combination with a decreased fraction of individuals experiencing no recourse resulting in overall higher recourse levels. Models from Meta show the reverse pattern: a higher fraction of no recourse and comparable levels of full recourse suggest that agreement among Meta models arises primarily from shared false negatives. Because these observations vary by task and by provider, consistent generalizations are difficult. We find some evidence that restricting to a single provider can amplify the extreme levels of recourse compared to those computed across all models in the Rashomon set.

C.1.2 Disaggregation by Model Variants (base vs. instruction-tuned)

In our work we include both base and instruction-tuned variant when available because both variants might be attractive for decision-makers for various reasons. Importantly, as shown in Figure 6, we find that across many tasks, there is often no clear performance gap between base models and instruction-tuned variants. Consequently, models of both types are present in the Rashomon sets of all tasks. Including both variants thus provides a more comprehensive picture. In this section, we disaggregate the results from the main paper by model variant as it can serve to analyze the impact of instruction-tuning on model agreement and recourse opportunities experienced by individuals.

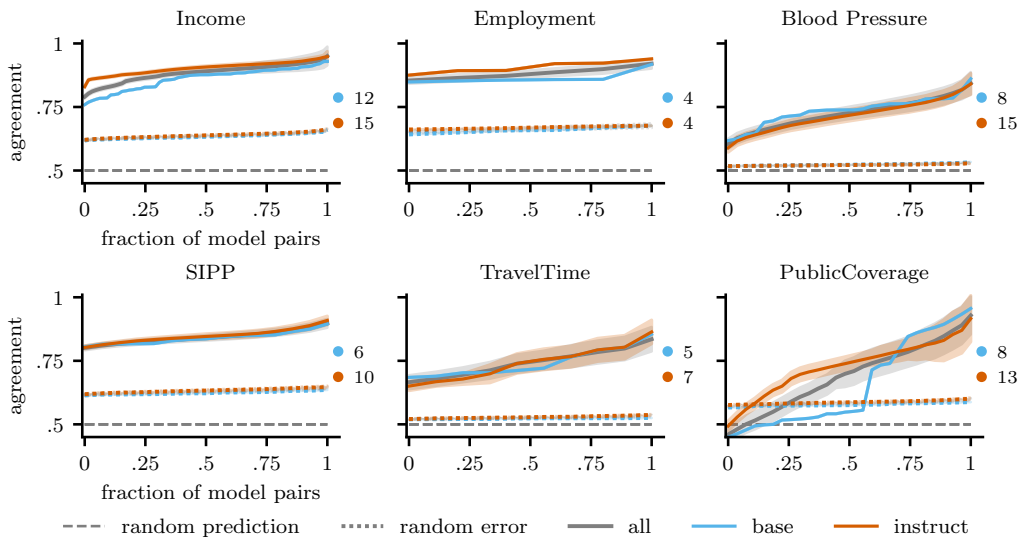

Figure 11: **Agreement rates grouped by model variants:** x% of model pairs agree on at most y% of the positive instances. The Rashomon set is split into base and instruction-tuned models. For both groups, observed agreement (solid colored) is higher than under random errors (corresponding dotted) and random predictions (dashed gray), but remain below strict monoculture (red). The solid gray line depicts agreement among all models in the Rashomon set, while the dotted gray line shows the corresponding random-error baseline. For most tasks, average agreement among instruction-tuned variants in the Rashomon set is slightly above that among base models.

Agreement rates. When comparing agreement rates between base and instruction-tuned models, we find them largely comparable, with a slight tendency for pairwise agreement to be higher among instruction-tuned models (Figure 11). This is most pronounced for ACSIncome and ACSEmployment. Across tasks, instruction-tuned models account for at least half of each group and, in most cases, constitute the majority. This suggests that instruction-tuning improves balanced accuracy. However, the effect on predictive similarity across models is moderate. A notable exception is ACSPublicCoverage, where a subset of base model pairs exhibits distinctly low, partially anti-correlated predictions with agreement rates below random, while other models shows relatively high agreement (mean 65.6%). In contrast, most instruction-tuned model pairs are highly correlated, with an mean agreement rate of 72.62%. These differences may be related to low predictive signal and class imbalance of this task.

Recourse levels. When disaggregating recourse levels by model variant, the overall picture remains stable: both no recourse (receiving an unfavorable outcome from all models) and full recourse (receiving a favorable outcome from all models) occur much more frequently than would be expected under random errors (Figure 12). Even when restricted to a specific model variant, disagreement between models creates opportunities for individuals to find recourse in other models. Across tasks, recourse curves are largely comparable to those computed across all models in the Rashomon set. On ACSIncome and ACSEmployment, instruction-tuned models show a slightly increased fraction of individuals correctly receiving a favorable outcome from all models, suggesting that higher agreement rates stem primarily from shared true positives. The opposite pattern is observed for BRFSS Blood Pressure, where full recourse is slightly more prevalent among base models. For ACSPublicCoverage, instruction-tuned models tend to produce positive predictions more frequently, resulting in a larger fraction of individuals experiencing substantive recourse compared to base models. Nevertheless, within both base and instruction-tuned models, a substantial fraction of individuals still faces no recourse.

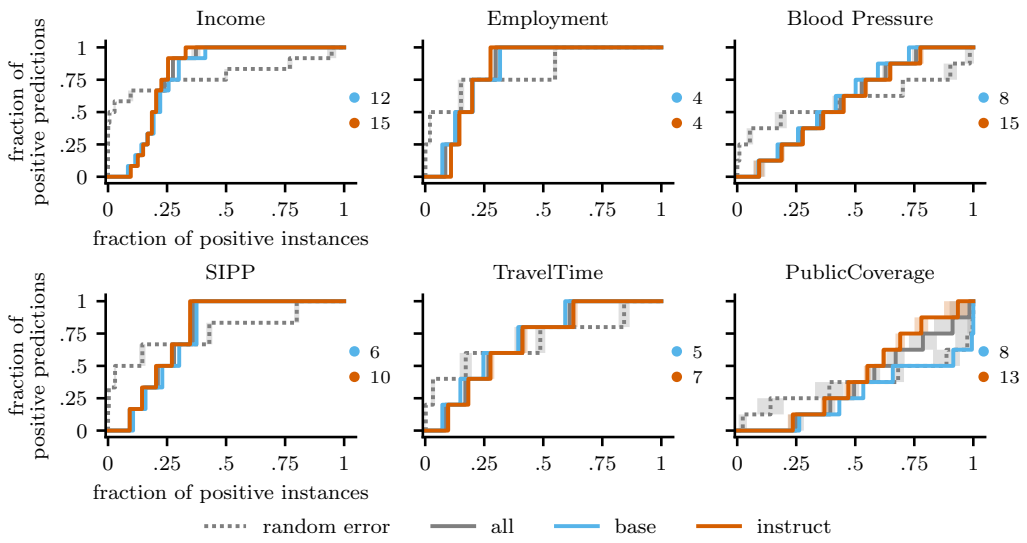

Figure 12: **Recourse curves grouped by model variants:** x% of the positive instances are accepted by at most y% of the models. The Rashomon set is split into base and instruction-tuned models. The solid gray line depicts observed recourse among all models in the Rashomon set, while the dotted gray line shows the corresponding baseline under random errors. Observed recourse curves for base and instruction-tuned models are depicted in blue and orange respectively.

C.2 Results by Demographic Attributes

In this section, we provide fine-grained results for different demographic groups: sex, race and age, features that are present among all tasks. For the plots we define a unified mapping to obtain similar groups (if present and defined) across tasks. We report results for demographic groups that account for at least 5% of the available data.

C.2.1 Disaggregation by Sex

Across all tasks, sex (or gender in the case of SIPP) is encoded as a binary variable with two categories: male and female. When stratifying individuals by this variable, we observe that agreement rates are largely comparable across tasks, with models showing slightly higher agreement for male individuals in most cases (Figure 13). This is also the case when looking at recourse levels (Figure 14). The largest group-level disparity is observed in the BRFSS Blood Pressure task. Although agreement rates between male and female individuals are nearly identical, an examination of recourse levels reveals systematically higher recourse for males. This pattern may indicate a shared bias among models that could disadvantage female individuals in downstream decision-making. Alternatively, the observed pattern could reflect intersectional effects between sex and age on the prevalence of blood pressure, which we have not investigated further.

C.2.2 Disaggregation by Race

To analyze results by race, we standardize the encoding across tasks resulting in the follwing categories: American Indian or Alaska Native, Asian, Black, Multiracial, Native Hawaiian or Other Pacific Islander, Other, and White. We report outcomes only for race groups representing at least 5% of the test data, but disparities are also prevalent for smaller groups. Note that, following the preprocessing procedure from `tableshift`, only binarized race information is available for BRFSS Blood Pressure. In this case, individuals get mapped to the categories White and Non-White, with the latter labeled as "Other" in the plots, since it does not correspond to a specific racial group.

Disaggregating results by race reveals substantial variation between groups in both within-group agreement rates (Figure 15) and within-group recourse levels (Figure 16). Across tasks, the majority of individuals in the datasets are White. For most tasks, agreement rates for this majority group closely

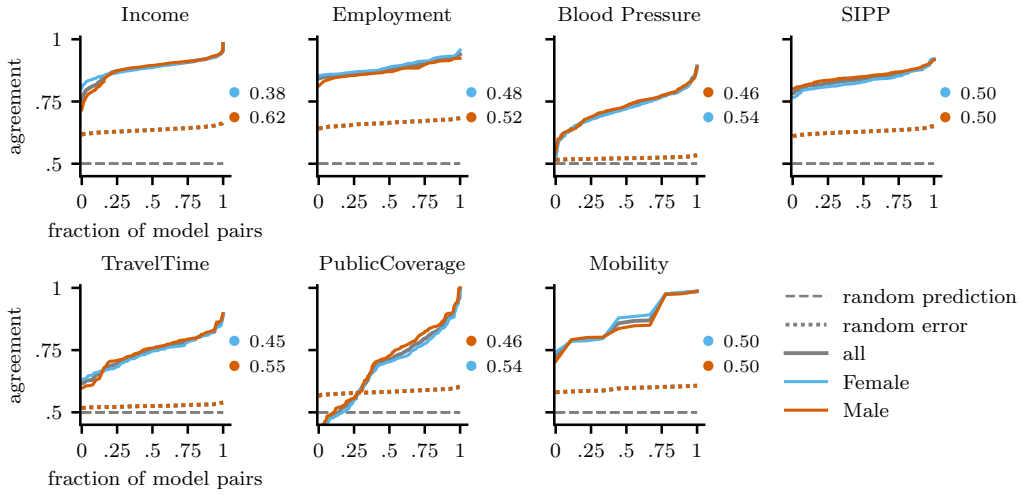

Figure 13: **Agreement rates by sex:** x% of model pairs agree on at most y% of the positive instances, with data grouped by sex. For both groups, observed agreement (solid colored) exceeds that expected under random errors (dotted) and random predictions (dashed gray), but remains below strict monoculture (red). The solid gray line represents agreement across all models in the Rashomon set, while the dotted gray line indicates the corresponding random-error baseline. Overall, agreement rates are comparable between groups across tasks. Task legends indicate group sizes.

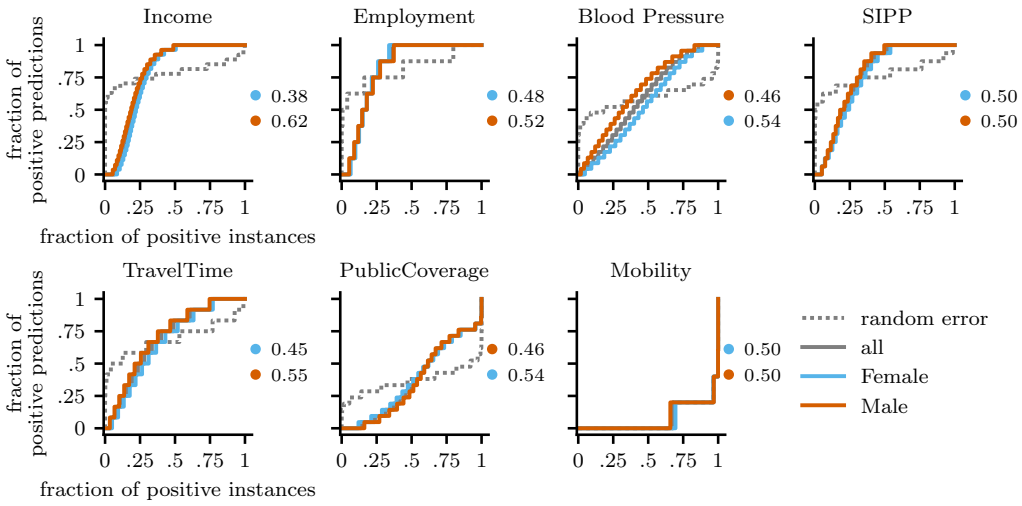

Figure 14: **Recourse curves grouped by sex:** x% of the positive instances are accepted by at most y% of the models, with data grouped by sex. Task legends indicate group sizes. Observed recourse curves are comparable between groups.

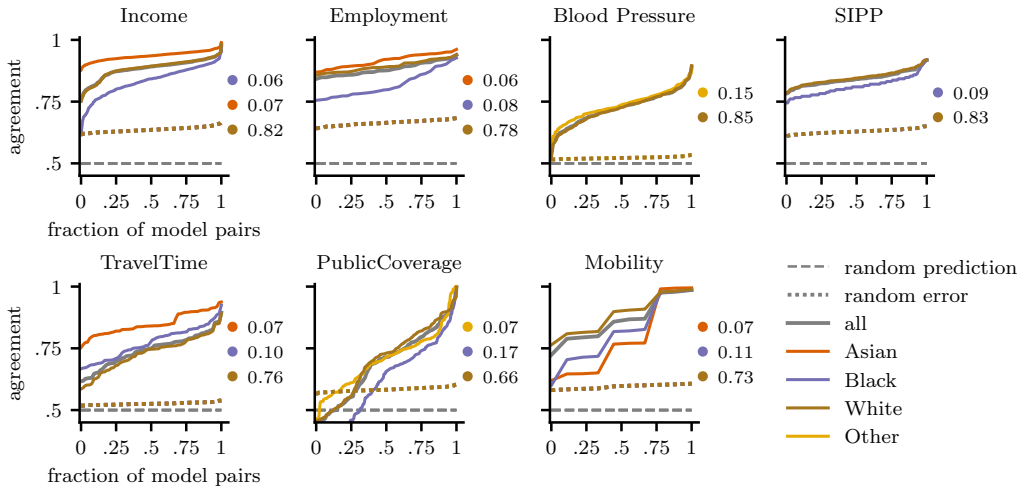

Figure 15: **Agreement rates grouped by race:** x% of model pairs agree on at most y% of the positive instances, with data grouped by race. For all groups, observed agreement (solid colored) is higher than under random errors (corresponding dotted) and random predictions (dashed gray), but remain below strict monoculture (red). The solid gray line depicts agreement among all models in the Rashomon set. For most tasks, average agreement among instruction-tuned variants in the Rashomon set is slightly above that among base models.

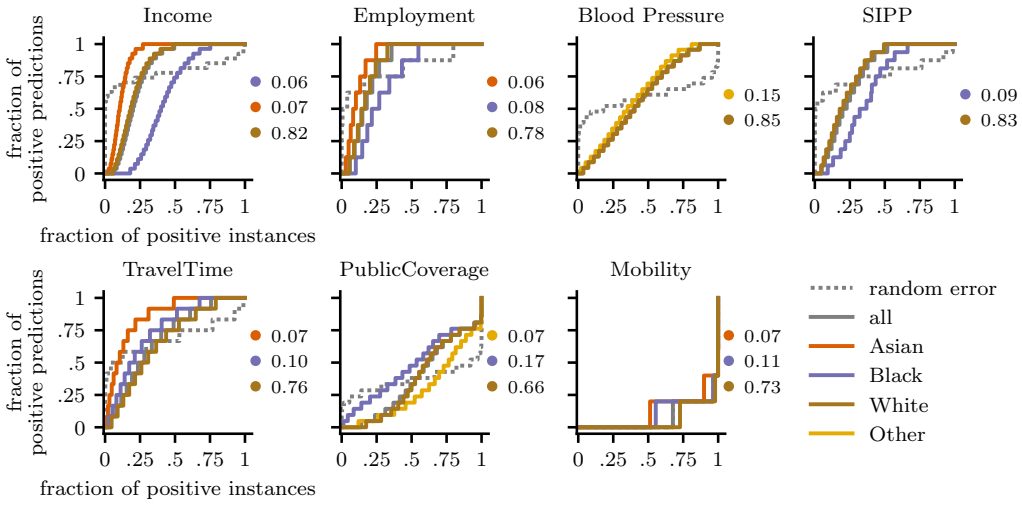

Figure 16: **Recourse curves grouped by race:** x% of the positive instances are accepted by at most y% of the models, with data grouped by race. Task legends indicate group sizes. Observed recourse curves vary systematically between groups.

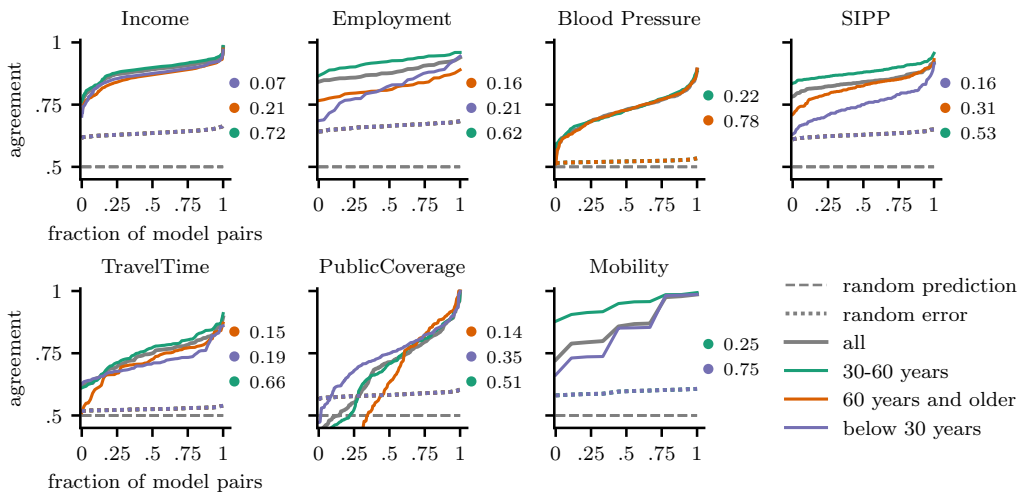

Figure 17: **Agreement rates grouped by age:** x% of model pairs agree on at most y% of the positive instances, with data grouped by age. Task legends indicate group size. Observed agreement (solid colored) lies between consistently the random-error baseline (corresponding dotted) and strict monoculture (red). Agreement rates vary systemically between groups.

match overall agreement rates. An exception is ACSMobility, where overall pairwise agreement rates are elevated for individuals identified as White. Recourse levels generally follow the same pattern. For tasks including individuals identified as Black, agreement rates for this group tend to be lower, a trend consistent across all tasks except ACSTravelTime. This suggests that Black individuals experience greater multiplicity in the examined model set. In terms of recourse, we observe that for ACSIncome, ACSEmployment and SIPP, Black individuals face systemic exclusion (no recourse) at much higher rates and receive full recourse at lower rates than other groups. This pattern is particularly pronounced in ACSIncome, despite this task having the largest empirical Rashomon set. By contrast, on ACSTravelTime and ACSPublicCoverage, a higher fraction of Black individuals experiences high levels of recourse. The elevated agreement rates on ACSTravelTime can be explained by a larger proportion of individuals receiving full recourse.When present, individuals identified as Asian generally experience higher agreement rates among models in the Rashomon set compared to both the overall agreement rates and those of other groups. This is primarily driven by an increase in shared true positives, as reflected in the fraction of individuals receiving full recourse. Correspondingly, the fraction of Asian individuals facing no opportunity for recourse tends to be lower than for other groups.

Overall, the disaggregated results indicate that agreement and recourse patterns vary systematically by race: White and Asian individuals often experience higher agreement and recourse levels, whereas Black individuals exhibit lower agreement and are more frequently subject to systemic exclusion across several tasks.

C.2.3 Disaggregation by Age

We categorize individuals into three age brackets: younger than 30 years, 30–60 years, and 60 years or older. Note that some tasks are restricted to specific age ranges. For instance, most ACS tasks consider only individuals aged 16 and older, ACSMobility restricts the population to those younger than 35, and the BRFSS Blood Pressure dataset includes only individuals aged 50 and above.

Disaggregating results by age reveals systematic variation in both within-group agreement rates (Figure 17) and within-group recourse levels (Figure 18). Across tasks, the majority of individuals are aged 30–60, with the exceptions of BRFSS Blood Pressure and ACSMobility due to the aforementioned age restrictions. For most tasks, agreement rates within this age group are higher and closely match or exceed overall agreement rates. Recourse levels follow a similar pattern, with a larger fraction of individuals in this age range receiving full recourse. Individuals younger than 30 tend to experience lower recourse levels across multiple tasks, although these disparities are generally

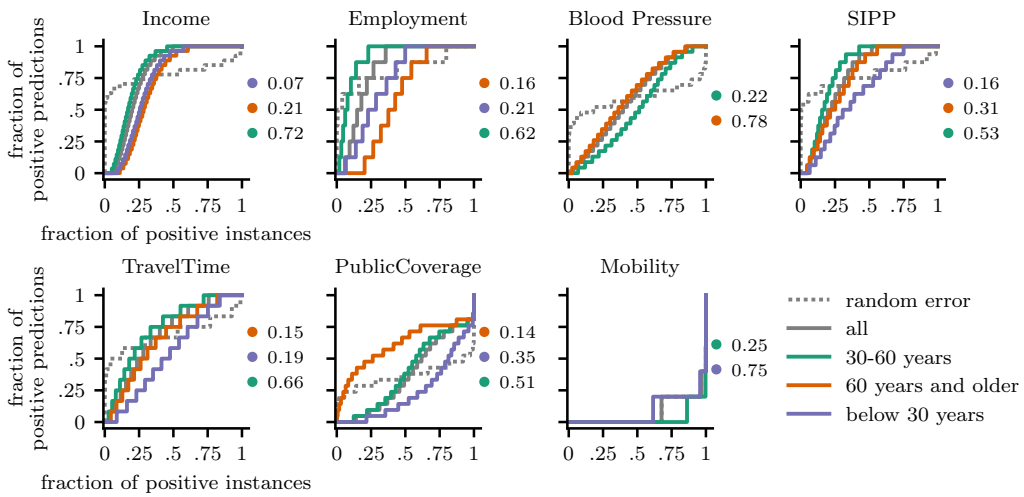

Figure 18: **Recourse curves grouped by age:** x% of the positive instances are accepted by at most y% of the models, with data group by age. Task legends indicate group sizes. Observed recourse curves vary systematically between groups.

less pronounced than those observed for the 30–60 age group. The most notable disparities occur in ACSEmployment, where younger individuals are disproportionately classified as false negatives, receive lower levels of recourse overall, and exhibit lower agreement rates. Individuals aged 60 years or older show a recourse curve with a slope similar to that of younger individuals, but a substantially larger fraction of positive instances (nearly 25%) experience no recourse. Overall, the disaggregated results indicate that agreement and recourse patterns vary systematically by age: individuals aged 30–60 generally experience higher agreement and greater access to recourse, whereas younger individuals face lower agreement and more limited recourse across several tasks. The relationship of the recourse curve for individuals aged 60 years or older to those of other age groups appears task-dependent.

The disaggregated analyses may serve as an initial step toward understanding how recourse patterns vary across demographic groups, providing indicative evidence that some groups may face disproportionate barriers to recourse in the present language model ecosystem. It remains an important avenue for future research to further investigate the mechanisms underlying such disparities and ways to mitigate them.

D Zero-Shot Prompting with Identical Prompts: Additional Results

In this section, we provide additional details on our results when zero-shotting different models with identical prompts. Section D.1 presents a more detailed discussion of results across tasks, Section D.2 shows the effect of varying ϵ, which determines the restrictiveness of the Rashomon set, and Section D.3 examines the impact of selecting models based on their balanced accuracy.

D.1 Agreement and Recourse Levels Across Tasks

To complement the results presented in the main paper, we provide a more detailed analysis across tasks in this section, including agreement rates between model pairs (Figure 19).

As reported in detail for ACSIncome, we observe generally high recourse levels among the positive instances, with a small, but considerable fraction experiencing no recourse (Figure 19, bottom panel). Compared to a random-error baseline, extreme outcomes – either no or full recourse – are more likely, reflecting shared inductive biases between models. Nevertheless, the distribution of recourse levels remains far from a strict monoculture, such that individuals largely retain the opportunity to obtain favorable outcomes by switching models. We find these patterns to be generally consistent across tasks (Figure 2 and 19). Empirically observed agreement rates significantly exceed those

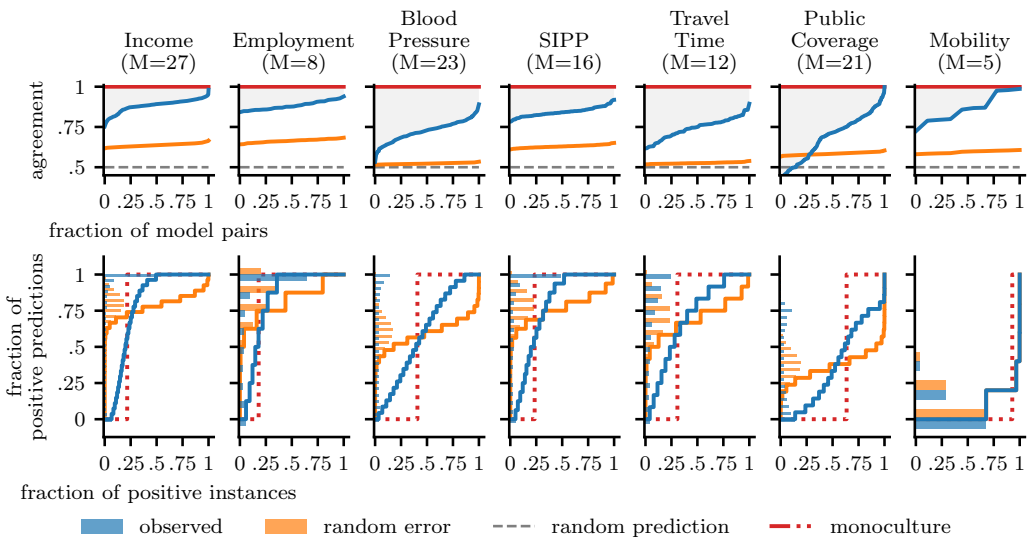

Figure 19: Agreement rates and recourse curves for accuracy-based selection of models. **Top.** Agreement curves across tasks: x% of model pairs agree on at most y% of the positive instances. Observed agreement (blue) is higher than under random errors (orange) and random predictions (dashed gray), but well below monoculture (red). **Bottom.** Recourse curves across tasks: x% of the positive instances are accepted by at most y% of the models. We zero-shot models and select those that achieve accuracy within $\epsilon = 0.05$ from the best. For example, on ACSIncome we observe (blue) 20% of positive instances being accepted by at most 50% of the models. Under random errors (orange) this would rarely happen. Under strict monoculture (red, dotted) individuals only experience no or full recourse. Here, the mean TPR is used for illustration. The bar plot on the y-axis shows density function of recourse level in the population.

expected under random errors, showing substantial predictive similarity across language models, but generally remain below complete consensus. As as result, both agreement rates as well as recourse curves consistently occupy the middle ground between the extremes of strong multiplicity and strict monoculture, reflecting some degree of diversity in model behavior even within the Rashomon set.

Two tasks, ACSMobility and ACSPublicCoverage, deviate notably from the overall pattern, showing higher rates of no recourse and smaller proportions of individuals with substantial recourse. These deviations likely reflect the influence of two underlying characteristics common to both tasks, whose combination may further amplify the observed differences. First, they exhibit low predictive signal [Ding et al., 2022]. We assess this by examining the performance gap between a majority-class predictor and XGBoost, a strong baseline for tabular data (Figure 6). Small performance gaps indicate that models hardly improve over trivial baselines. Second, both tasks display high class imbalance, with a low prevalence of the positive class. Because the Rashomon set is constructed based on overall accuracy, it tends to favor models that prioritize performance on the majority class. This effect is particularly pronounced for ACSMobility, where the (small) Rashomon set includes mainly models behaving highly similar to the constant majority-class predictor. Consequently, agreement among some models is close perfect and a large fraction of individuals receive no recourse. Notably, for ACSMobility, recourse levels align closely with those observed under the respective the random-error baseline. For ACSPublicCoverage, recourse levels are generally lower than in other tasks, reflecting the high negative prediction rates of models in the empirical Rashomon set (see Figure 6, right panel). Nonetheless, 38% of positive instances receive substantial recourse. This task also exhibits elevated levels of ambiguity and prediction discrepancy (Figure 20), consistent with unexpectedly low agreement rates, that for a subset of models even fall below the baseline of random predictions. Notably, the lowest agreement rates are observed for model pairs comprising one large and one small model, suggesting that differences in model capacity may contribute to the divergence in predictions.

As discussed in Appendix D.3 (Figure 22), selecting models using balanced accuracy rather than overall accuracy partially mitigates these effects, reducing the fraction of individuals receiving no

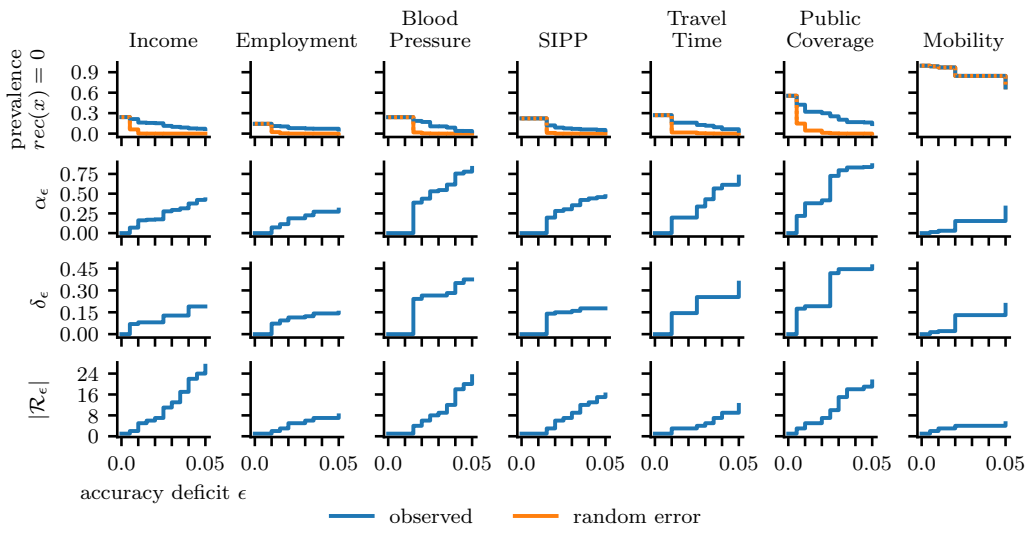

Figure 20: Severity of monoculture and predictive multiplicity as a function of accuracy deficit ϵ from the best model. Each column corresponds to one task. Predictions are obtained via 0-shot prompting. Models are selected based on overall accuracy. **Top.** The fraction of positive instances experiencing no recourse (blue) is consistently higher than would be expected under random errors (orange). Although this gap persists, it decreases as ϵ increases. **Second Row.** Ambiguity increases with ϵ, highlighting the fraction of individuals affected by model choice. Due to monotonicity, ambiguity is likely higher for the full Rashomon set. **Third Row.** Discrepancy increases with ϵ, potentially affording opportunities for recourse for some individuals. Due to monotonicity, discrepancy is likely higher for the full Rashomon set. **Bottom.** Number of models in the Rashomon set as ϵ set increases.

recourse and increasing substantial or full recourse. Taken together, these observations suggest that task-dependent factors such as class balance, the predictive signal of the prediction task, and the selection criterion for the Rashomon set can influence the observed levels of recourse individuals receive.

D.2 Effect of the Choice of Accuracy Deficit ϵ

In the main analysis, we consider models whose accuracy lies within an absolute margin of $\epsilon = 0.05$ of the best-performing model. This choice aligns with ranges commonly explored in prior work [Marx et al., 2020, Hsu et al., 2023, Watson-Daniels et al., 2023] and strikes a practical balance between including plausible deployment candidates from multiple major language model providers and avoiding overly narrow empirical Rashomon sets. Because a 0.05 difference in accuracy can correspond to a substantial change in error depending on the task and the best model's performance, this section explores the sensitivity of our findings to more restrictive values of ϵ.

Figure 20 reproduces results from the main analysis, showing the prevalence of no recourse and discrepancy across varying values of ϵ, now also including ambiguity as an additional measure of multiplicity. At $\epsilon = 0.0$, the Rashomon set contains only the best model found. Across tasks, we observe that small increases in ϵ rapidly expand the empirical Rashomon set as well as the observed discrepancy (δ_ϵ) and ambiguity (α_ϵ), indicating that many models achieve comparable accuracy while producing divergent individual-level predictions. This diversity creates potential avenues for recourse and reduces the likelihood that decision makers converge on a single model. Nevertheless, for all ϵ values where the Rashomon set contains more than one model, a non-negligible fraction of individuals experiences no recourse—substantially higher than would be expected under random errors.

To further examine the effect of ϵ, we provide agreement and recourse curves for varying ϵ values in Figure 21. The results show a pattern consistent with the main analysis. For small ϵ, the empirical Rashomon set typically contains very few models, often only the best-performing one (see also

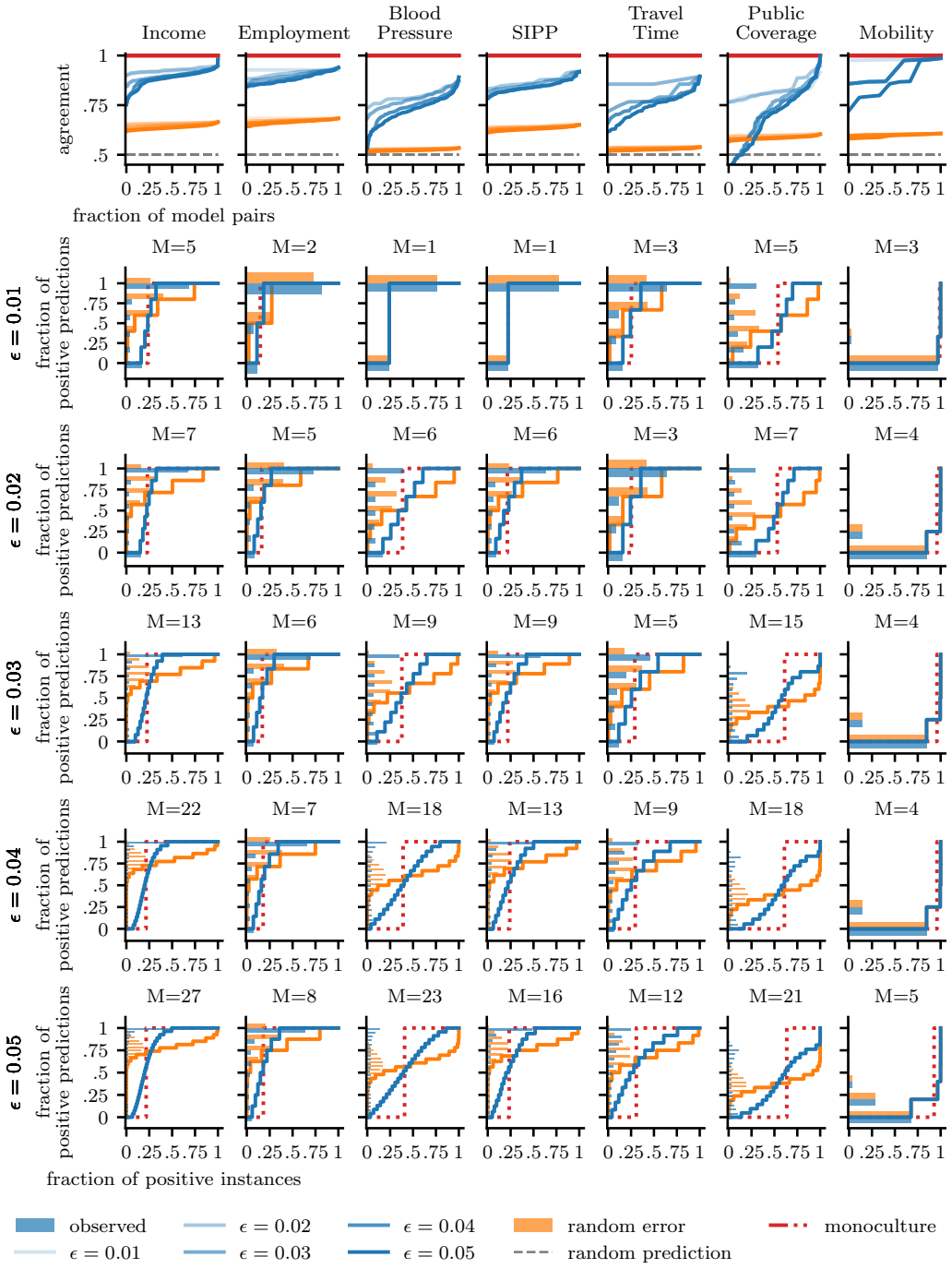

Figure 21: Agreement rates and recourse curves for different values of ϵ. **First row.** Agreement curves: x% of model pairs agree on at most y% of the positive instances. Observed agreement (blue) consistently lies between strong multiplicity (orange) and strict monoculture (red). **Rows 2-6.** Recourse curves for varying values of ϵ: x% of the positive instances are accepted by at most y% of the models. We zero-shot models and select those that achieve accuracy within ϵ from the best. For small ϵ, the empirical Rashomon set can become restrictively small.

Table 7: Recourse levels and measures of multiplicity for all tasks under zero-shot prompting when using balanced accuracy as selection criterion for the Rashomon set.

task	$\lvert\mathcal{R}_\epsilon\rvert$	no recourse	substantial recourse	full recourse	ambiguity	discrepancy
ACSIncome	32	0.05	0.81	0.56	0.39	0.15
ACSEmployment	8	0.06	0.82	0.64	0.30	0.15
BRFSS Blood Pressure	25	0.03	0.59	0.11	0.86	0.37
SIPP	15	0.05	0.78	0.49	0.46	0.18
ACSTravelTime	13	0.03	0.77	0.29	0.67	0.25
ACSPublicCoverage	17	0.09	0.51	0.22	0.69	0.37
ACSMobility	14	0.12	0.46	0.07	0.81	0.43

Figure 20, bottom row). When the Rashomon set includes at least two models, agreement rates exceed what would be expected under random errors, indicating substantial predictive similarity across language models. Nevertheless, agreement generally remains below complete consensus, even for more restrictive Rashomon sets. Consistently across varying levels of ϵ and across tasks, the observed recourse curve lies in between the extremes of strong multiplicity and strict monoculture (see Appendix D.1 for a discussion of task differences). In particular, the observed likelihood for both extreme outcomes, no recourse and full recourse, is elevated compared to the random-error baseline even for more restrictive, smaller Rashomon sets. Unsurprisingly, we do observe, that opportunities for recourse increase as ϵ and thus the size of the Rashomon set increases. Nevertheless, we observe distribution of recourse levels differs considerably from a strict monoculture, wherein the recourse curve would collapse into a step function determined by the false negative rate of the sole remaining model.

Across varying values of ϵ and tasks, the observed recourse curve lies between the extremes of strong multiplicity and strict monoculture (see Appendix D.1 for a discussion of task differences). In particular, the likelihood of both extreme outcomes – no recourse and full recourse – is elevated compared to the random-error baseline, even for smaller Rashomon sets. As expected, opportunities for recourse increase as ϵ and thus the size of the Rashomon set grow. Nonetheless, the distribution of recourse levels differs substantially from a strict monoculture, even for more restrictive values of ϵ, reflecting some degree of diversity in model behavior within the Rashomon set.

D.3 Impact of Thresholding and Model Selection Metrics

In the main paper, we take a conservative approach: to obtain model predictions, the decision threshold t is tuned to maximize balanced accuracy, converting continuous risk scores into binary outcomes. However, when constructing the Rashomon set, models are selected based on overall accuracy rather than balanced accuracy. An alternative would be to align the selection criterion with the objective used to fit the threshold by employing the same metric for both steps. In the following sections, we report additional results obtained when from using either balanced accuracy for both threshold tuning and model selection, or overall accuracy in both steps. We note that the objective used by decision makers to convert risk scores into discrete predictions may vary across contexts and is generally unknown.

D.3.1 Thresholding and Model Selection based on Balanced Accuracy

In the main analysis, we optimize model predictions by tuning a threshold t on a validation subset of $n = 2000$ samples to maximize balanced accuracy, which is then used to convert risk scores into binary class predictions. Models are subsequently selected for inclusion in the Rashomon set based on their overall accuracy. In this section, we use *balanced accuracy* for both threshold tuning and model selection for the Rashomon set. Applying it also as the selection criterion ensures that the chosen models perform well across both classes. The selected models, along with their corresponding balanced accuracy scores, are highlighted in black in Figure 7.

Overall, the findings remain consistent (Figure 22). Across tasks, agreement rates consistently fall between two extremes: (i) strong multiplicity, as expected under random errors, and (ii) strict monoculture, in which all model predictions are identical. Likewise, the observed recourse curves clearly lie between these two extremes across tasks, with elevated likelihood of no and full recourse.

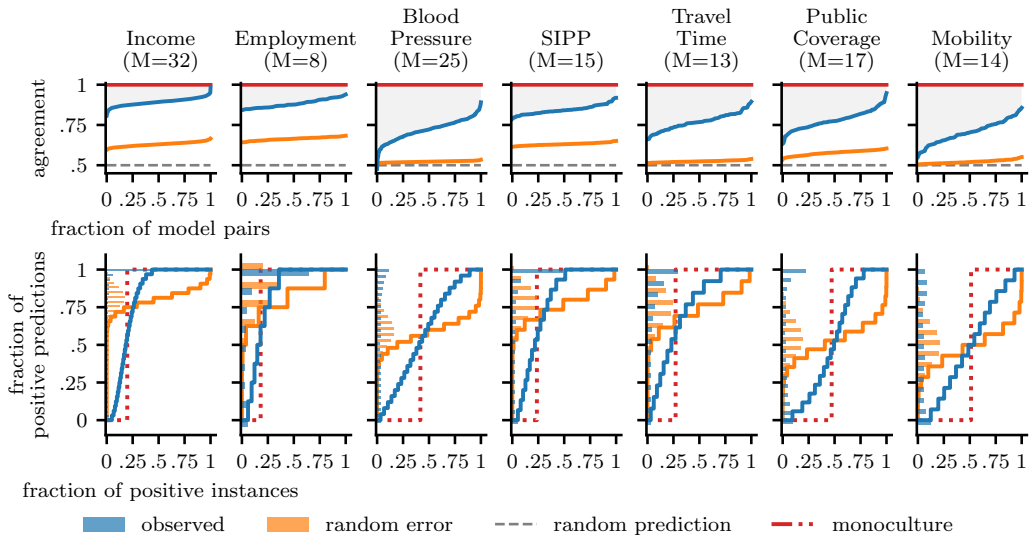

Figure 22: Agreement rates and recourse curves under zero-shot prompting when selecting models based on balanced accuracy. **Top.** Agreement curves across tasks: x% of model pairs agree on at most y% of the positive instances. Observed agreement (blue) is higher than under random errors (orange) and random predictions (dashed gray), but well below monoculture (red). **Bottom.** Recourse curves across tasks: x% of the positive instances are accepted by at most y% of the models. We zero-shot models and select those that achieve balanced accuracy within $\epsilon = 0.05$ from the best. For example, on ACSIncome we observe (blue) 20% of positive instances being accepted by at most 50% of the models. Under random errors (orange) this would rarely happen. Under strict monoculture (red, dotted) individuals only experience no or full recourse. Here, the mean TPR is used for illustration. The bar plot on the y-axis shows the density function of recourse level in the population.

Furthermore, we observe that the size of the Rashomon set increases for most tasks compared to the accuracy-based sets. This is likely because it not only includes models that are already high-performing overall, but also models that achieve more balanced performance across classes, placing more emphasis on the minority class. Despite this increase in set sizes, recourse levels remain stable across those tasks with balanced datasets (see Table 7 and Figure 23). We observe stronger differences for tasks with highly imbalanced datasets, particularly ACSMobility and ACSPublicCoverage. For both tasks, smaller fractions of individuals experience no recourse, while larger fractions experience substantial or even full recourse. On ACSMobility, discrepancy increases compared to accuracy-based selection, indicating that the Rashomon set includes models with more diverse predictive behavior. In contrast, discrepancy decreases for ACSPublicCoverage. Recall that under accuracy-based selection, a subset of models exhibits anti-correlated predictions. When selection is based on balanced accuracy, model agreement rates are consistently higher than expected under random errors, suggesting that the inductive biases of models in the empirical Rashomon set are more aligned. This alignment also results in lower discrepancy.

D.3.2 Thresholding and Model Selection based on Accuracy

In the main analysis, we optimize model predictions by tuning a threshold t on a validation subset of $n = 2000$ samples to maximize balanced accuracy, which is then used to convert risk scores into binary class predictions. Models are subsequently selected for inclusion in the Rashomon set based on their overall accuracy. In this section, we instead use the same metric, *overall accuracy*, for both threshold tuning and model selection. Overall, the results remain largely consistent, except for ACSMobility, where the empirical Rashomon set includes all evaluated models and very high agreement rates up to perfect agreement are observed. As discussed in Appendix D.1, this dataset is highly imbalanced and has low predictive signal. Consequently, most models basically behave like the constant majority class predictor, further amplified by tuning the threshold for overall accuracy. This example illustrates that the choice of metrics for thresholding and the Rashomon set should

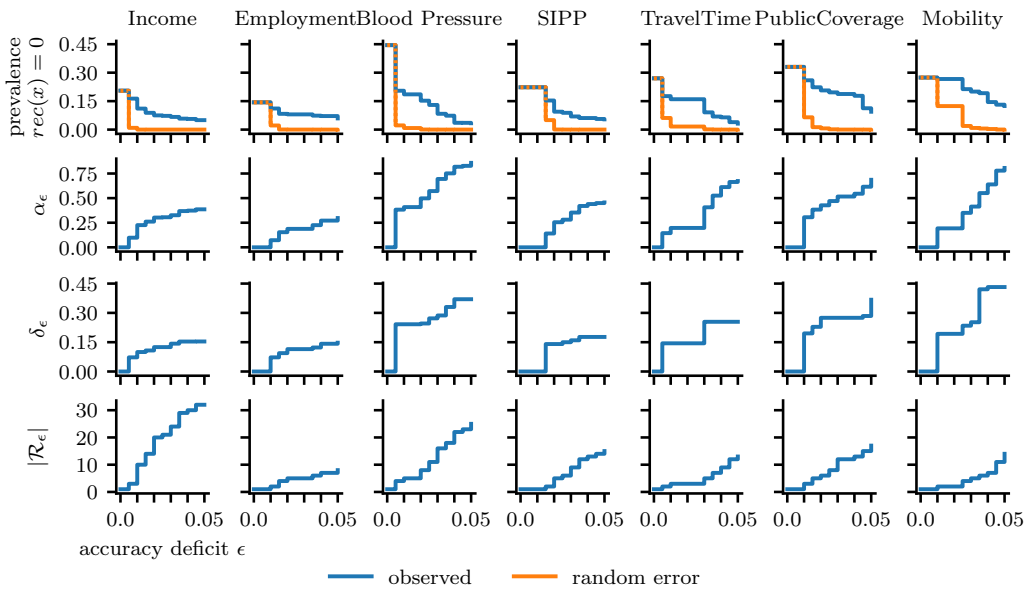

Figure 23: Severity of monoculture and predictive multiplicity as a function of accuracy deficit ϵ from the best model. Each column corresponds to one task. Predictions are obtained via 0-shot prompting. Models are selected based on balanced accuracy. **Top.** The fraction of positive instances that experience no recourse (blue) is consistently higher than what would be expected under random errors (orange). Although this gap persists, it decreases as ϵ increases. **Second Row.** Ambiguity increases with ϵ, highlighting the fraction of individuals affected by model choice. Due to monotonicity, ambiguity is likely higher for the full Rashomon set. **Third Row.** Discrepancy increases with ϵ, potentially affording opportunities for recourse for some individuals. Due to monotonicity, discrepancy is likely higher for the full Rashomon set. **Bottom.** Number of models in the Rashomon set as ϵ set increases.

Table 8: Recourse levels and measures of multiplicity for all tasks under zero-shot prompting when using overall accuracy as selection criterion for the Rashomon set.

task	$\lvert\mathcal{R}_\epsilon\rvert$	no recourse	substantial recourse	full recourse	ambiguity	discrepancy
ACSIncome	34	0.13	0.64	0.31	0.56	0.25
ACSEmployment	8	0.07	0.78	0.61	0.32	0.15
BRFSS Blood Pressure	22	0.02	0.72	0.22	0.76	0.35
SIPP	14	0.06	0.78	0.51	0.44	0.17
ACSTravelTime	15	0.08	0.49	0.10	0.82	0.37
ACSPublicCoverage	22	0.38	0.27	0.03	0.59	0.26
ACSMobility	50	0.88	0.00	0.00	0.12	0.05

be considered in a task-dependent manner. For completeness, Table 8 reports the exact fractions of individuals receiving no, substantial or full recourse as well as ambiguity and discrepancy as measured on the empirical Rashomon set.

D.4 Results for Negative Instances

In this work, we focus on positive instances, individuals with a true label of $y = 1$ who should be accepted or approved. Nevertheless, the notion of a *(*favorable outcome) varies depending on the context. For instance, in the ACSIncome dataset, a positive prediction may indicate loan eligibility, where approval is the advantageous outcome. Conversely, when the same data is used to assess eligibility for financial aid, in which case a negative prediction (reflecting lower income) would instead result in a beneficial outcome for the individual. Similarly, in tasks like fraud detection,

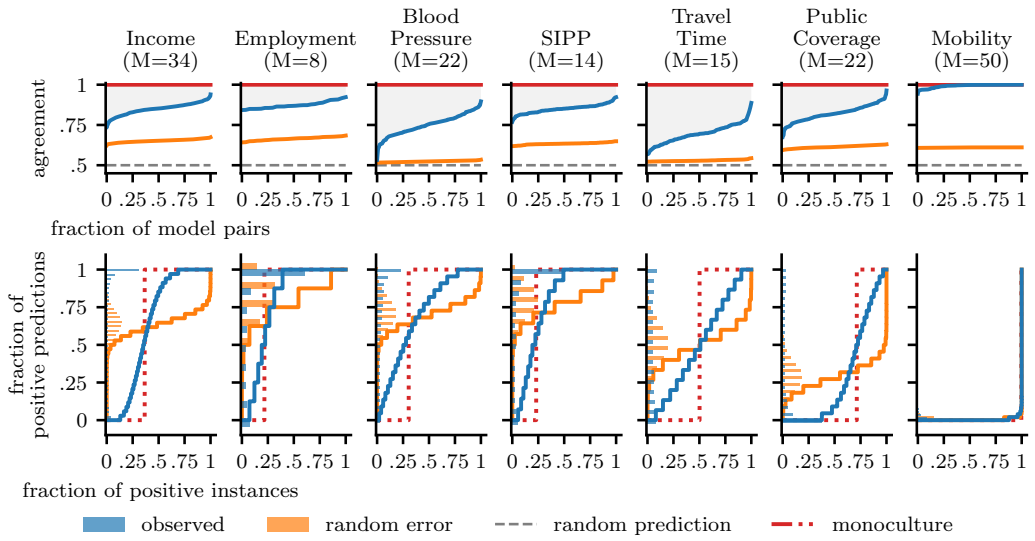

Figure 24: Agreement rates and recourse curves under zero-shot prompting when selecting models based on overall accuracy. **Top.** Agreement curves across tasks: x% of model pairs agree on at most y% of the positive instances. Observed agreement (blue) is higher than under random errors (orange) and random predictions (dashed gray), but well below monoculture (red). **Bottom.** Recourse curves across tasks: x% of the positive instances are accepted by at most y% of the models. We zero-shot models and select those that achieve balanced accuracy within $\epsilon = 0.05$ from the best. For example, on ACSIncome we observe (blue) 20% of positive instances being accepted by at most 50% of the models. Under random errors (orange) this would rarely happen. Under strict monoculture (red, dotted) individuals only experience no or full recourse. Here, the mean TPR is used for illustration. The bar plot on the y-axis shows the density function of recourse level in the population.

Table 9: Recourse levels and measures of multiplicity for negative instances. We zero-shot models and use their overall accuracy as selection criterion for the Rashomon set.

task	$\lvert\mathcal{R}_\epsilon\rvert$	no recourse	substantial recourse	full recourse	ambiguity	discrepancy
ACSIncome	27	0.08	0.76	0.52	0.40	0.14
ACSEmployment	8	0.13	0.75	0.61	0.26	0.16
BRFSS Blood Pressure	23	0.04	0.66	0.13	0.83	0.39
SIPP	16	0.07	0.76	0.45	0.48	0.20
ACSTravelTime	12	0.09	0.53	0.20	0.70	0.29
ACSPublicCoverage	21	0.00	0.91	0.33	0.67	0.22
ACSMobility	5	0.00	1.00	0.77	0.23	0.14

the desirable outcome corresponds to a negative prediction (i.e., being classified as not fraudulent). In such settings, it becomes more appropriate to evaluate recourse for negative instances, as these individuals represent those potentially excluded from desirable opportunities or favorable decisions.

In this section, we present results for zero-shot prompting under the assumption that the favorable outcome corresponds to a negative prediction. Consequently, our analysis focuses on negative instances. In this context, full recourse generalizes the notion of the true negative rate across models, that is, it captures the extent to which all models consistently classify an individual as belonging to the negative class. Models are included in the empirical Rashomon set if their overall accuracy lies with a margin of $\epsilon = 0.05$ of the best-performing model.

Results for negative instances mirror those for positive ones: The empirical pairwise agreement rates fall between the two extremes – substantially higher than what would be expected under strong multiplicity, yet notably lower than under full strict monoculture (Figure 25, top panel). The corresponding recourse curve reveals an overall high level of recourse, with most negative instances

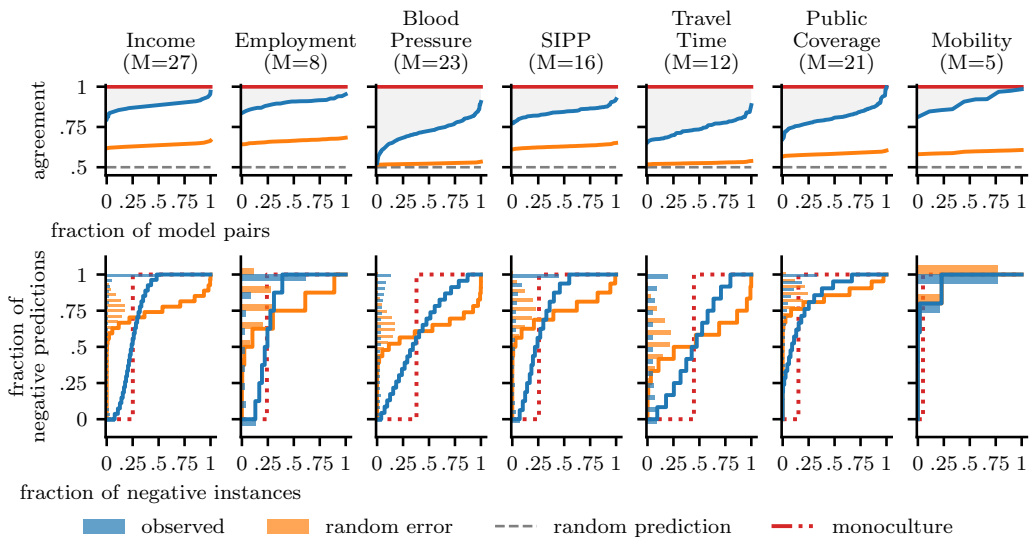

Figure 25: Agreement rates and recourse curves under zero-shot prompting for accuracy-based selection of models. Results are shown for **negative instances**. **Top.** Agreement curves across tasks: x% of model pairs agree on at most y% of the negative instances. For most tasks, observed agreement (blue) is higher than under random errors (orange) and random predictions (dashed gray), but well below monoculture (red). **Bottom.** Recourse curves across tasks: x% of the negative instances are correctly rejected by at most y% of the models. We zero-shot models and select those that achieve accuracy within$\epsilon = 0.05$ from the best.

experiencing substantial recourse potential (Figure 25, bottom panel, and Table 9). Complementary to the results for positive instances, recourse levels for ACSMobility and ACSPublicCoverage are shifted toward higher values; no recourse does not occur. Across tasks, both extremes, no recourse and full recourse, occur more frequently than expected under the random-error baseline. Nevertheless, the overall recourse curves remain situated between the poles of strict monoculture and strong multiplicity, underscoring the potential for many individuals to find recourse by turning to a different model.

E Few-Shot Prompting with Identical Prompts

We repeat our analysis using 10-shot prompting, providing each model with the same class-balanced set of examples – five labeled positive and five labeled negative. This uniform prior is chosen to ensure that models are exposed to both classes. Note, prior work has shown that language model behavior can be sensitive to the specific composition of few-shot examples [Zhao et al., 2021], which might further increase variability in model predictions (see also Section F). Consistent with the zero-shot analysis, we report results based on two selection strategies for constructing the empirical Rashomon set: one defined by overall accuracy and the other by balanced accuracy, corresponding to the criterion used to tune the decision threshold.

E.1 Impact of the number of shots

Moving from zero-shot to few-shot prompting has been found to generally improve performance; here, we examine its effects on predictive similarity among models. Since the exact number of shots is a design choice, we tested 4-, 8-, and 10-shot prompting on ACSIncome and observed consistent results (Figure 26), with a slight increase in average accuracy among models selected for the Rashomon set as the number of shots increased. For the main analysis, we focus on a larger number of shots to allow for in-context variations and ensure class balance. Beyond these considerations, the choice of the number of shots appears to have little impact on the observed patterns.

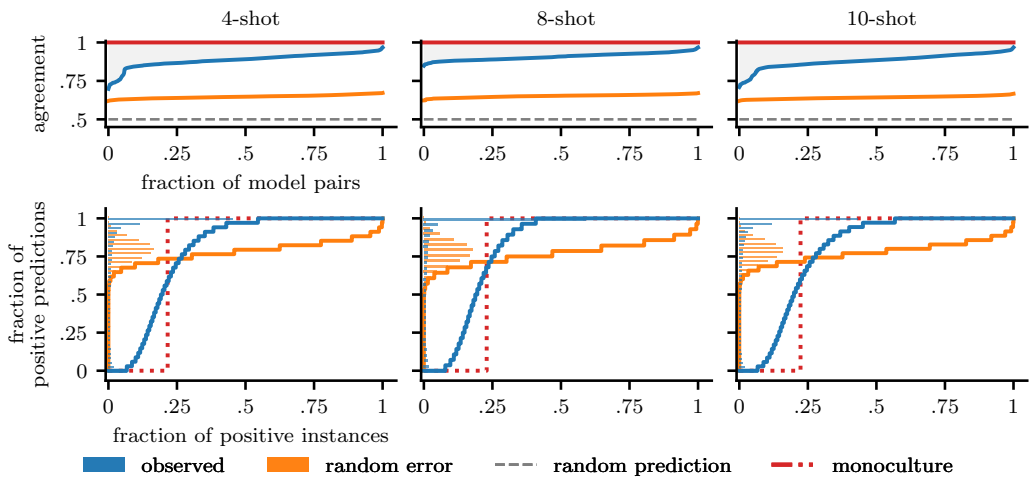

Figure 26: Agreement rates and recourse curves under few-shot prompting for varying numbers of shots. **Top.** Agreement curves across tasks: x% of model pairs agree on at most y% of the negative instances. For most tasks, observed agreement (blue) is higher than under random errors (orange) and random predictions (dashed gray), but well below monoculture (red). **Bottom.** Recourse curves across tasks: x% of the positive instances are correctly accepted by at most y% of the models. We few-shot models and select those that achieve accuracy within$\epsilon = 0.05$ from the best.

E.2 Aggregate Performance

As in zero-shot prompting, aggregate performance under 10-shot prompting varies considerably across models and tasks (Figure 27 for accuracy; Figure 28 for balanced accuracy). All thresholds were tuned to maximize balanced accuracy. We observe a tendency for additional context in few-shot prompting to benefit larger models, while also reducing performance variability among them. Effects on smaller models are mixed. Consequently, Rashomon sets are mostly comprised of large models.

At the task level, accuracy improves slightly for most models on ACSIncome and ACSEmployment. In contrast, tasks with more imbalanced label distributions – particularly ACSMobility and ACSPublicCoverage – show substantial variability in performance, with some models experiencing notable declines in accuracy under 10-shot prompting. We hypothesize that this degradation may be attributed to the use of a uniform label prior when selecting few-shot examples. This might be further amplified as thresholds are tuned to maximize balanced accuracy.

E.3 Agreement and Recourse Levels Across Tasks

E.3.1 Selection based on overall accuracy

Overall, the analysis of predictive similarity under 10-shot prompting yields results consistent with those observed in the zero-shot setting (Figure 29). Across tasks, we observe slightly higher agreement rates, though they continue to fall short of complete consensus as would be expected under strict monoculture. Notably, for ACSPublicCoverage, a small subset of models exhibits distinctly low agreement rates. Closer inspection of the Rashomon set shows that these model pairs all involve the single small model in the Rashomon set, Llama 3.2 1B, which effectively behaves like a constant majority-class predictor. As shown in Figure 30, this pattern can be addressed by choosing a more restrictive value of ϵ, without affecting the overall results.

Similarly, recourse curves under 10-shot prompting exhibit characteristics comparable to those observed under zero-shot prompting. The curves consistently lie between the two extremes of strong multiplicity and strict monoculture, while the distribution of recourse levels shows a higher likelihood of observing both extremes relative to the random-error baseline. For the tasks with class imbalance and low predictive signal, ACSPublicCoverage and ACSMobility, we observe a more varied pattern of model similarity. In particular, notably larger fractions of positive instances receive substantial

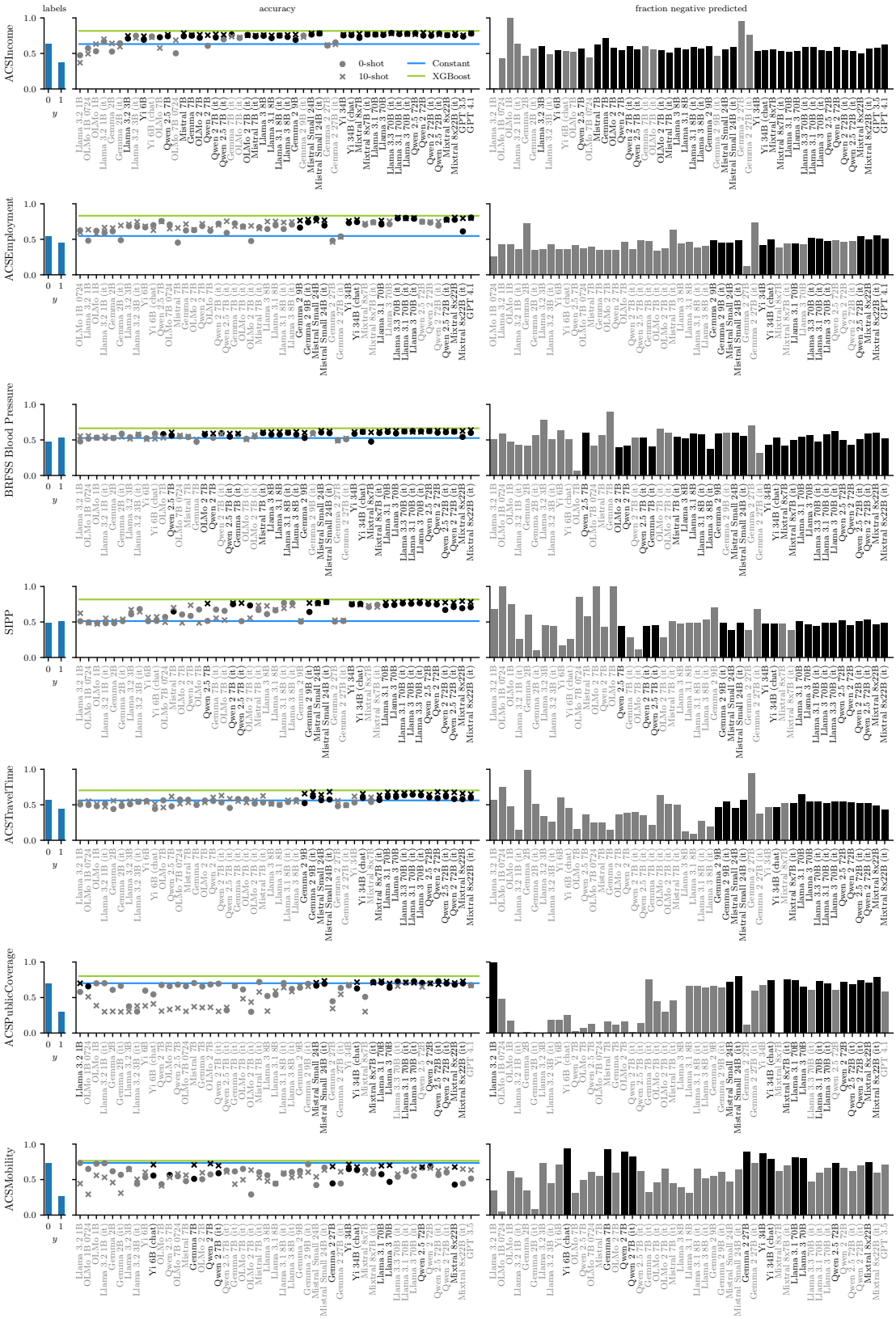

Figure 27: **10-shot performance** on the test set. Each row is a task; columns show the ground truth label distribution, **accuracy**, and the fraction of negative predictions. Models are ordered by parameter size. In the middle panel, the blue line marks the constant majority-class predictor, and the green line indicates XGBoost performance. Models included in the Rashomon set ($\epsilon = 0.05$) are shown in black; others in gray. For reference, the 0-shot performance is shown in round markers.

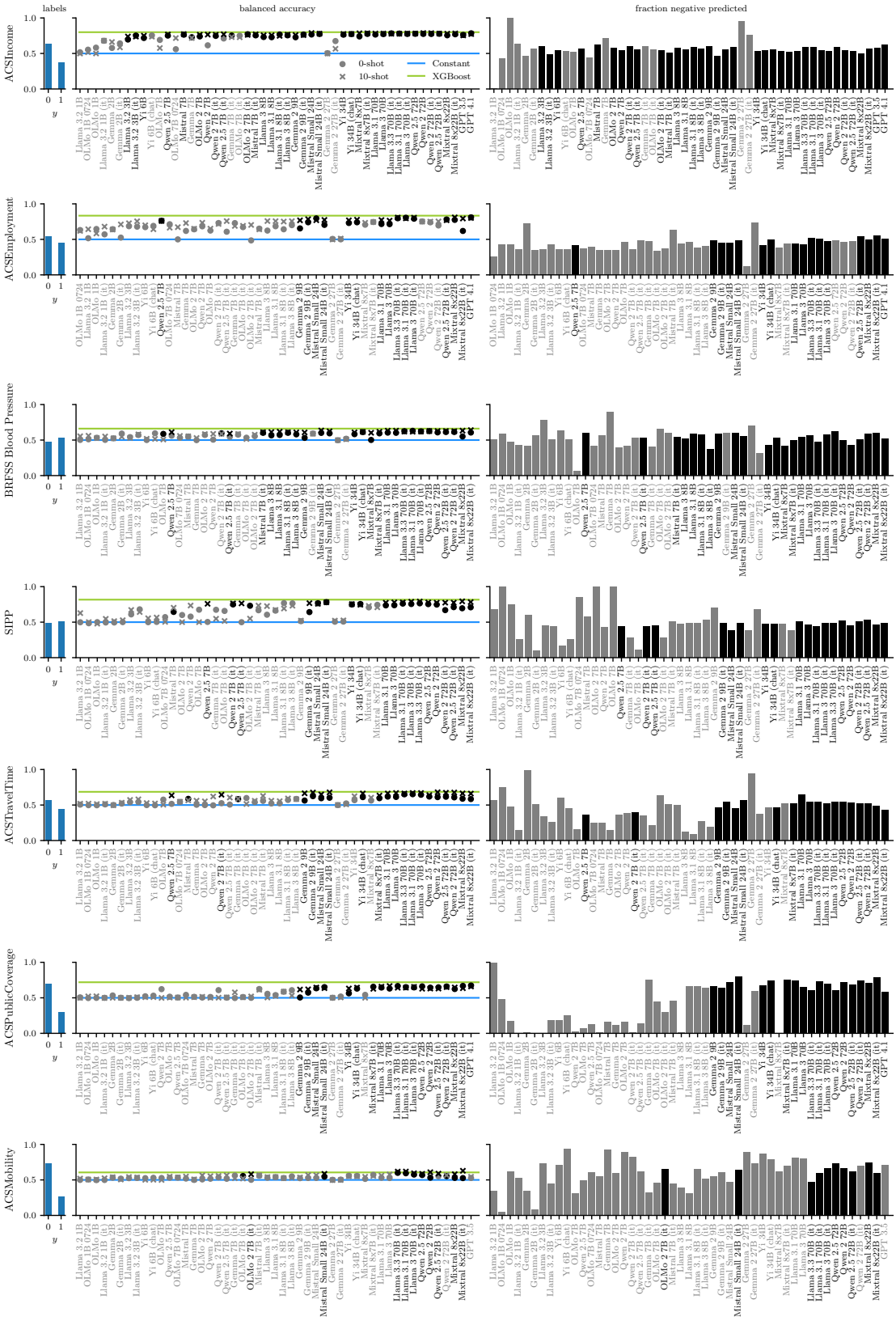

Figure 28: **10-shot performance** on the test set. Each row is a task; columns show the ground truth label distribution, **balanced accuracy**, and the fraction of negative predictions. Models are ordered by parameter size. In the middle panel, the blue line marks the constant majority-class predictor, and the green line indicates XGBoost performance. Models included in the Rashomon set ($\epsilon = 0.05$) are shown in black; others in gray. For reference, the 0-shot performance is shown in round markers.

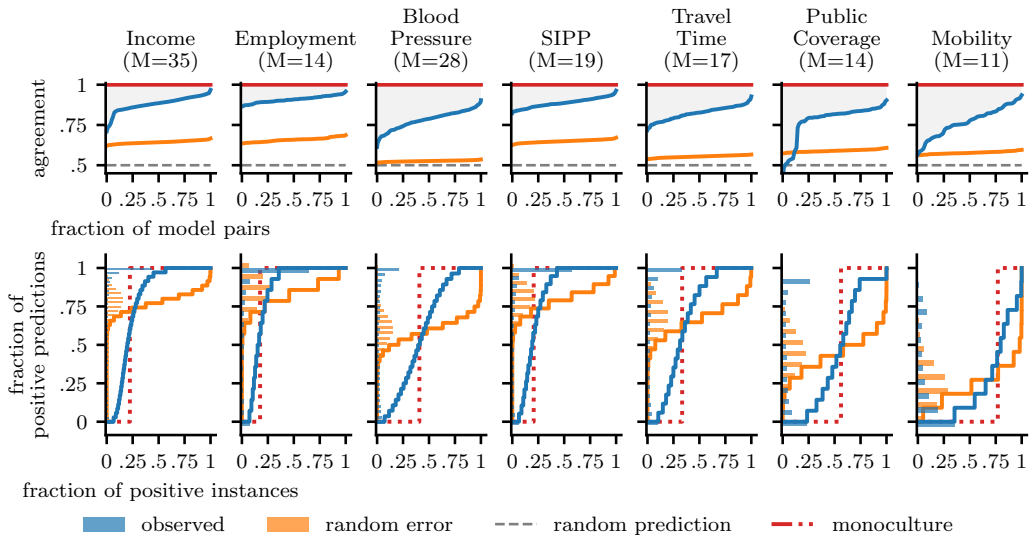

Figure 29: Agreement rates and recourse curves under 10-shot prompting for accuracy-based selection of models. **Top.** Agreement curves across tasks: x% of model pairs agree on at most y% of the positive instances. For most tasks, observed agreement (blue) is higher than under random errors (orange) and random predictions (dashed gray), but well below monoculture (red). **Bottom.** Recourse curves across tasks: x% of the positive instances are accepted by at most y% of the models. We zero-shot models and select those that achieve accuracy within$\epsilon = 0.05$ from the best. For example, on ACSIncome we observe (blue) 20of positive instances being accepted by at most 50% of the models. Under random errors (orange) this would rarely happen. Under strict monoculture (red, dotted) individuals only experience no or full recourse. Here, the mean TPR is used for illustration. The bar plot on the y-axis shows density function of recourse level in the population.

Table 10: Recourse levels and measures of multiplicity for all tasks under 10-shot prompting. Models are selected for the empirical Rashomon set based on overall accuracy.

task	$\lvert\mathcal{R}_\epsilon\rvert$	no recourse	substantial recourse	full recourse	ambiguity	discrepancy
ACSIncome	35	0.07	0.80	0.43	0.50	0.24
ACSEmployment	14	0.07	0.83	0.64	0.28	0.12
BRFSS Blood Pressure	28	0.07	0.59	0.21	0.72	0.27
SIPP	19	0.06	0.81	0.57	0.37	0.15
ACSTravelTime	17	0.10	0.69	0.33	0.56	0.25
ACSPublicCoverage	14	0.24	0.42	0.00	0.76	0.39
ACSMobility	11	0.35	0.20	0.00	0.65	0.26

recourse. Comparing the Rashomon sets for these tasks between the zero-shot and 10-shot settings reveals that larger models are more frequently included under the 10-shot condition, likely because they benefit most from the additional context (see Section E.2).

While 10-shot prompting largely mirrors zero-shot results, the data reveal a mild trend toward monoculture, as higher fractions of positive instances fall into one of the extremes for most tasks (Table 10). This is particularly notable given that the empirical Rashomon sets are larger than those identified under the zero-shot setting, making it less likely that a larger set of models will agree unless they are highly aligned. This pattern is also reflected in the overall lower discrepancy observed across Rashomon sets.

Together, these results suggest that while 10-shot prompting slightly aligns model behavior, substantial heterogeneity in model predictions remains, preserving opportunities for individuals to find recourse by switching to another model.

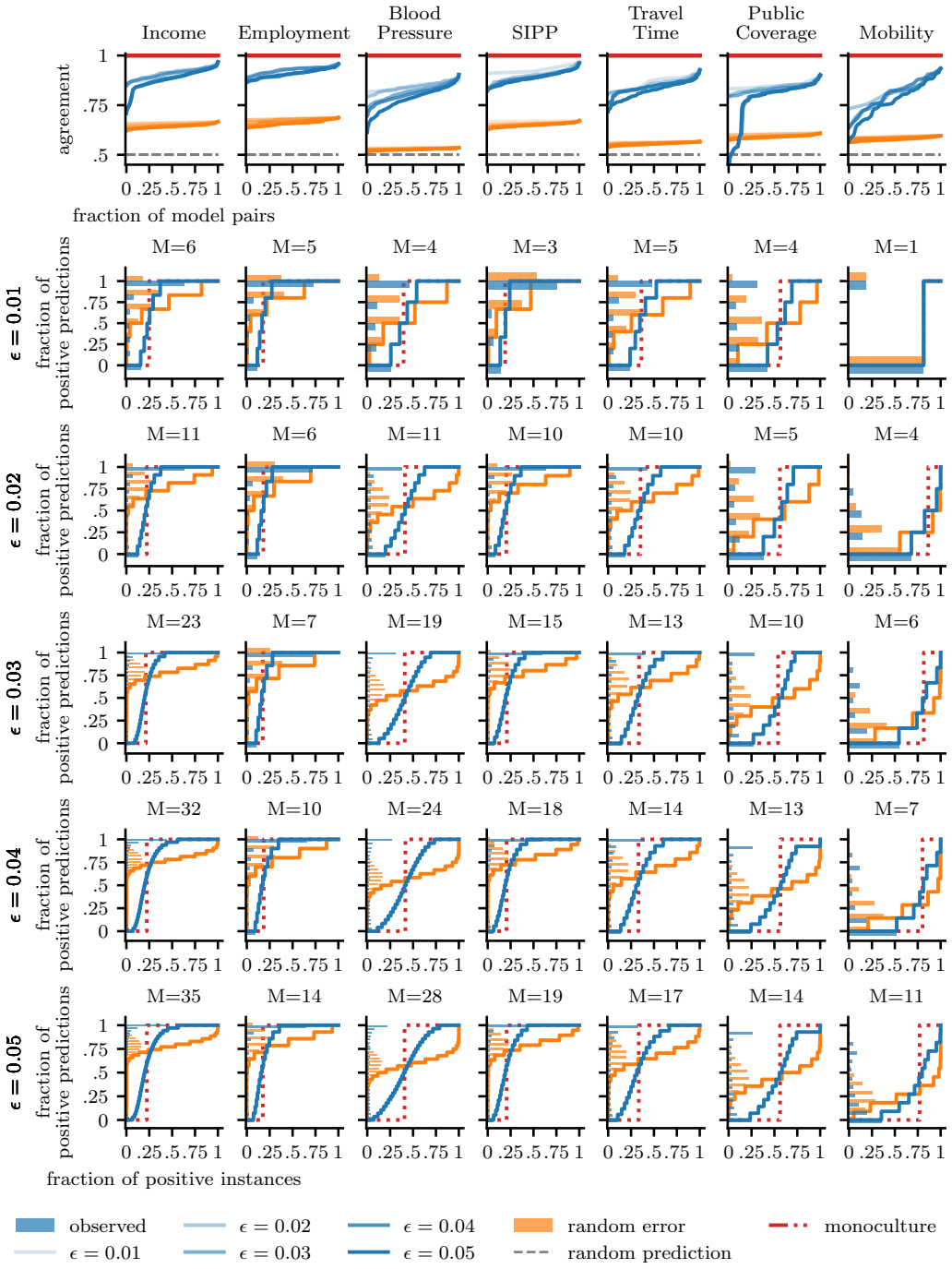

Figure 30: Agreement rates and recourse curves under 10-shot prompting for different values of ϵ. **First row.** Agreement curves: x% of model pairs agree on at most y% of the positive instances. Observed agreement (blue) consistently lies between strong multiplicity (orange) and strict monoculture (red). **Rows 2-6.** Recourse curves for varying values of ϵ: x% of the positive instances are accepted by at most y% of the models. We zero-shot models and select those that achieve accuracy within ϵ from the best. For small ϵ, the empirical Rashomon set can become restrictively small.

E.3.2 Selection based on balanced accuracy

In the main analysis, model predictions are optimized by tuning a threshold t on a validation subset of $n = 2000$ samples to maximize balanced accuracy, which is then used to convert continuous risk scores into binary class predictions. Models are subsequently selected for inclusion in the Rashomon set based on their overall accuracy. Given that few-shot examples are provided in a class-balanced manner, this section presents 10-shot prompting results when the model selection criterion is aligned with the optimization objective—that is, when the empirical Rashomon set is defined based on balanced accuracy.

Analogous to observations in the zero-shot setting, selecting models based on balanced accuracy primarily affects tasks with highly imbalanced datasets, namely ACSMobility and ACSPublicCoverage. For both tasks, we observe a distribution shift toward higher levels of recourse, with fewer individuals experiencing no recourse and larger fractions experiencing substantial or full recourse. In contrast to accuracy-based selection, agreement rates for ACSPublicCoverage are more homogeneous and higher, likely reflecting the emphasis on minority-class performance enforced by balanced accuracy. As a result, the Llama 3.2 1B model, which was included in the accuracy-based Rashomon set and effectively behaves like a constant majority-class predictor, is no longer part of the Rashomon set. While the Rashomon set size does not noticeably increase compared to accuracy-based selection in the 10-shot setting, it increases relative to the corresponding zero-shot balanced-accuracy-based Rashomon sets. Pairwise agreement rates are generally higher, and despite the larger set sizes, instances of no recourse and full recourse are more frequent, indicating a mild trend toward monoculture. Recourse curves remain situated between strong multiplicity and strict monoculture, showing that opportunities for recourse persist for a large fraction of positive instances.

Taken together, these results largely mirror those observed under zero-shot prompting. Slightly higher alignment in model predictions under 10-shot prompting suggests a mild trend toward monoculture, but substantial disagreement between models remains, preserving opportunities for individuals to find recourse by switching to another model.

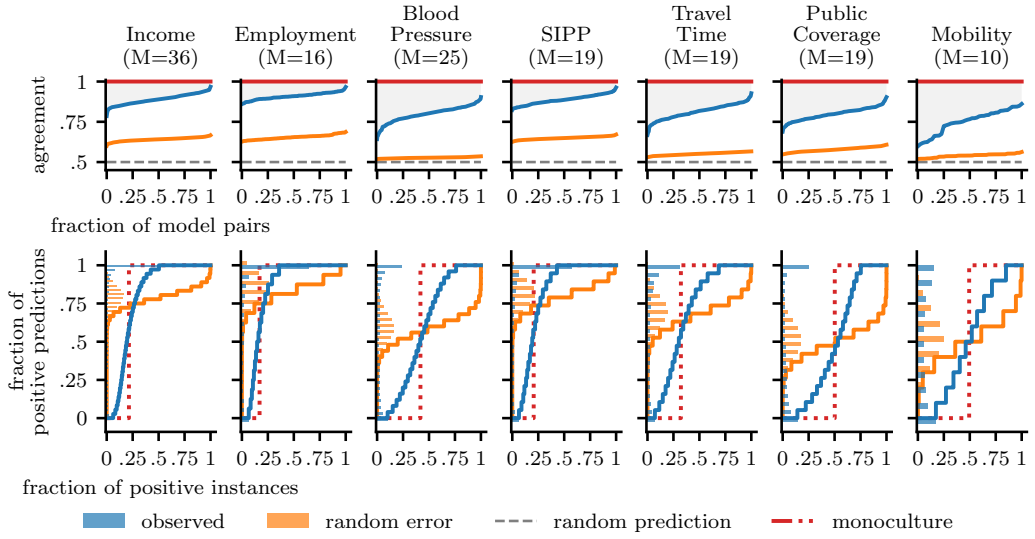

Figure 31: Agreement rates and recourse curves under 10-shot prompting for models selected by balanced accuracy. **Top.** Agreement curves across tasks: x% of model pairs agree on at most y% of the positive instances. For most tasks, observed agreement (blue) is higher than under random errors (orange) and random predictions (dashed gray), but well below monoculture (red). **Bottom.** Recourse curves across tasks: x% of the positive instances are accepted by at most y% of the models. We zero-shot models and select those that achieve balanced accuracy within $\epsilon = 0.05$ from the best. For example, on ACSIncome we observe (blue) 21of positive instances being accepted by at most 50% of the models. Under random errors (orange) this would rarely happen. Under strict monoculture (red, dotted) individuals only experience no or full recourse. Here, the mean TPR is used for illustration. The bar plot on the y-axis shows density function of recourse level in the population.

Table 11: Recourse levels and measures of multiplicity for all tasks under 10-shot prompting. Balanced accuracy is used as selection criterion for the Rashomon set.

task	$\lvert\mathcal{R}_\epsilon\rvert$	no recourse	substantial recourse	full recourse	ambiguity	discrepancy
ACSIncome	36	0.06	0.80	0.50	0.44	0.17
ACSEmployment	16	0.07	0.84	0.64	0.29	0.13
BRFSS Blood Pressure	25	0.10	0.58	0.24	0.66	0.27
SIPP	19	0.06	0.81	0.57	0.37	0.15
ACSTravelTime	19	0.07	0.70	0.31	0.62	0.23
ACSPublicCoverage	19	0.14	0.47	0.25	0.61	0.27
ACSMobility	10	0.17	0.48	0.15	0.67	0.34

F Few-Shot Prompting with Varying Prompts

Few-shot prompting introduces additional sources of variation, such as the choice, ordering, and class composition of examples. To investigate this, we repeat our analysis of prompt variations under 10-shot prompting. Consistent with prior work [Zhao et al., 2021, Lu et al., 2022, Gao et al., 2021, Schick and Schütze, 2021], we find that both the order of examples and the class composition of few-shot examples affect overall model performance (Figure 32), resulting in accuracy differences of up to three percentage points even when changing a single aspect. In addition, models remain sensitive to the same minor prompt variations tested in the zero-shot setting, exhibiting similar fluctuations in accuracy.

Since decision-makers may vary in more than one aspect of how they construct prompts, we evaluate all four models on a subsample of the variations possible for few-shot prompting. To compare effects of prompt variations and model changes directly, we fix the number of prompts and models to be the same: we randomly sample M prompt styles ($M = 35$ for ACSIncome) and evaluate agreement and recourse across 100 independent repetitions (Figure 33). We observe that pairwise agreement across models increases under prompt variations (blue), though it remains similar to the agreement observed when varying the model under identical prompting (gray line), suggesting that substantial disagreement persists in both settings. Comparing recourse curves (bottom panel), we find that prompt variation leads to a higher fraction of positive instances with no and full recourse. This suggests a mild trend toward monoculture; nevertheless, prompt variations alone still enable recourse for a considerable fraction of individuals, with the majority experiencing substantial recourse.

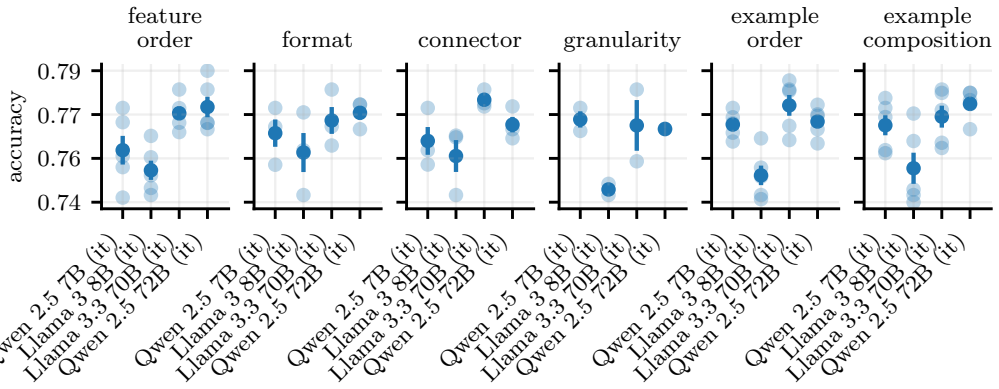

Figure 32: Minor prompt variations induce changes in accuracy of up to 3 percentage points, consistently across models on ACSIncome. Each subplot varies a single aspect of the prompt, keeping the others fixed to default. Light blue dots show accuracy for individual variations, dark blue dots indicate the mean accuracy with error bars.

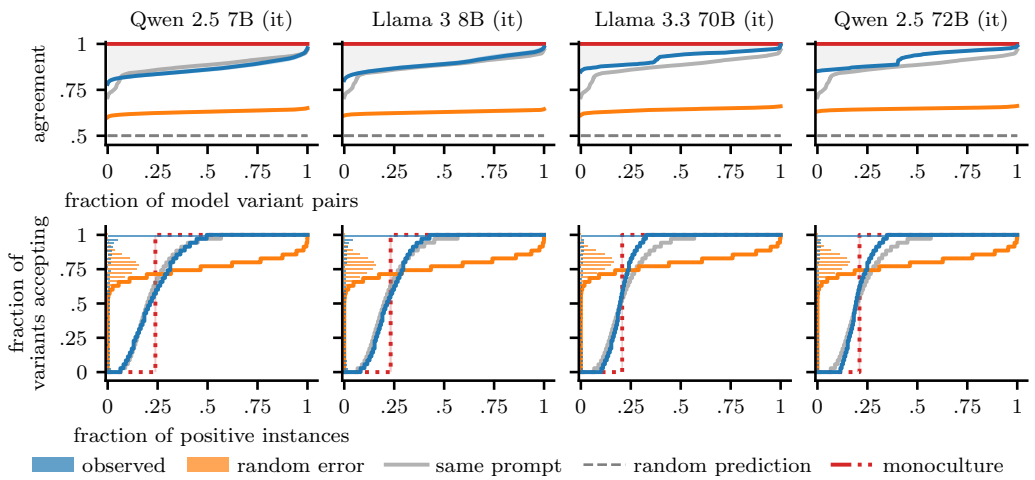

Figure 33: Agreement and recourse across prompt variations on ACSIncome. We 10-shot models with varying prompting styles, subsampling $V = 35$ prompt variations per model to match the number of different models in Figure 29. **Top.** Agreement curve: x% of prompt variation pairs agree on up to y% of the positive instances. Observed agreement rates (blue) surpass those under random errors (orange), but remain similar to agreement rates observed with identical prompting across different models (gray, solid). **Bottom.** Recourse curve: x% of the positive instances are accepted by at most y% of the prompt variations. For example, 25% of the positive instances are accepted by at most 50% of the variations. Observed recourse (blue), random errors (orange). Under strict monoculture (red, dotted) individual only experience no or full recourse. Here, the mean TPR is used for illustration. Bar plot on the y-axis shows density function of recourse level in the population.

G Prompt Variations

In this section we provide further details and examples for the prompt variations tested in this work. By default, information about an individual is presented as a bulleted list, where each feature name is followed by its value using the verb *is*. The default feature order and feature-to-text mapping are taken as provided from `folktexts`. As running example consider:

```
Information:
- age is 48 years old
- class of worker is Working for a non-profit organization
- highest educational attainment is Doctorate degree
- marital status is Married
- occupation is Education and childcare administrators
- place of birth is California
- relationship to the reference person in the survey is Brother or
sister
- usual number of hours worked per week is 45 hours
- sex is Female
- race is Asian
```

To assess how sensitive models are to subtle changes in prompting, we modify four minor aspects of prompt construction – specifically, the way information about individuals is presented. The following paragraphs detail each variation.

Feature order. We test five arbitrary orders in which features of an individual are presented - the default order given by `folktexts`, its reverse, and three random samples from the $d!$ possible orders, where d is the number of features. On ACSIncome, the resulting feature orders are:

- default: AGEP, COW, SCHL, MAR, OCCP, POBP, RELP, WKHP, SEX, RAC1P
- reverse: RAC1P, SEX, WKHP, RELP, POBP, OCCP, MAR, SCHL, COW, AGEP

- random 1: RAC1P, WKHP, AGEP, SCHL, MAR, SEX, RELP, POBP, COW, OCCP
- random 2: WKHP, OCCP, RAC1P, MAR, AGEP, RELP, SCHL, POBP, COW, SEX
- random 3: AGEP, SCHL, OCCP, MAR, COW, WKHP, RAC1P, RELP, SEX, POBP

The rows in the above example are reordered according to the given feature order.

Format. Features are presented either as *bullet* list, as *comma*-separated list (CSV style) or as simple *text* using the format '<feature name> <connector> <feature value>.'.

Information provided as comma-separated list:

```
Information:
age is 48 years old, class of worker is Working for a non-profit
organization, highest educational attainment is Doctorate degree,
marital status is Married, occupation is Education and childcare
administrators, place of birth is California, relationship to the
reference person in the survey is Brother or sister, usual number of
hours worked per week is 45 hours, sex is Female, race is Asian
```

Information provided as simple text

```
Information:
The age is 48 years old. The class of worker is Working for a non-
profit organization. The highest educational attainment is Doctorate
degree. The marital status is Married. The occupation is Education and
 childcare administrators. The place of birth is California. The
relationship to the reference person in the survey is Brother or
sister. The usual number of hours worked per week is 45 hours. The sex
 is Female. The race is Asian.
```

Connector. We vary the symbol used between <feature name> and <feature value>, choosing among '*is*', '=' or ':'.

Information provided with ':' as connector:

```
Information:
- age: 48 years old
- class of worker: Working for a non-profit organization
- highest educational attainment: Doctorate degree
- marital status: Married
- occupation: Education and childcare administrators
- place of birth: California
- relationship to the reference person in the survey: Brother or
sister
- usual number of hours worked per week: 45 hours
- sex: Female
- race: Asian
```

Information provided with '=' as connector:

```
Information:
- age = 48 years old
- class of worker = Working for a non-profit organization
- highest educational attainment = Doctorate degree
- marital status = Married
- occupation = Education and childcare administrators
- place of birth = California
- relationship to the reference person in the survey = Brother or
sister
- usual number of hours worked per week = 45 hours
- sex = Female
- race = Asian
```

Granularity. We toggle between the original feature mapping provided by `folktexts` and a lower resolution mapping. For most features, the lower-resolution version aggregates categories or bins numerical values – for example, exact age is grouped into age ranges, and detailed occupational statuses are mapped to broader occupational categories as defined in the ACS data documentation[1].

Information provided in lower resolution:

```
Information:
- age is 40-49 years old
- class of worker is Employed
- highest educational attainment is Graduate or professional degree
- marital status is Married
- occupation is Management Occupations
- place of birth is West USA
- relationship to the reference person in the survey is Siblings
- usual number of hours worked per week is 40-49 hours
- sex is Female
- race is Asian
```

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: All claims are empirically backed in Section 4 and 5.

   Guidelines:.

   - The answer NA 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 NA 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: Scope and assumptions are clearly stated, we discuss limitations in Section 6.

   Guidelines:

   - The answer NA 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: [NA]

Justification: Paper does not include theoretical results.

Guidelines:

- The answer NA 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: Steps to reproduce all experimental results are provided in the code.

   Guidelines:

   - The answer NA 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: Code and data to reproduce the main results are provided with the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 3 and supplementary material.

   Guidelines:

   - The answer NA 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: Provided in Figure 5

   Guidelines:

   - The answer NA 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: The paper includes details on computational resources used in the supplementary material in Section A.3.

   Guidelines:

   - The answer NA 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:

   Guidelines:

   - The answer NA 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:

    Guidelines: Societal impact discussed in 1.

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper is focused on evaluation and does not release any new assets.

    Guidelines:

    - The answer NA 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: [Yes]

    Justification: Original paper that produced the code package or dataset is cited.

    Guidelines:

    - The answer NA 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: [No]

    Justification: Paper does not release new assets

    Guidelines:

    - The answer NA 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: [NA]

    Justification: Paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: Paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA 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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were primarily used for editing.

Guidelines:

- The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.