DOES REASONING MAKE LANGUAGE MODELS AGREE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models face two opposing concerns: *monoculture*, in which models make the same mistakes, and *multiplicity*, in which equally accurate models make conflicting decisions. The rise of reasoning models has complicated this debate: On the one hand, reasoning may increase model agreement if models anchor on the same salient features or follow similar reasoning paths. On the other hand, stochastic reasoning traces may introduce additional branch points leading to more diverse predictions. We resolve this tension by distinguishing reasoning as an intervention from reasoning as an observed behavior. We compare agreement under controlled changes in reasoning effort with agreement across observed reasoning lengths, measuring both against the agreement expected from model accuracy. Across nine models from two developers and three classification tasks, enabling reasoning increases agreement between models evaluated on the same individuals. Yet the individuals who elicit longer reasoning traces receive less consistent predictions across models. Both patterns persist after accounting for differences in accuracy. Asking models to reason more thus promotes agreement, while observing them reason longer signals disagreement.

1 INTRODUCTION

When different models assess the same individual in consequential settings, two seemingly opposing concerns arise. Under *algorithmic monoculture*, models make the same mistakes, so using a different model may not correct an erroneous assessment of an individual (Creel & Hellman, 2022; Bommasani et al., 2022a). Under *model multiplicity*, similarly accurate models make conflicting predictions, so model choice can determine an individual's outcome (Marx et al., 2020; Black et al., 2022).

Prior work on language models finds correlated errors across models (Kim et al., 2025; Goel et al., 2025). On predictions about individuals, similarly accurate language models agree more often than an independent-error baseline would suggest, but remain well below complete agreement (Gorecki & Hardt, 2025). That study examines base and instruction-tuned models, leaving open how reasoning changes this picture. *Does more reasoning push predictions towards greater monoculture or greater multiplicity?*

Reasoning introduces additional computation before a model commits to an answer, with demonstrated gains on tasks such as mathematics and code (OpenAI et al., 2024; DeepSeek-AI et al., 2025). Its implications for agreement are less clear. Models could converge by attending to similar features or following similar reasoning patterns, consistent with evidence of correlated errors across increasingly capable models (Kim et al., 2025; Goel et al., 2025). Alternatively, stochastic intermediate generations could introduce variability. Questions eliciting longer traces have been found to receive less consistent answers from a single model (Hägele et al., 2026). These findings make both greater agreement and greater disagreement plausible, but concern different forms of consistency and do not settle what reasoning does to agreement between models.

We resolve this apparent tension by distinguishing reasoning as a treatment from reasoning length as observed behavior. Under *reasoning as an intervention*, we vary reasoning settings while holding the model and prompt fixed. Under *reasoning as behavior*, we compare individuals who elicit different trace lengths within a fixed setting. The first analysis measures the effect of enabling reasoning; the second identifies where models disagree.

This distinction yields a clear empirical picture (Figure 1): asking models to reason longer promotes agreement, while observing them reason for longer about an individual signals less consistent predictions across models. Both patterns persist after accounting for accuracy.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

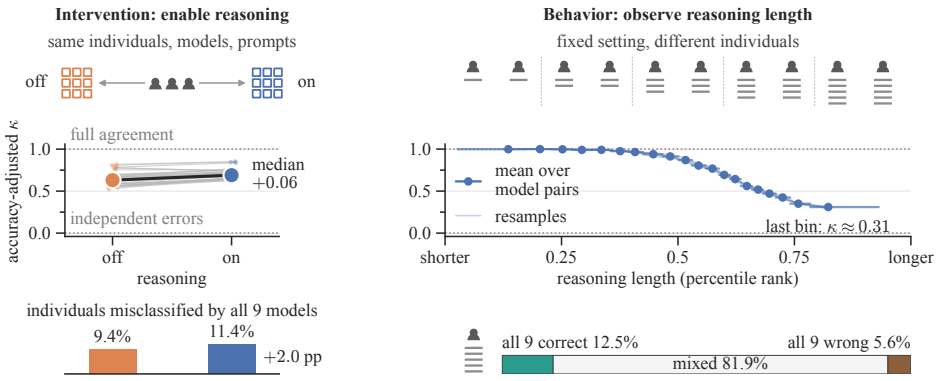


Figure 1: ACSIncome: enabling reasoning increases agreement; longer observed traces accompany less agreement.

1.1 Our Contributions

In this work, we distinguish the effect of enabling reasoning from the association between realized reasoning length and model agreement. This separates intervention effects from observed associations, without treating disagreement on longer traces as an effect of reasoning more.

Across nine models, two developers, and three prediction tasks, these comparisons reveal opposing patterns that persist after accounting for accuracy: enabling reasoning increases agreement (Figure 2), while individuals eliciting longer traces receive less consistent predictions (Figure 4). We further connect pairwise agreement to individual recourse: increased agreement concentrates outcomes at both unanimous correctness and unanimous error (Figure 3), raising the share of individuals misclassified by all nine models by 2.0–2.4 percentage points. Conversely, longer traces accompany more mixed predictions, yet unanimous errors remain more frequent than the independent-error baseline predicts (Figure 5). Thus, greater agreement does not guarantee access to a correct prediction, and greater disagreement does not eliminate shared failure.

1.2 Related Work

Monoculture and Multiplicity. Work on algorithmic monoculture examines shared errors and exclusion when decision makers rely on the same or similar models (Bommasani et al., 2022b;a), as well as aggregate welfare losses (Kleinberg & Raghavan, 2021; Kleinberg et al., 2026). Empirical studies find shared failures across commercial models (Chen et al., 2022; Toups et al., 2024) and repeated rejection of applicants assessed by a shared hiring vendor (Bommasani et al., 2026). Work on model multiplicity studies conflicting predictions among similarly accurate models (Breiman, 2001; Marx et al., 2020; Black et al., 2022; D'Amour et al., 2022; Ganesh et al., 2025). The normative implications of agreement and disagreement depend on the decision context (Creel & Hellman, 2022; Jain et al., 2024; Hedden & Raghavan, 2026; Gur-Arieh & Lee, 2025; Gur-Arieh et al., 2026; Jain et al., 2026). Our focus is empirical: how reasoning changes agreement and shared errors.

Prior work measures agreement relative to independent errors at observed accuracies, both between model pairs (Mania et al., 2019; Goel et al., 2025; Kim et al., 2025) and across model sets (Toups et al., 2024). Baselines that account for item difficulty can yield different conclusions about excess agreement (Jo et al., 2026). We use an accuracy-based baseline and discuss this limitation in Section 6.

Language models exhibit homogeneity in open-ended generation (Wu et al., 2025; Jiang et al., 2025) and correlated errors on benchmarks and simulated hiring tasks (Goel et al., 2025; Kim et al., 2025). Most directly related, Gorecki & Hardt (2025) study pairwise agreement and individual recourse for base and instruction-tuned models on tabular prediction tasks. They find agreement above an independent-error baseline but below strict monoculture. We extend this analysis to reasoning models, studying reasoning settings and variation in observed trace length.

Reasoning. Chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022) and reinforcement learning for reasoning (OpenAI et al., 2024; DeepSeek-AI et al., 2025) have significantly driven performance on tasks with verifiable answers. Beyond whether models reason, studies differ in how they control or measure reasoning:

Reasoning can be varied through effort settings or token budgets (Mei et al., 2026; Wani et al., 2026), prompt instructions (Renze & Guven, 2024; Lee et al., 2025; Nayab et al., 2026; Han et al., 2025; Gema et al., 2025), or forced continuation (Muennighoff et al., 2025). Another approach samples and aggregates multiple reasoning paths, within a model (Wang et al., 2023; Snell et al., 2025) or across models (Liu, 2026). We study provider-exposed reasoning settings without aggregating predictions across generations. Other studies examine the reasoning lengths models produce. Longer observed traces are associated with lower accuracy on competition math, even after controlling for problem difficulty (Ballon et al., 2026), and lower model confidence (Devic et al., 2025). Separately, extending reasoning can reduce accuracy on tasks designed to expose overthinking (Gema et al., 2025). Hägele et al. (2026) find that, within a fixed reasoning budget, longer observed traces are associated with greater within-model error-incoherence, while increasing reasoning budgets modestly reduces it for most tested models. Their error-incoherence metric measures the fraction of error attributable to variance across samples. The coexistence of these findings underscores the ambiguity of claims about "more reasoning". We make the treatment–behavior distinction central to the study of inter-model agreement and show how it reconciles increased agreement under a reasoning intervention with disagreement among individuals who elicit longer traces.

Trace length need not measure computational difficulty (Valmeekam et al., 2026), and stated reasoning need not faithfully explain a prediction (Turpin et al., 2023). We therefore treat length as an observed marker, without inferring reasoning mechanisms from trace content.

Another strand of work connects reasoning models to the concerns of monoculture and multiplicity more directly through output diversity and shared errors. Post-training on reasoning traces can reduce output diversity, which is not restored by suppressing chain-of-thought at inference (Karouzos et al., 2026). Across reasoning models, Liu (2026) find lower error correlations between model families than within them, alongside persistent shared errors on some tasks. These findings concern verifiable tasks or open-ended generation. We examine predictions about individuals, connecting inter-model agreement to their access to a correct prediction.

2 Measuring Agreement Beyond Accuracy

Preliminaries. We consider n individuals with features x_i and binary labels $y_i \in \{0, 1\}$, and a set of M models $h_1, \ldots, h_M$, all evaluated on the same individuals under the same reasoning setting (e.g., reasoning disabled). We call a model–reasoning-setting pair a *configuration* and evaluate each configuration repeatedly. A *resample*, indexed by s, comprises one generated response per individual for each evaluated configuration; we omit s below for notational simplicity. Each model returns a risk score, which we threshold to obtain a binary prediction $\hat{y}_{im} = h_m(x_i) \in \{0, 1\}$ (Section 3). Let $a_m = \frac{1}{n}\sum_i \mathbb{1}[\hat{y}_{im} = y_i]$ denote the accuracy of model h_m. All quantities below are computed on the individuals included in the respective analysis. For subsets such as reasoning-length bins, we recompute accuracies and both baselines using only those individuals.

Based on the model predictions, we ask two questions. Do models err on the same individuals, leaving them without recourse (monoculture)? Or do equally accurate models assign them conflicting predictions (multiplicity)? We address them at two levels: between pairs of models, and for each individual across the whole model set.

Pairwise Agreement. The observed agreement between models h_j and h_k is the fraction of individuals on which they make the same prediction,

$$p_o(h_j, h_k) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[\hat{y}_{ij} = \hat{y}_{ik}]. \tag{1}$$

For readability, we drop h and write $p_o(j, k)$ instead of $p_o(h_j, h_k)$, and likewise for p_{base}, ABC and κ. Models can agree on both correct and incorrect predictions. Observed agreement is hard to interpret on its own, because it is constrained by accuracy: In the binary setting, two models with accuracies a_j and a_k agree on at least $|a_j + a_k - 1|$ and at most $1 - |a_j - a_k|$ of individuals (Fréchet bounds). Two models with 0.8 accuracy each must therefore agree on at least 0.6 of individuals and, when all errors align, on all of them.

Accuracy-based null model. We compare observed agreement with an **accuracy-based null model** p_{base} in which models place errors independently and uniformly across individuals while preserving their accuracies:

$$p_{\text{base}}(j, k) = a_j \cdot a_k + (1 - a_j) \cdot (1 - a_k). \tag{2}$$

This baseline is widely used in the literature (Geirhos et al., 2020; Goel et al., 2025; Gorecki & Hardt, 2025). Following Gorecki & Hardt (2025), we interpret it as strong predictive multiplicity. Under this baseline, two

models with 0.8 accuracy agree on 0.68 of individuals by chance. The *agreement beyond chance* $\mathrm{ABC}(j,k) = p_o(j,k) - p_{\mathrm{base}}(j,k)$ measures excess agreement due to shared errors.

Throughout this work, we report the ***accuracy-adjusted*** κ, which scales excess by perfect agreement:

$$\kappa(h_j, h_k) = \tfrac{p_o - p_{\mathrm{base}}}{1 - p_{\mathrm{base}}}. \tag{3}$$

Unlike Cohen's κ, which uses positive prediction rates, the accuracy-adjusted κ derives expected agreement from model accuracies. It thus asks whether two models err on the same individuals more often than their accuracies imply. $\kappa = 0$ means they agree as often as independent errors would, $\kappa > 0$ that they are more aligned than the baseline, and $\kappa = 1$ that they agree on every individual. When $p_o = p_{\mathrm{base}} = 1$, the ratio is undefined and we set $\kappa = 1$ by convention in both the effort and length analyses. For binary predictions, κ equals the error consistency of Geirhos et al. (2020), which is defined on agreement in correctness rather than in predictions.

Because p_{base} depends on model accuracies, the denominator of equation 3 and the largest attainable κ vary across pairs, a known difficulty of chance-corrected agreement (Feinstein & Cicchetti, 1990; Byrt et al., 1993). We therefore also report all components p_o, p_{base} and ABC separately in the Appendix. Throughout the paper, for each model pair, we compute p_o and p_{base} for every combination of the two models' resamples, average both components over all 5×5 combinations in both analyses, and form κ from these averages. This is more stable than averaging per-resample κ values when $1 - p_{\mathrm{base}}$ is small. We exclude self-pairs.

Individual-level outcomes and recourse. Pairwise agreement does not show how errors are distributed across individuals. Following Gorecki & Hardt (2025, Def. 5), we therefore consider *recourse within the model set*: an individual's ability to obtain a favorable outcome by switching models. We operationalize a favorable outcome as a correct prediction and additionally report results conditioned on the true label. For each individual i, let

$$m_i = \textstyle\sum_{j=1}^{M} \mathbf{1}[\hat{y}_{ij} = y_i] \in \{0, \ldots, M\} \tag{4}$$

denote the number of models that predict their label correctly. The *recourse level* $\frac{m_i}{M}$ is the fraction of models that provide a correct prediction. Higher values indicate more opportunities to obtain a correct decision by switching models. Individuals with $m_i = 0$ are misclassified by every model and have no recourse, which is the *systemic failure* associated with monoculture (Bommasani et al., 2022a). Individuals with $m_i = M$ are classified correctly by every model and have full recourse. Intermediate values indicate that the outcome depends on which model is used, which is the concern raised by multiplicity.

Independent-error baseline. The same independent-error assumption yields a baseline for the recourse distribution (Bommasani et al., 2022a; Toups et al., 2024; Gorecki & Hardt, 2025). If each model h_j is correct independently with probability a_j, then m_i follows a Poisson-binomial distribution:

$$\Pr[m_i = k] = \sum_{S \subseteq [M], |S| = k} \prod_{j \in S} a_j \prod_{j \notin S} (1 - a_j), \tag{5}$$

where $[M] = \{1, \ldots, M\}$ and S indexes the models predicting correctly. For $M = 2$, the probability of a unanimous outcome, $\Pr[m_i \in \{0, 2\}]$, equals p_{base} in equation 2, so this baseline generalizes the pairwise one. We report the distribution of m_i across individuals and its *excess* over the baseline: the observed fraction at each k minus the probability in equation 5. Positive excess at $k = 0$ and $k = M$ means that unanimous outcomes (no recourse and full recourse) occur more often than independent errors at the observed accuracies would imply. Because unanimity is easier to reach with fewer models, we compare groups of equal size M evaluated on the same individuals. For each resample s, we count one prediction per model and individual, ensuring that m_i counts models rather than repeated generations. We compute the distribution and its corresponding baseline separately for each resample, then average both across resamples.

Both baselines assume that each model is equally likely to be correct on every individual. Null models that also account for item difficulty can lead to different conclusions (Jo et al., 2026). Comparisons across reasoning settings and length bins thus remain relative to this accuracy-based baseline.

3 Experimental Design

Overview. We evaluate nine language models from two developers on three binary prediction tasks: ACSIncome, ACSEmployment, and SIPP. Using the measures in Section 2, we assess pairwise agreement and individual-level

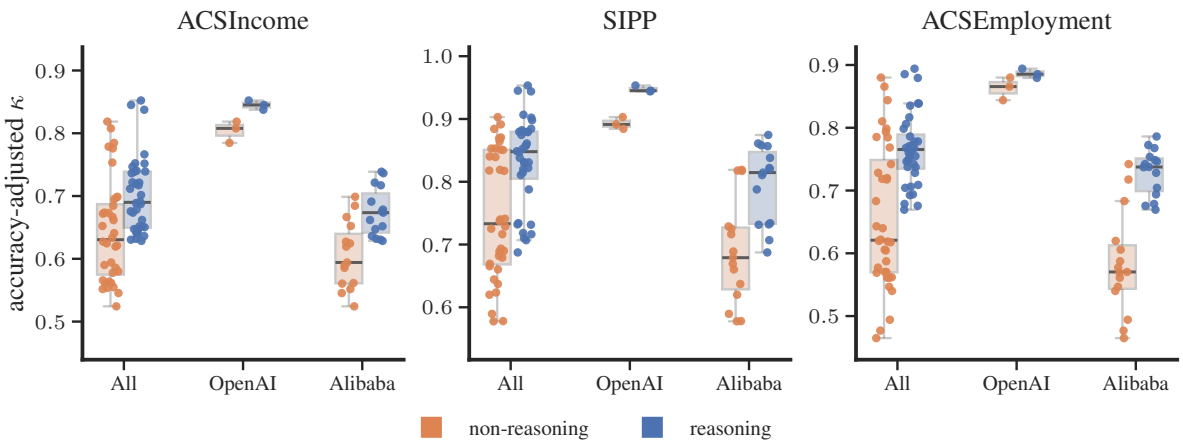


Figure 2: Accuracy-adjusted agreement κ between distinct model pairs in the same reasoning condition: disabled (orange) or enabled at the highest evaluated effort (blue). Each dot represents one pair. Observed and baseline agreement are averaged over five matched resamples before computing κ.

recourse under changes in reasoning settings (*reasoning as an intervention*) and across individuals grouped by realized reasoning length (*reasoning as behavior*).

Models and prompting. We evaluate nine reasoning language models that support disabling reasoning for matched comparisons: six from Alibaba Cloud (Qwen 3 4B, 14B, 32B and Qwen 3.5 4B, 9B and 27B) and three from OpenAI (GPT 5.1, 5.2 and 5.4). Appendix A.3 provides details. We evaluate OpenAI models at effort levels `none`, `low`, `medium`, and `high`, and Alibaba models with reasoning enabled and disabled.

We use the default numeric QA template in `folktexts` (Cruz et al., 2024) to convert tabular features of individuals into standardized, zero-shot prompts requesting a probability. We adapt the package to support reasoning models and add an answer-formatting instruction to the system prompt. Answer probabilities are extracted using regular expressions. All models receive the same zero-shot prompt, without additional instructions to control reasoning length or effort. Appendix A.3.1 provides details on prompts, reasoning controls, decoding settings, token limits, and response processing. We convert extracted probabilities into class predictions using a decision threshold t fitted separately for each model, reasoning setting, and resample to maximize balanced accuracy on a validation set of 2000 individuals, following Cruz et al. (2024) and Gorecki & Hardt (2025).

Reasoning settings and measurement. Under the intervention lens, we compare `off` versus `on` for Alibaba models and `none` versus `high` for OpenAI models. Additional results for intermediate effort levels of OpenAI models (`low` and `medium`) are reported in the appendix. These comparisons capture the effect of enabling reasoning within each model, but we note that effort levels are not directly comparable across providers. Realized reasoning length varies across individuals within each setting, motivating the behavioral analysis in Section 5. Under the behavioral lens, we count tokens within `<think></think>` blocks emitted by Alibaba models and use the reasoning-token counts reported in the API response metadata for OpenAI models. We record the available reasoning traces or summaries, token counts, responses, and extracted answers. Appendix A.3.1 provides details on answer extraction and reasoning-length measurement.

Binary prediction tasks. We evaluate three binary classification tasks from two data sources. ACSIncome and ACSEmployment predict whether an individual's annual income exceeds $50,000 and whether they are employed, respectively. Both tasks are based on the U.S. Census Bureau's American Community Survey (ACS) Public Use Microdata Sample (PUMS, U.S. Census Bureau, 2018), with task definitions from Ding et al. (2021) and natural-text feature mappings from `folktexts` (Cruz et al., 2024). We select these tasks because prior work found that base and instruction-tuned language models underperform an XGBoost baseline, leaving room for potential gains from reasoning (Gorecki & Hardt, 2025). SIPP uses the longitudinal Survey of Income and Program Participation (U.S. Census Bureau, 2014) to predict whether an individual's income is significantly above

the Official Poverty Measure (OPM). We adapt its column-to-text mappings from Gorecki & Hardt (2025). See Appendix A.2 for a detailed description of all tasks.

To limit computational costs, we randomly sample 8323 instances from ACSIncome (5% of the full dataset) and 5000 from ACSEmployment (1.5%), and use all 3972 SIPP test records (10% of the full dataset), excluding the 2000 validation individuals used for threshold fitting. All models and evaluated reasoning configurations receive the same inputs. We generate five responses per configuration and input to capture generation stochasticity, aggregating agreement across resamples as described in Section 2.

Our analysis proceeds in two stages, corresponding to the intervention and behavioral lenses. We first examine how enabling reasoning changes agreement between models and individuals' opportunities to obtain a correct prediction by switching models. We then hold reasoning settings fixed and examine how these outcomes vary across individuals who elicit shorter or longer reasoning traces.

4 Reasoning as Intervention: Agreement vs. Reasoning Effort

We first ask whether enabling reasoning increases model agreement. Reasoning settings provide a practical way to vary inference-time computation: provider documentation describes higher effort as a trade-off between response quality, latency, and token usage (Comanici et al., 2025; OpenAI, 2025c; Anthropic, 2026). We compare the same models and prompts with reasoning disabled and enabled at the highest evaluated setting. Unless stated otherwise, numerical ranges in the following results span the three tasks.

Aggregate accuracy varies considerably between models for all three tasks (Figure A1) and enabling reasoning does not consistently improve aggregate accuracy. Nevertheless, most configurations fall within an empirical Rashomon set with $\epsilon = 0.05$ (Marx et al., 2020; Semenova et al., 2022), defined in Appendix A.4. Only a few configurations, in particular the smaller, mostly 4B-parameter Qwen models with reasoning disabled, fall outside of it. We keep them to preserve the matched setting, noting that this yields broader Rashomon sets with $\epsilon \in [0.06, 0.10]$ across tasks, based on the worst-performing resample.

Pairwise agreement. Across all three tasks, enabling reasoning (`on` / `high`) shifts accuracy-adjusted pairwise agreement towards higher κ values (Figure 2). Because we account for accuracy, this increase cannot be explained by increased model performance alone. Across all model pairs, the median κ increases by 0.06–0.15. The increase is largest for pairs in which both models were developed by Alibaba Cloud (0.08–0.17, `on` vs `off`). OpenAI model pairs, which already agree strongly with reasoning disabled (median κ 0.81–0.89), gain less (0.02–0.05, `none` vs `high`). Across all pairs, κ values are also less dispersed with reasoning enabled on every task. The positive κ values throughout indicate that observed agreement across model pairs markedly exceeds the baseline agreement expected if models made random errors at their respective accuracy levels. Note that the increase in agreement is not accompanied by systematically more extreme risk-score distributions (Figure A2). These patterns are also reflected in raw observed agreement. All model pairs in the non-reasoning condition agree on at least 80% of individuals, whereas reasoning model pairs agree on at least 86% of individuals.Decomposing pairwise agreement by prediction correctness (Figure A5) shows that for Alibaba model pairs reasoning increases agreement on both correct and incorrect predictions. Results for OpenAI model pairs vary by task: on ACSIncome, agreement shifts towards incorrect predictions; on SIPP, agreement on both correct and incorrect predictions increases; on ACSEmployment, agreement on correct predictions increases while agreement on incorrect predictions decreases. Figure A4 reports all components of the accuracy-adjusted κ, namely observed agreement, baseline agreement and their difference. Among OpenAI models, which support intermediate effort levels, κ increases with effort on SIPP and ACSEmployment and peaks at `medium` on ACSIncome (Figure A3).

Individual-level agreement and shared errors. To determine whether greater agreement reflects correct predictions or shared errors, we examine individual recourse levels across all nine models in each reasoning condition (Section 2). With reasoning enabled, we observe larger fractions of individuals receiving unanimous predictions, either correct (full recourse, $m = 9$) or incorrect (no recourse, $m = 0$), while few individuals have recourse levels near 0.5, where disagreement among models is highest (Figure A6). Averaged across the five resamples, the share with no recourse rises from 9.44% to 11.41% on ACSIncome, from 11.29% to 13.62% on SIPP, and from 8.13% to 10.52% on ACSEmployment. These increases amount to approximately 2.0–2.4 percentage points, or 21–29% relative to reasoning disabled. The increase is positive in each of the five resamples on every task. Because these shares partly reflect changes in accuracy, Figure 3 shows how much they exceed

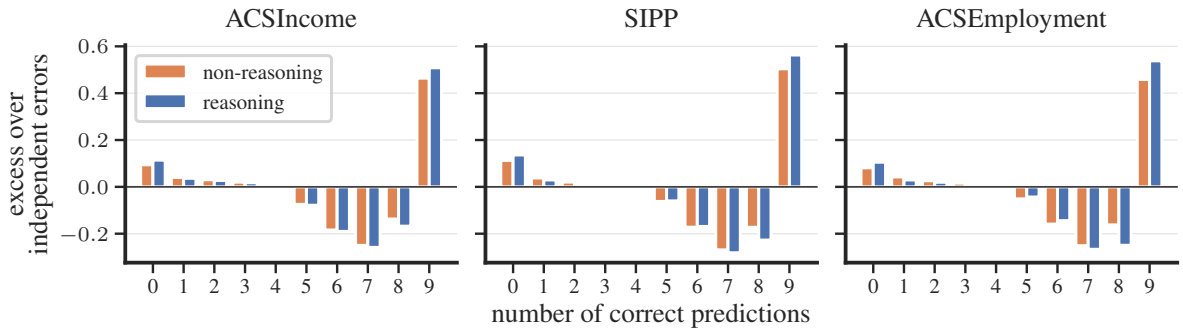


Figure 3: Excess over the independent-error baseline by number of correct models (m out of nine), with reasoning disabled (orange) or enabled (blue). Excess is the observed share minus the share expected from model accuracies under independent errors; observed distributions appear in Figure A6.

the shares expected if models erred independently at their respective accuracies. Both no and full recourse occur more often than this baseline predicts, and the excess grows when reasoning is enabled. This holds for individuals with either true label (Figure A7), so the concentration at the extremes is not driven by one outcome class. In the high-recourse regime ($m \in \{5, \ldots, 8\}$), however, where the majority of models predicts the correct label, the reasoning condition shows a more negative deviation from the baseline than the non-reasoning condition does. Thus, greater agreement concentrates outcomes at both full and no recourse, rather than uniformly improving access to a correct prediction. The exact transitions for individuals when moving from non-reasoning to reasoning are shown in Figure A8. Most individuals keep their recourse level when reasoning is enabled (67–76%); among the others, increases outnumber decreases. The most frequent change on each task is from 8 to 9 correct models. A transition from 1 to 0 is also among the most frequent changes on every task, showing that some individuals lose their sole source of a correct prediction.

Conclusions. Enabling reasoning increases both agreement and the share of individuals misclassified by all nine models. We next examine how agreement varies with the reasoning lengths individuals elicit.

5 Reasoning as Behavior: Agreement vs. Realized Reasoning Length

The preceding section examined how enabling reasoning changes agreement. We now ask: within a fixed reasoning setting, how does agreement vary across individuals who elicit different reasoning lengths? These lengths vary across individuals for both developers and overlap substantially across OpenAI effort levels (Figure A9). We use the same runs and predictions as in the preceding section, but group individuals by the lengths of the reasoning traces they elicit. Because length is observed rather than controlled, comparisons between these groups measure association, not the effect of extending reasoning.

Pairwise agreement. Across all three tasks and all model groups in Figure 4, individuals who elicit relatively longer reasoning traces receive less consistent predictions across model pairs, whereas those who elicit shorter traces experience high pairwise agreement. We group individuals by their median reasoning-length percentile rank across all resamples of all nine models. We report percentile ranks, which measure relative length within each model, since absolute token counts are not directly comparable across models. We recompute accuracies and the agreement baseline within each bin, so κ measures agreement beyond what those accuracies imply (Section 2). Bins in which all models are correct on every individual have $\kappa = 1$ by convention (Section 2). Note that observed agreement also declines (Figure A12), so the drop is not an artifact of the accuracy adjustment.

For individuals up to the median reasoning length, $\kappa \approx 1$ in nearly all bins, i.e., models agree in their prediction for almost all of them. Above the median, agreement declines towards the longest reasoning lengths, down to about 0.31 (ACSIncome), 0.34 (SIPP), and 0.39 (ACSEmployment) across all nine reasoning models in the longest bin, which is still well above the value of 0 expected under independent errors. Within the same bins, agreement tends to be lower across all nine models than among the three OpenAI models at fixed effort levels.The pooled group includes cross-developer pairs, so this difference also reflects which models are compared. Sets of models

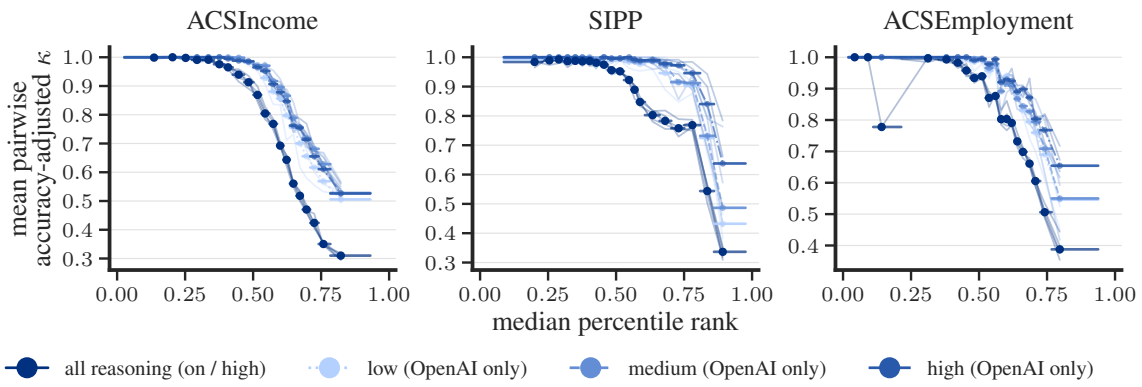


Figure 4: Accuracy-adjusted agreement by median within-model reasoning-length percentile rank across nine models and all resamples. Bins contain the same individuals across curves, allowing within-bin comparisons. Dots mark bin medians, horizontal segments span bins, and thin lines show individual resamples. Dark blue: all nine models (Alibaba `on`, OpenAI `high`); lighter blues: OpenAI models at each indicated effort.

from a single developer might be more aligned because they share much of their training data and post-training procedures. At the long-reasoning tail, pairwise agreement among OpenAI models also differs by effort level: on SIPP and ACSEmployment, κ in the longest bins is lower at `low` than at `high` effort, whereas on ACSIncome the three levels end at similar values. This matches the intervention result in Section 4, where agreement among OpenAI models increases with effort level on SIPP and ACSEmployment (Figure A3).

The decline is robust to alternative binning schemes and length measures, with one exception: It persists when bins are formed separately for each developer (Figure A10a). When each model pair bins individuals by its own reasoning lengths, the same decline appears on ACSIncome and ACSEmployment, while on SIPP the trend is flat for Alibaba model pairs, whose reasoning lengths vary little across individuals (Figure A10b). Measuring length in raw tokens or as a per-model z-score instead of percentile ranks yields the same trend (Figure A11).

Taken together, individuals who elicit longer traces therefore face greater dependence on model choice, a pattern associated with multiplicity. We next examine how this disagreement coexists with shared errors.

Individual-level agreement and shared errors. To distinguish disagreement from shared errors, we count how many of the nine models predict each individual correctly. Across all three tasks, we find that individuals who elicit short reasoning traces almost always receive the same prediction from all nine models (correct or incorrect), whereas most individuals who elicit long traces receive a mix of correct and incorrect predictions (Section 5). Within each length bin, we compare the distribution of correct-prediction counts with its Poisson-binomial baseline, computing both separately for each resample and averaging across resamples (Section 2).

Below the median reasoning length, nearly all individuals receive either unanimously correct predictions (87–94%) or unanimously incorrect predictions with no recourse by switching models (5–11%). Intermediate recourse levels are almost absent (1–3%). Above the median, the share of individuals with intermediate recourse levels grows steadily, and in the longest bin only about 10–16% of individuals receive a correct prediction from all nine models (ACSIncome: 13% / SIPP: 10% / ACSEmployment: 16%). No recourse is most common for individuals around the median length (17–19%) and less common in the longest bin (6–10%). Relative to the independent-error baseline (top row of Figure 5), no and full recourse exceed it in every bin in which models make errors. The excess at full recourse peaks close to the median length and declines towards the longest bin, while the excess at no recourse remains positive in bins where models make errors. In contrast, maximally split predictions (4 or 5 correct predictions) occur roughly as often as the baseline suggests below the median and fall further below it above the median. Figure A13 reports the excess for every recourse level, and Figure A14 shows that the pattern holds also when conditioning on the true label.

This individual-level view complements the pairwise results: Longer traces thus mark greater dependence on model choice without eliminating shared errors: even in the longest bin, all nine models misclassify 6–10% of individuals. Multiplicity and systemic failure coexist within this group.

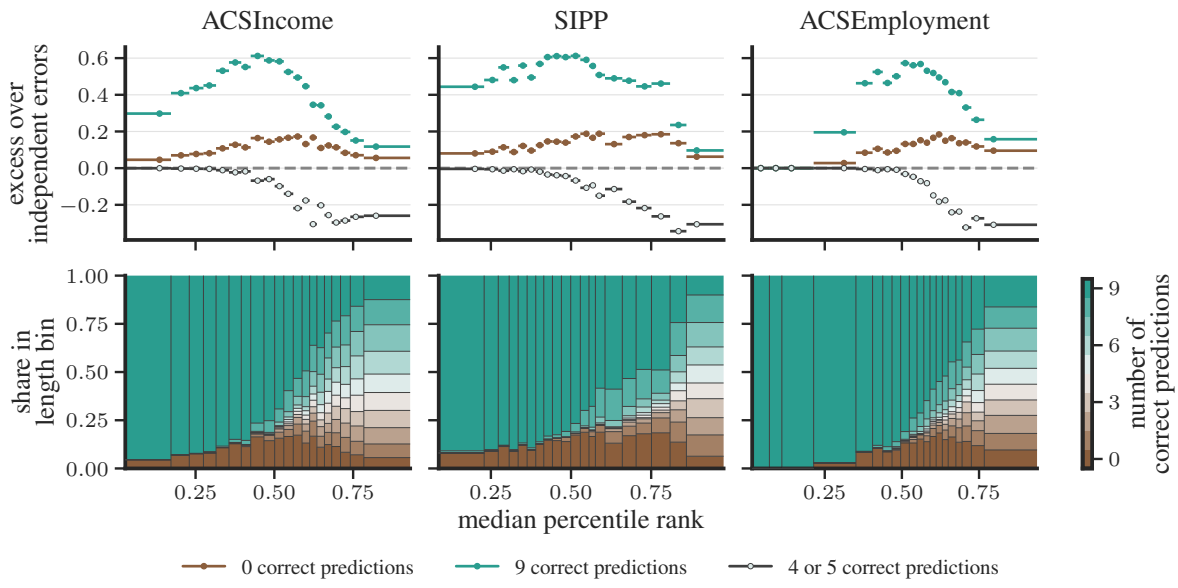


Figure 5: Recourse by reasoning-length bin, as in Figure 4, with all nine models reasoning (Alibaba `on`, OpenAI `high`). *Top:* excess over the independent-error baseline at no recourse, full recourse, and maximal split (4 or 5 correct). *Bottom:* shares by number of correct models, from 0 (brown) to 9 (green). Each distribution uses one resample per model; distributions and baselines are averaged over five resamples.

Conclusions. Longer observed traces identify individuals whose predictions depend more on model choice, even after accounting for accuracy. Yet unanimous errors remain more frequent than the independent-error baseline predicts, so greater disagreement does not eliminate shared failure.

6 Discussion

Does more reasoning push predictions towards monoculture or multiplicity? Our results show why these tendencies can coexist. Enabling reasoning increases agreement beyond an accuracy-based baseline and concentrates outcomes at both full and no recourse: the share of individuals misclassified by all nine models rises by 2.0–2.4 percentage points. Yet within fixed reasoning settings, individuals eliciting longer traces receive less consistent predictions across models. This extends prior observations of within-model inconsistency on long traces (Hägele et al., 2026) to agreement between models. The two lenses thus distinguish what enabling reasoning changes from where disagreement occurs; even among individuals eliciting long traces, unanimous errors exceed the baseline.

Our findings cover nine models from two developers and three prediction tasks. The interventions concern reasoning at inference: disabling it does not undo the effects of reasoning training on output diversity (Karouzos et al., 2026). Observed trace length is not controlled, so its association with disagreement may also reflect differences between individuals. Our baseline adjusts for accuracy but not item difficulty, which may change estimates of excess agreement (Jo et al., 2026). Difficulty is relative to the task, model and available information, but even when it explains shared errors, affected individuals still lack recourse through model choice.

For decisions about individuals, the central implication is that greater agreement does not necessarily mean greater reliability, and access to multiple models does not guarantee access to a correct prediction. Whether shared errors reflect model similarities or difficulty, affected individuals still lack recourse through model choice. Although these experiments do not measure deployment harms, they motivate evaluating who remains consistently misclassified alongside average accuracy. Future work should extend the analysis to more diverse models and decision settings, and compare increased reasoning effort with resampling and answer aggregation. Such comparisons could identify mechanisms that improve individuals' access to correct predictions. Separating within-model sampling variation from differences between models would further clarify when switching models offers a meaningful alternative.

AI USE STATEMENT

We used generative AI tools to implement methods, provide feedback on research methodology and experiments, refine and stress-test hypotheses, and interpret results. We also used these tools to create or modify scientific figures and suggest alternative data visualizations at the authors' direction, edit and debug software code, edit the paper to improve readability, and check references, summarize and analyze prior work, and identify related literature and connections between studies, using the authors' reference list as a starting point and supplementing their own literature search. We did not use generative AI to generate synthetic data sets, propose hypotheses, help develop theoretical models or the conceptual framework, or clean and reformat data sets. Formulating mathematical claims, contributing key steps to their proofs, assisting with proof writing or translation, and supporting qualitative data analysis do not apply to this work. The language models whose predictions we analyze are the object of study, not tools used to conduct it. We reviewed all AI-assisted work, carefully revised AI-assisted text, checked statements about cited work against the original papers, and reviewed and tested AI-assisted code. We take responsibility for the final content, including all AI-assisted text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the tasks and data preparation in Appendix A.2, the evaluated models in Appendix A.3, and the prompting, generation, and response-processing protocol in Appendix A.3.1. Code to reproduce the analysis and a summary of computational resource usage are provided in Appendix A.1.

REFERENCES

Anthropic. Effort. Claude Platform Documentation, 2026.

Marthe Ballon, Andres Algaba, and Vincent Ginis. The relationship between reasoning and performance in large language models—O3 (mini) thinks harder, not longer. *Scientific Reports*, 16(1):20896, May 2026. ISSN 2045-2322. doi: 10.1038/s41598-026-50923-2.

Barry Becker and Ronny Kohavi. Adult Dataset, 1996.

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, pp. 850–863, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533149.

Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? In *Advances in Neural Information Processing Systems*, volume 35, pp. 3663–3678. Curran Associates, Inc., 2022a. doi: 10.52202/068431-0265.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, et al. On the Opportunities and Risks of Foundation Models, July 2022b.

Rishi Bommasani, Sarah H. Bana, Kathleen A. Creel, Dan Jurafsky, and Percy Liang. Algorithmic Monocultures in Hiring. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, pp. 6351–6382, New York, NY, USA, June 2026. Association for Computing Machinery. ISBN 979-8-4007-2596-8. doi: 10.1145/3805689.3812400.

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

Ted Byrt, Janet Bishop, and John B. Carlin. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5): 423–429, 1993. doi: 10.1016/0895-4356(93)90018-V.

Lingjiao Chen, Zhihua Jin, Sabri Eyuboglu, Christopher Ré, Matei Zaharia, and James Zou. HAPI: A large-scale longitudinal dataset of commercial ML API predictions. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 24571–24585, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, December 2025.

Kathleen Creel and Deborah Hellman. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*, 52(1):26–43, 2022. doi: 10.1017/can.2022.3.

André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 37, 2024.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, et al. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Nature*, 645(8081): 633–638, September 2025. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-025-09422-z.

Siddartha Devic, Charlotte Peale, Arwen Bradley, Sinead Williamson, Preetum Nakkiran, and Aravind Gollakota. Trace Length is a Simple Uncertainty Signal in Reasoning Models, October 2025.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990. doi: 10.1016/0895-4356(90)90158-L.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019. ISSN 1533-7928.

Prakhar Ganesh, Afaf Taik, and Golnoosh Farnadi. Systemizing Multiplicity: The Curious Case of Arbitrariness in Machine Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, October 2025. doi: 10.1609/aies.v8i2.36609.

Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit Fraser-Taliente, Henry Sleight, Linda Petrini, et al. Inverse Scaling in Test-Time Compute. *Transactions on Machine Learning Research*, July 2025. ISSN 2835-8856.

Shashwat Goel, Joschka Strüber, Ilze Amanda Auzina, Karuna K. Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, et al. Great Models Think Alike and this Undermines AI Oversight. In *Forty-Second International Conference on Machine Learning*, June 2025.

Mila Gorecki and Moritz Hardt. Monoculture or Multiplicity: Which Is It? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025.

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, pp. 3336–3349, New York, NY, USA, June 2025. Association for Computing Machinery. ISBN 979-8-4007-1482-5. doi: 10.1145/3715275.3732215.

Shira Gur-Arieh, Angelina Wang, and Sina Fazelpour. Ambiguity Collapse by LLMs: A Taxonomy of Epistemic Risks. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26. Association for Computing Machinery, 2026. doi: 10.1145/3805689.3806469.

Alexander Hägele, Aryo Pradipta Gema, Henry Sleight, Ethan Perez, and Jascha Sohl-Dickstein. The hot mess of AI: How does misalignment scale with model intelligence and task complexity? In *The Fourteenth International Conference on Learning Representations*, 2026.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-Budget-Aware LLM Reasoning, June 2025.

Brian Hedden and Manish Raghavan. Algorithmic Monoculture and its Critics, April 2026.

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*, pp. 197–206, June 2024. doi: 10.1145/3630106.3658899.

Shomik Jain, Jack Lanchantin, Maximilian Nickel, Candace Ross, Karen Ullrich, Ashia Wilson, and Jamelle Watson-Daniels. Task-Dependent Evaluation of LLM Output Homogenization: A Taxonomy-Guided Framework, April 2026.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, et al. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Nathanael Jo, Nikhil Garg, and Manish Raghavan. The Subjectivity of Monoculture, 2026.

Constantinos Karouzos, Xingwei Tan, and Nikolaos Aletras. Where does output diversity collapse in post-training?, April 2026.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated Errors in Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30038–30066. PMLR, 2025.

Michael P. Kim and Moritz Hardt. Is your model predicting the past? In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, pp. 1–8. Association for Computing Machinery, 2023. doi: 10.1145/3617694.3623225.

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.

Robert Kleinberg, Erald Sinanaj, and Éva Tardos. Price of Anarchy of Algorithmic Monoculture, April 2026.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 22199–22213, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.

Ayeong Lee, Ethan Che, and Tianyi Peng. How Well do LLMs Compress Their Own Chain-of-Thought? A Token Complexity Approach, April 2025.

Ning Liu. LLMs as a Jury: Cross-Model Consensus Can Outperform Process Reward Models for LLM Reasoning, August 2026.

Horia Mania, John Miller, Ludwig Schmidt, Moritz Hardt, and Benjamin Recht. Model Similarity Mitigates Test Set Overuse. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Sho, and Anirudha Majumdar. Reasoning about Uncertainty: Do Reasoning Models Know When They Don't Know? In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3408–3458, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.178.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, et al. S1: Simple test-time scaling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1025.

Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. Concise Thoughts: Impact of Output Length on LLM Reasoning and Cost. *Information Sciences*, 757:123939, December 2026. ISSN 00200255. doi: 10.1016/j.ins.2026.123939.

OpenAI. GPT-5.1 instant and GPT-5.1 thinking system card addendum. Technical report, OpenAI, November 2025a.

OpenAI. Update to GPT-5 system card: GPT-5.2. Technical report, OpenAI, December 2025b.

OpenAI. Reasoning models (reasoning_effort parameter), 2025c.

OpenAI. GPT-5.4 thinking system card. Technical report, OpenAI, March 2026.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, et al. OpenAI O1 System Card, December 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026.

Matthew Renze and Erhan Guven. The Benefits of a Concise Chain of Thought on Problem-Solving in Large Language Models. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pp. 476–483, November 2024. doi: 10.1109/FLLM63129.2024.10852493.

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, pp. 1827–1858, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533232.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters for Reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

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, pp. 51178–51201, Red Hook, NY, USA, May 2024. Curran Associates Inc.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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.

Karthik Valmeekam, Vardhan Palod, Kaya Stechly, Atharva Gundawar, and Subbarao Kambhampati. Beyond Semantics: The Unreasonable Effectiveness of Reasonless Intermediate Tokens. *Transactions on Machine Learning Research*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*, 2023.

Sachin Gopal Wani, Ajay Dholakia, and David Ellison. The Reasoning Tax: Token Economics of LLM Reasoning Across Task Types and Deployment Contexts, August 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 24824–24837, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.

Fan Wu, Emily Black, and Varun Chandrasekaran. Generative Monoculture in Large Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report. Technical Report arXiv:2505.09388, Qwen Team, May 2025.

A Appendix

A.1 Experimental Details

Code and reproducibility. Code to reproduce the analysis is available at https://anonymous.4open.science/r/reasonance_-2347/. Appendix A.2 describes the datasets and evaluation samples, Appendix A.3 lists the evaluated models, and Appendix A.3.1 details the prompts, generation settings, resampling seeds, and response processing.

Compute resources and usage. Open-weight models were served locally with vLLM on a shared compute cluster, using one GPU per job; closed-weight models were queried through their providers' APIs. Table A1 summarizes resource usage for the numeric-QA runs of the six Qwen and three OpenAI models used in the reported analyses, with exploratory Claude runs listed separately. Local inference used 7,776 GPU-hours. API usage for these runs totaled approximately 652.4 million prompt tokens and 883.6 million completion tokens. Token counts are rounded to 0.1 million.

Table A1: Resource usage for the numeric-QA runs, aggregated by model group. GPU-hours cover local inference; prompt and completion counts cover API inference and are reported in millions of tokens (M). Dashes indicate quantities not reported for that inference backend. Exploratory Claude runs do not enter the reported analyses.

Model group	GPU-hours	Prompt (M)	Completion (M)
Qwen (six models)	7776	–	–
OpenAI (three models)	–	501.6	825.7
Claude (exploratory)	–	150.8	57.9
Total	7776	652.4	883.6

A.2 Tasks

We evaluate three binary classification tasks from two U.S. Census Bureau data sources: the American Community Survey (ACS) Public Use Microdata Sample (PUMS) (U.S. Census Bureau, 2018) and the Survey of Income and Program Participation (SIPP) (U.S. Census Bureau, 2014).

Evaluation samples. We randomly sample 8323 instances from ACSIncome (5% of the full dataset) and 5000 from ACSEmployment (1.5%), and use all 3972 SIPP test records (10% of the full dataset). Evaluation samples exclude the 2000 validation individuals used for threshold fitting. All models and reasoning configurations receive the same inputs.

ACS prediction tasks. We use person-level data from the 2018 ACS one-year survey, adopting the ACSIncome and ACSEmployment task definitions, feature sets, and population filters from `folktables` (Ding et al., 2021). We encode features using the natural-text mappings provided by `folktexts` (Cruz et al., 2024). Table A2 lists the features and example encodings. Prior work found a larger performance gap between language models and traditional machine learning models on these tasks (Gorecki & Hardt, 2025).

ACSIncome. ACSIncome predicts whether annual income (PINCP) exceeds $50,000 among employed U.S. residents older than 16. Its features are AGEP, COW, SCHL, MAR, OCCP, POBP, RELP, WKHP, SEX, and RAC1P. The task was proposed as a successor to the UCI Adult dataset (Becker & Kohavi, 1996), widely used in algorithmic fairness research.

ACSEmployment. ACSEmployment predicts employment status from the ESR target column for U.S. residents. Its features are AGEP, SCHL, MAR, RELP, DIS, ESP, CIT, MIG, MIL, ANC, NATIVITY, DEAR, DEYE, DREM, SEX, and RAC1P.

SIPP. SIPP is a longitudinal survey of income, employment, household composition, and program participation. Following Kim & Hardt (2023), we use Waves 1 and 2 of the 2014 panel and adopt their data cleaning and

preprocessing. The task predicts whether an individual's income is well above the Official Poverty Measure (OPM), a cash-income-based measure of poverty, using the Wave 2 OPM ratio as the basis for the binary target. We construct 50 features from Wave 1 variables, covering demographics, income, health, hardship, and program participation, and convert them to natural text using the SIPP codebook. Table A3 provides feature descriptions and example encodings.

SIPP preprocessing. We adapt the pipeline of Gorecki & Hardt (2025), retaining the data source, panel, wave split, and target definition. We correct personal income to avoid counting investment and property income twice, and exclude identifiers and the Wave 1 poverty measure from the predictors. We also preserve the codebook's living-quarter categories and revise feature descriptions for clarity. The prediction question explicitly refers to household income, consistent with the target's household income-to-poverty ratio. All reported results use this preprocessing.

Table A2: Description of column-to-text mappings for the ACS features used by ACSIncome or ACSEmployment. 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 31 years old
COW	Class of worker	class of worker is Working for a for-profit private company or organization
SCHL	Educational attainment	highest educational attainment is Bachelor's degree
MAR	Marital status	marital status is Married
OCCP	Occupation	occupation is Cashiers
POBP	Place of birth	place of birth is California
RELP	Relationship	relationship to the reference survey respondent in the survey is The reference person itself
WKHP	Work-hours per week	usual number of hours worked per week is 32 hours
SEX	Sex	sex is Female
RAC1P	Race	race is Black or African American
DIS	Disability status	disability status is No disability
ESP	Employment status of parents	employment status of parents is N/A (not own child of householder, and not child in subfamily)
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 a different house in the United States one year ago
MIL	Military service	military service status is Never served in the military
ANC	Ancestry	ancestry is Single ancestry
NATIVITY	Nativity	nativity is Native
DEAR	Hearing	hearing status is No hearing difficulty
DEYE	Vision	vision status is With vision difficulty.
DREM	Cognition	cognition status is No cognitive difficulty

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

Table A3: 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 Kim & Hardt (2023) for the full list of available variables.[1]

Column	*Description*	*Example*
demographics		
AGE	age	age is `49 years old`
GENDER	gender	gender is `Male`
RACE	race	self-identified race(s) is `White only`
EDUCATION	highest educational attainment	highest level of education completed is `High School Graduate (diploma or GED or equivalent)`
MARITAL_STATUS	marital status	marital status is `Married, spouse present`
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 `4 persons`
ORIGIN	of Spanish, Hispanic, or Latino origin	Spanish, Hispanic, or Latino origin is `No`
income		
INCOME	total annual personal income from earnings, investment and property, transfers, social insurance, and other sources	total personal income from personal earnings, investment and property, means-based transfer income, social insurance payments and other sources in given year is `$14498`
HOUSEHOLD_INC	total household income in the reference year	total annual income of all household members is `$61422`
RECEIVED_WORK_COMP	receives worker's compensation payments	receipt of worker's compensation payments is `No`
INCOME_FROM_ASSISTANCE	total annual income from transfers, social insurance, and other non-earned sources	total income from transfers, social insurance and other non-earned sources in given year is `$0`
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	receipt of 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 `0.00%`
MEDICAID_ASSISTANCE	percentage of year in which individual received assistance from MEDICAID	percentage of the year the respondent was covered by Medicaid is `0.00%`
HEALTHDISAB	has a physical, mental or other health condition limiting the amount of work they can do	work-limiting health condition status is `Without a physical, mental or other health condition`

(Continued on next page)

(Continued from previous page)

Column	*Description*	*Example*
DAYS_SICK	number of days spent in bed due to illness or injury in the last year	number of days spent in bed due to illness or injury in the last year is `10 days`
HOSPITAL_NIGHTS	number of nights in hospital in the last year	number of nights in hospital last year is `0 nights`
PRESCRIPTION_MEDS	did take any prescription medications	use of prescription medication is `Takes prescription medications`
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 on medical care	amount paid for non-premium medical out-of-pocket expenditures on medical care is `$1000`
HEALTH_HEARING	has serious difficulty hearing	hearing status is `Deaf or with hearing difficulty`
HEALTH_SEEING	serious difficulty seeing	vision status is `No vision difficulty`
HEALTH_COGNITIVE	has serious difficulty concentrating, remembering, or making decisions	cognition status is `No cognitive difficulty`
HEALTH_AMBULATORY	has serious difficulty walking or climbing stairs	ambulatory status is `No ambulatory difficulty`
HEALTH_SELF_CARE	has difficulty dressing or bathing	self-care status is `No self-care difficulty`
HEALTH_ERRANDS_DIFFICULTY	has difficulty doing errands alone	independent-living status is `No difficulty doing errands alone`
HEALTH_CORE_DISABILITY	has at least one of six core disability measures	core disability status is `With a core disability`
HEALTH_SUPPLEMENTAL_DISABILITY	answered positively to at least one core questions, three child disability questions, or two work disability questions	supplemental disability status is `With a supplemental 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 `2 visits`
VISIT_DENTIST_NUM	number of visits to a dentist or other dental professional	number of dentist visits is `0 visits`
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	receipt of transportation assistance is `No`
UNEMPLOYMENT_COMP	received unemployment compensation payments at any time during the reference year	receipt of unemployment compensation payments is `No`
UNEMPLOYMENT_COMP_AMOUNT	total annual amount of unemployment compensation payments	amount of unemployment compensation in given year is `$0`
SOCIAL_SEC_BENEFITS	received social security benefits for themselves or on behalf of a child	receipt of social security benefits is `No`
VA_BENEFITS_AMOUNT	total annual amount of VA benefits and pension	total amount of VA benefits and pension in given year is `$0`
RETIREMENT_INCOME_AMOUNT	total annual amount of retirement income	total amount of retirement income in given year is `$0`
SURVIVOR_INCOME_AMOUNT	total annual amount of survivor benefits	total amount of survivor benefits in given year is `$0`

(Continued on next page)

(Continued from previous page)

Column	***Description***	***Example***
DISABILITY_BENEFITS_AMOUNT	total annual amount of payments due to sickness, accident or disability	total amount of disability benefits or income in given year is $0
SEVERANCE_PAY_PENSION	received any severance pay or lump sum payments from a pension or retirement plan during the reference period	receipt of severance pay or a lump sum payment from a pension or retirement plan is No
FOSTER_CHILD_CARE_AMT	amount of foster child care payments received in the reference year	amount of foster child care payments received in given year is $0
CHILD_SUPPORT_AMT	amount of child support payments received in the reference year	amount of child support payments received in given year is $0
ALIMONY_AMT	amount of alimony payments received in the reference year	amount of alimony payments received in given year is $0

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

A.3 MODELS

The reported analyses use nine models that support matched comparisons with and without reasoning. Exploratory Claude runs are included only in the resource-usage summary in Appendix A.1.

- **Qwen (6 models total)**: Qwen 3 4B, 14B, and 32B (Yang et al., 2025), and Qwen 3.5 4B, 9B, and 27B (Qwen Team, 2026), each with reasoning on/off.
- **OpenAI (3 models total)**: GPT 5.1 (OpenAI, 2025a), GPT 5.2 (OpenAI, 2025b), and GPT 5.4 (OpenAI, 2026), each at effort levels none/low/medium/high.

A.3.1 PROMPTING, GENERATION, AND RESPONSE PROCESSING

Models and reasoning controls. Inference was run between July and September 2026. We use the public instruction-tuned Qwen checkpoints and access the OpenAI models through the API. For Qwen, we enable or disable reasoning through the model's chat template. For OpenAI, we request `none`, `low`, `medium`, or `high` reasoning effort. These controls are provider-specific; we do not assume that effort levels are comparable across developers.

Prompt templates. We use zero-shot numeric-QA prompts based on `folktexts`, with the same prompt for every model and reasoning setting. The system message is

```
You are a helpful assistant.  You provide numeric probability
estimates based on the information provided.
```

followed by the answer-formatting instruction we add to `folktexts`:

```
Your response MUST end with your probability estimate in the
following format:
Probability:  X%
where X is a number between 0 and 100.
```

The user message describes the survey context, lists the individual's features in natural language, and asks the task-specific question:

> *ACSIncome:* What is the probability that this person's estimated yearly income is above $50,000 ?
> *ACSEmployment:* What is the probability that this person is an employed civilian?
> *SIPP:* What is the probability that the income of this individual's household is at least three times the poverty threshold based on the Official Poverty Measure (OPM)?

Feature mappings are given in Appendix A.2. We add no chain-of-thought instruction: reasoning is controlled through the model settings, while prompt wording is held fixed.

Generation settings. We use temperature 1.0 with reasoning enabled and greedy decoding (temperature 0.0) with reasoning disabled, to accommodate the models' supported decoding settings. Other sampling parameters retain their provider or backend defaults. The initial generation budget is 8000 tokens per response; input prompts are not truncated.

Repeated evaluation. We repeat each model–reasoning-setting configuration with different generation seeds while holding the evaluation data and prompts fixed. Qwen `on`/`off` and OpenAI `none`/`high` use five resamples; OpenAI `low`/`medium` use three. These repeats capture variability in generated responses, rather than variability from drawing new evaluation samples.

Reasoning-length measurement. For OpenAI, we use the API-reported reasoning-token count, rather than the length of the returned reasoning summary. For Qwen, we count tokens in the delimited reasoning portion of the response, including its boundary markers and excluding the final answer. Responses without a closing reasoning marker are assigned zero reasoning tokens. Counts therefore follow each provider's available representation of reasoning; relative-length analyses use within-model ranks or standardization.

Answer extraction. We extract probabilities from the final answer, excluding the reasoning trace. The extraction heuristic prioritizes the requested `Probability: X%` format, with fallbacks for other percentage or decimal expressions. When the selected format occurs more than once, we use its last occurrence to allow for revised estimates. We convert percentages to probabilities and reject values outside $[0, 1]$. Unparsable answers are treated as missing rather than assigned a default probability.

Missing responses. When no valid answer was returned because of truncation, a request failure, or an unparsable response, we sampled another response for the affected individual. This yielded valid answers for nearly all evaluated inputs. Analyses use only individuals with valid answers across all compared models and reasoning settings, so comparisons are made on the same individuals.

A.4 General Performance and Risk Scores

We assess accuracy comparability using an *empirical Rashomon set* (Breiman, 2001; Fisher et al., 2019): the evaluated configurations whose accuracy is within $\epsilon = 0.05$ (five percentage points) of the best evaluated configuration on each task. We retain configurations outside this set to preserve matched comparisons across reasoning settings.

Accuracy varies considerably across models (Figure A1). No configuration reaches the XGBoost baseline; OpenAI models come closest, in particular on ACSEmployment. Enabling reasoning improves accuracy for most Alibaba models, whereas for OpenAI models the differences between effort levels are small and not consistent in direction. Most configurations lie within $\epsilon = 0.05$ of the best configuration in each task; the exceptions, based on the plotted means, are Qwen 3 4B and Qwen 3.5 4B on ACSIncome, Qwen 3 4B on SIPP, and Qwen 3 14B and Qwen 3.5 4B on ACSEmployment, all with reasoning disabled.

Risk-score distributions differ more across tasks than between reasoning settings (Figure A2). On SIPP, scores concentrate near 0 and 1 in both conditions; on the ACS tasks, models use a wider range of scores, often concentrated at a few values. Reasoning changes these distributions without a uniform shift toward the extremes. The plots therefore do not suggest a general increase in score extremity that would by itself explain the higher agreement.

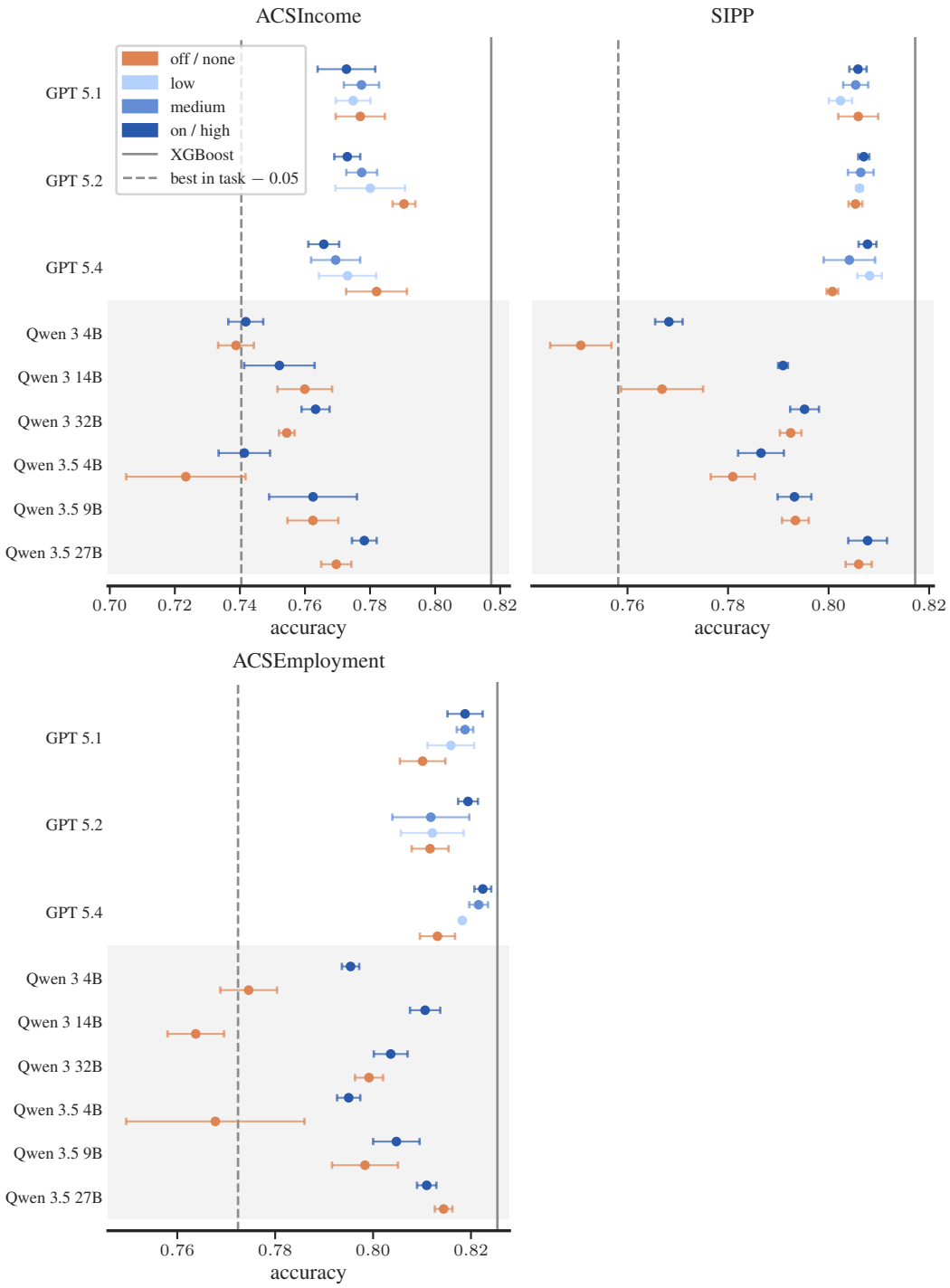
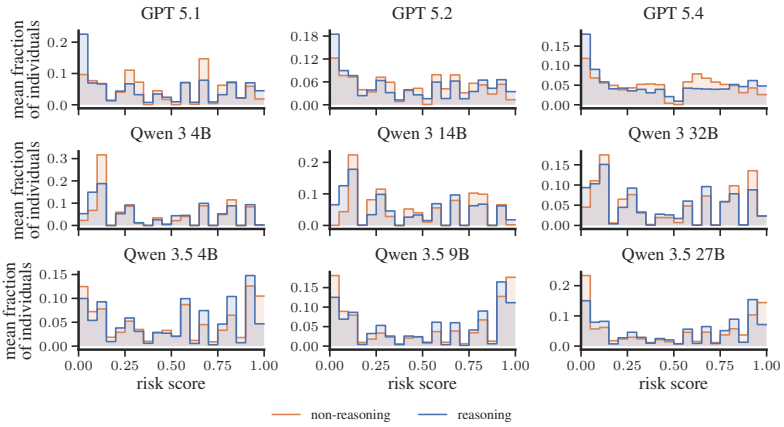
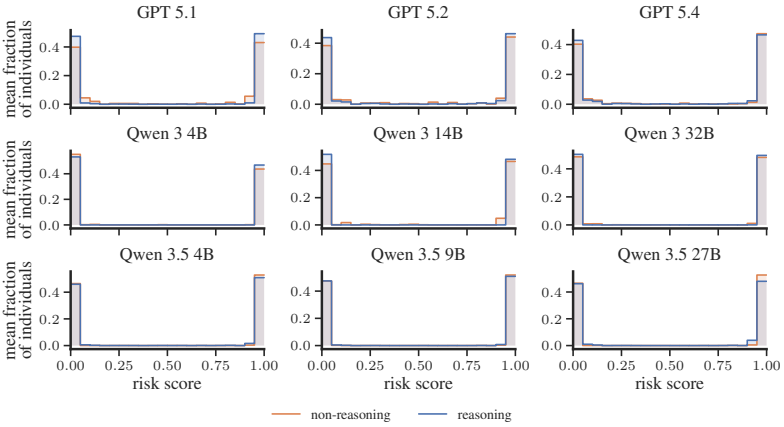


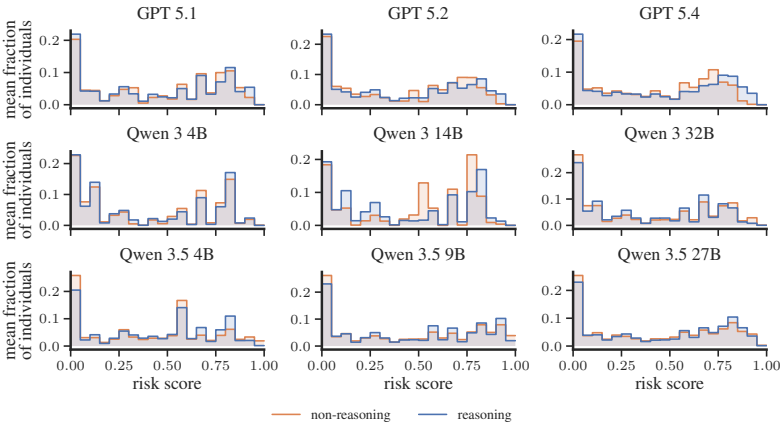
Figure A1: Accuracy of each model and reasoning setting per task. Dots show the mean over resamples and error bars the spread across resamples; the solid line is an XGBoost baseline and the dashed line the Rashomon-set cutoff (best configuration minus $\epsilon = 0.05$).



(a) ACSIncome



(b) SIPP



(c) ACSEmployment

Figure A2: Distribution of risk scores per model with reasoning disabled (orange) and enabled (blue), pooled over resamples.

A.5 Reasoning as Intervention: Additional Results

These analyses examine how enabling reasoning changes shared predictions. We first separate observed agreement from its accuracy-based baseline, then distinguish shared correct predictions from shared errors and examine the resulting changes in individual recourse.

A.5.1 κ for Varying Reasoning Levels

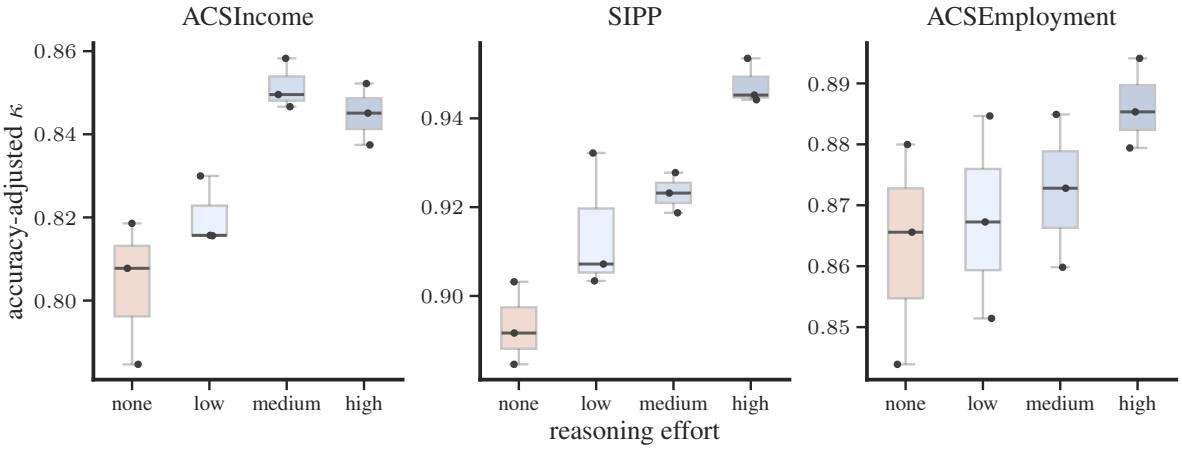


Figure A3: Accuracy-adjusted κ between pairs of different OpenAI models at each reasoning effort level.

Figure A3 examines whether agreement changes gradually with requested reasoning effort. Among the three OpenAI model pairs, agreement generally rises from `none` to `high` on SIPP and ACSEmployment. On ACSIncome, it peaks at `medium` before declining slightly. Thus, enabling reasoning and increasing it further need not have identical effects, and the pattern varies by task.

A.5.2 κ Components

Figure A4 separates changes in observed agreement from changes expected under the accuracy-based baseline. Reasoning raises observed agreement for most model pairs, while the baseline shifts less and in a task-dependent direction. Agreement beyond the baseline and the normalized measure κ generally increase, particularly for Alibaba pairs. The higher κ therefore reflects changes in joint predictions beyond those accounted for by individual model accuracies.

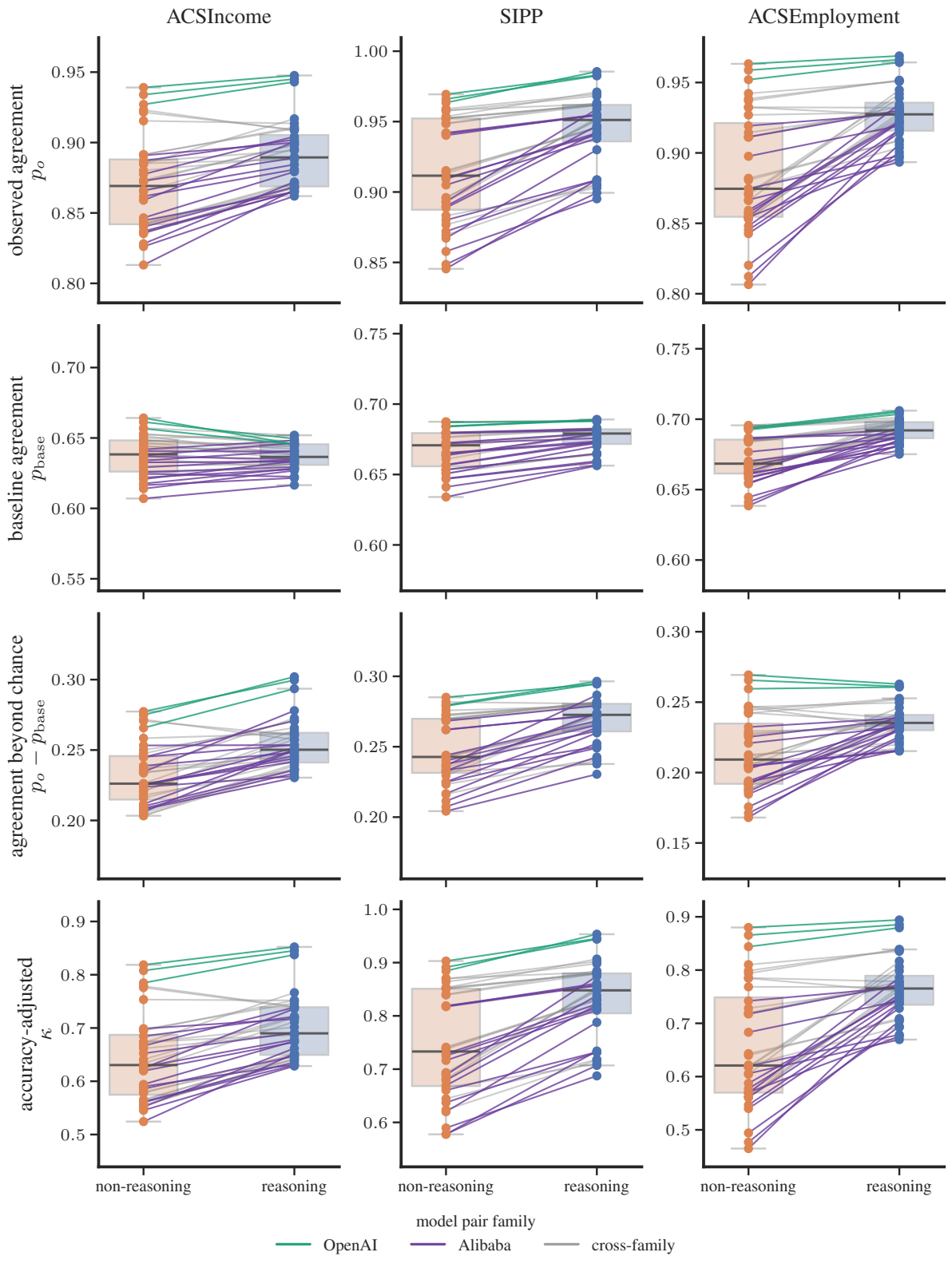


Figure A4: Components of accuracy-adjusted κ with reasoning disabled (orange) and enabled (blue): observed agreement p_o, baseline agreement p_{base}, their difference, and κ. Lines connect the same model pair (green: OpenAI, purple: Alibaba, gray: cross-developer). Pairs are formed only between configurations with the same reasoning condition (both disabled or both enabled).

A.5.3 Pairwise Change in Agreement

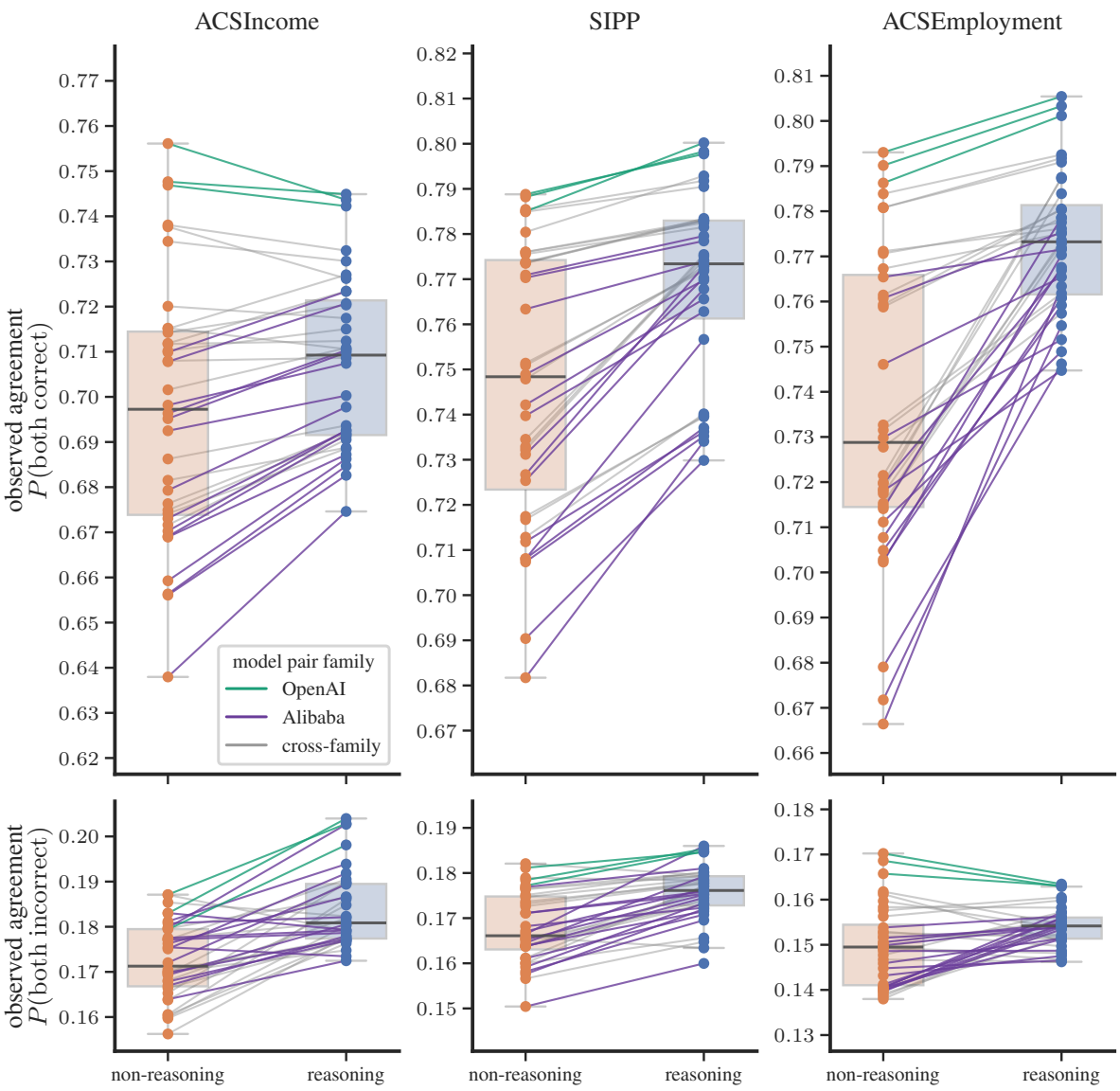


Figure A5: Observed pairwise agreement on correct predictions (*top*) and on incorrect predictions (*bottom*), with reasoning disabled (orange) and enabled (blue). Dots give the share of individuals for whom both models are correct (top) or both are incorrect (bottom). Lines connect the same model pair (green: OpenAI; purple: Alibaba; gray: cross-developer). Axis limits vary across panels, so visual slopes are not directly comparable. Pairs are formed only between configurations with the same reasoning condition (both disabled or both enabled).

Figure A5 separates agreement into both models being correct and both being incorrect. For Alibaba pairs, reasoning generally increases both components. OpenAI pairs show more task-dependent changes: on ACSIncome, agreement shifts toward shared errors, whereas on ACSEmployment shared correct predictions increase and shared errors decrease. Higher aggregate agreement can therefore accompany different changes in prediction quality.

A.5.4 Recourse Levels

Figure A6 shows how many of the nine models predict each individual correctly. Most individuals receive a correct prediction from every model, and reasoning increases this share on all three tasks. It also increases the share for

whom every model is wrong, while reducing the mass at seven or eight correct models. This concentration at both extremes explains why greater agreement can coexist with less opportunity to obtain a correct prediction by switching models.

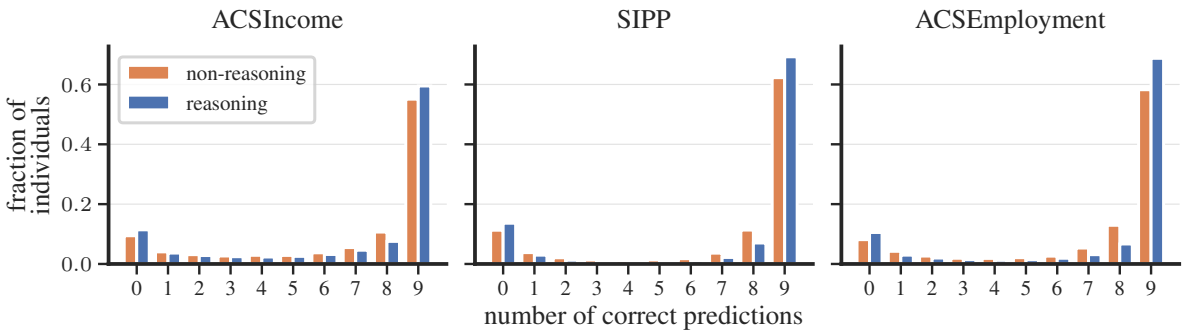


Figure A6: Distribution of recourse levels with reasoning disabled (orange) and enabled (blue): share of individuals by the number of models (out of $M = 9$) that predict their label correctly, from 0 (no recourse) to 9 (full recourse). Each distribution counts the nine models within one reasoning condition (all disabled or all enabled).

A.5.5 Recourse Levels by Label

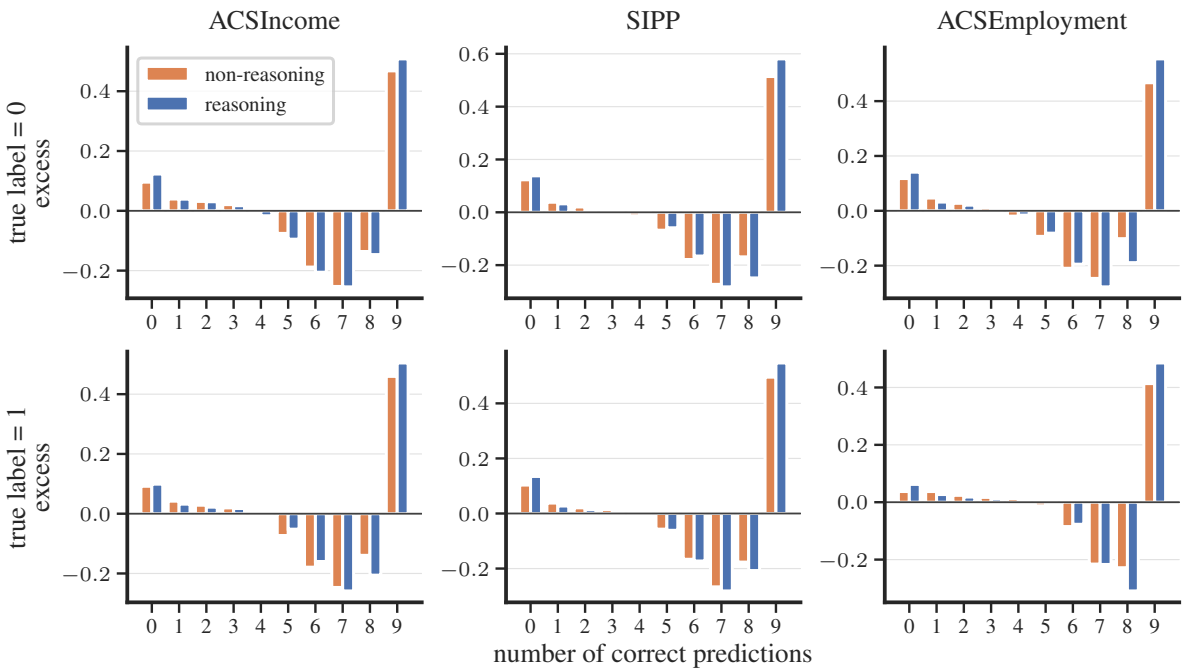


Figure A7: Excess over the independent-error baseline for each recourse level, conditioned on the true label (*top:* label 0, *bottom:* label 1), with reasoning disabled (orange) and enabled (blue). Each distribution counts the nine models within one reasoning condition (all disabled or all enabled).

Splitting the population by true label preserves the concentration at extreme recourse levels (Figure A7). For both labels, reasoning increases the excess over the plotted baseline at full recourse and, more modestly, at no recourse. Levels with most but not all models correct generally remain below the baseline. The aggregate pattern is therefore present within both outcome groups, although its magnitude differs across tasks and labels.

A.5.6 Recourse Level Transitions

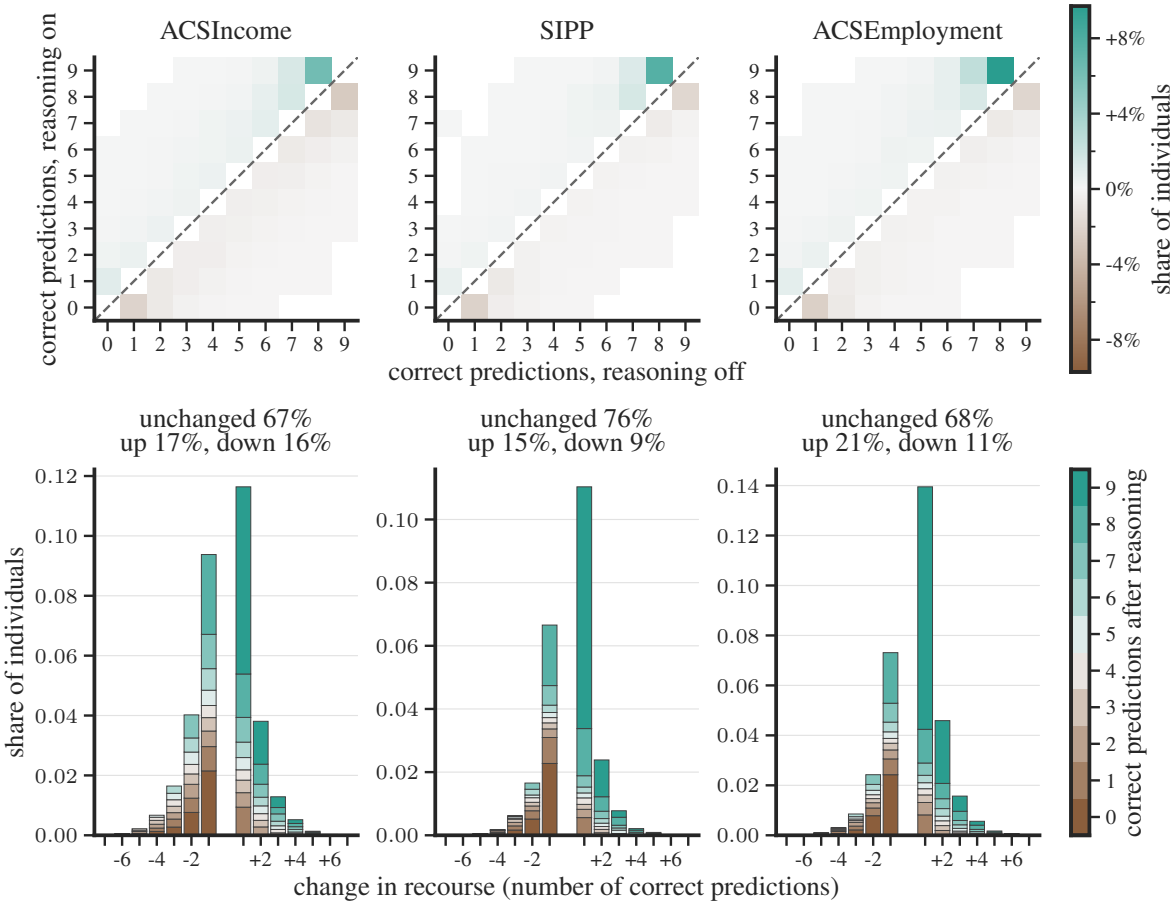


Figure A8: Change in recourse levels when reasoning is enabled. *Top:* transitions from the number of correct models with reasoning disabled (x-axis) to enabled (y-axis). Color magnitude gives the share of individuals making each transition; increases are positive (green), decreases negative (brown), and unchanged cases on the diagonal are omitted. *Bottom:* share of all individuals by the change in the number of correct models, excluding unchanged cases and stacked by the number correct with reasoning enabled.

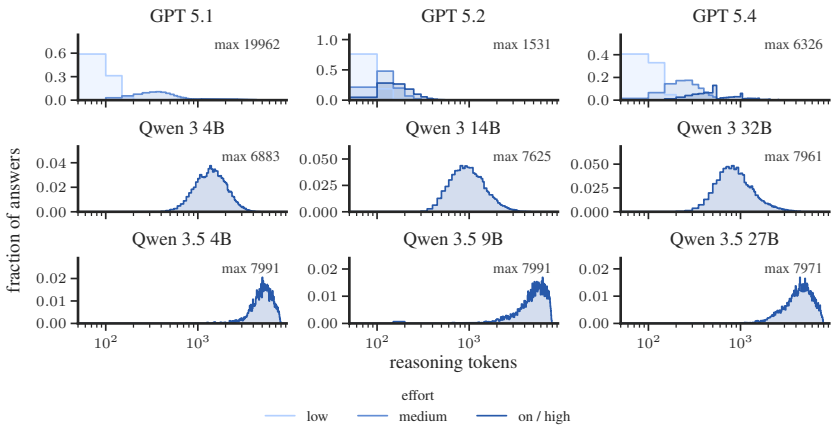
Most individuals retain the same recourse level when reasoning is enabled: 67% on ACSIncome, 76% on SIPP, and 68% on ACSEmployment (Figure A8). Increases are more common than decreases, and changes are usually by one or two correct models. The prominent transition from eight to nine correct models contributes to higher full recourse, while transitions from one correct model to none show how some individuals lose their only correct prediction.

A.6 Reasoning as Behavior: Additional Results

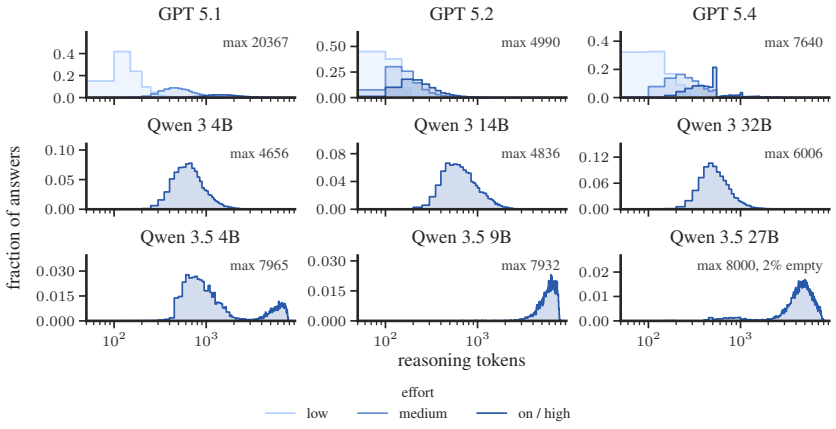
These analyses relate realized reasoning length to agreement and recourse within fixed reasoning settings. We check whether the patterns persist across model families, model pairs, and alternative length measures. These comparisons describe associations; they do not isolate the effect of making a model reason longer.

A.6.1 Reasoning Length by Effort

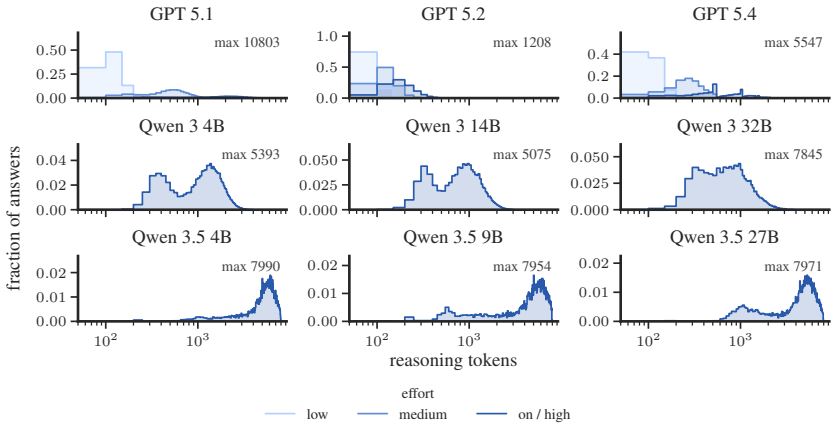
Figure A9 shows that requested effort and realized reasoning length are related but distinct. Higher OpenAI effort settings generally produce longer traces, with substantial overlap between settings. Lengths also vary markedly across models: Qwen 3.5 often generates several thousand tokens, while many OpenAI responses are much shorter.



(a) ACSIncome



(b) SIPP

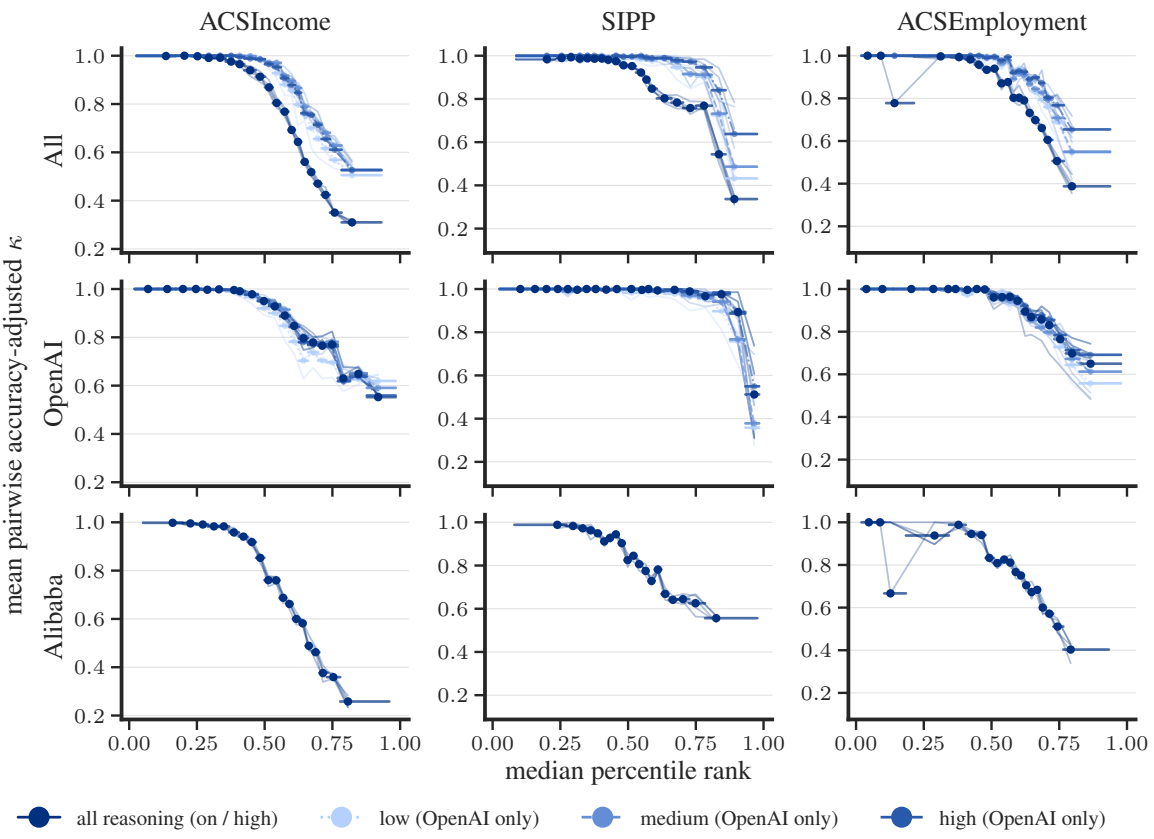


(c) ACSEmployment

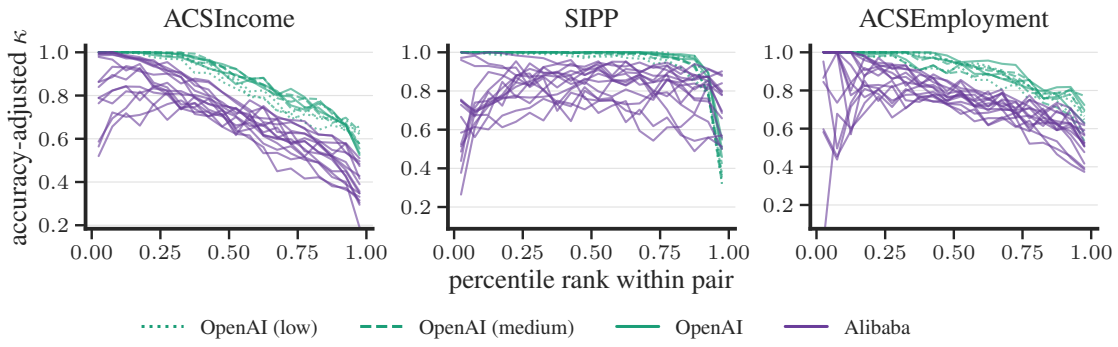
Figure A9: Distribution of reasoning tokens per answer for each model and enabled reasoning effort (logarithmic x-axis). The y-axis gives the fraction of answers in each token bin; colors denote OpenAI `low`, `medium`, and `high` effort or Alibaba `on`. Numbers give the largest observed length for each model and task.

These differences motivate comparing relative length within models rather than treating a fixed token count as equivalent across models.

A.6.2 Agreement and Reasoning Length by Developer



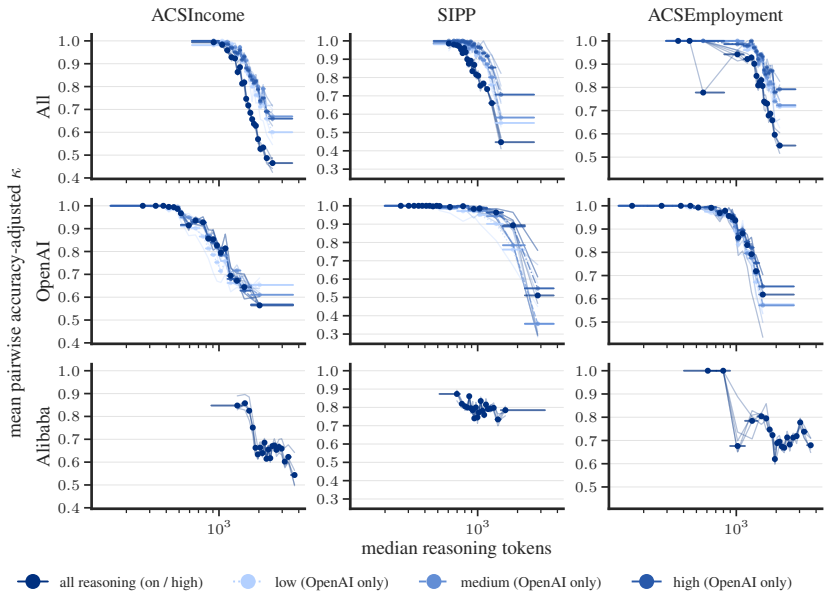
(a) Per developer.



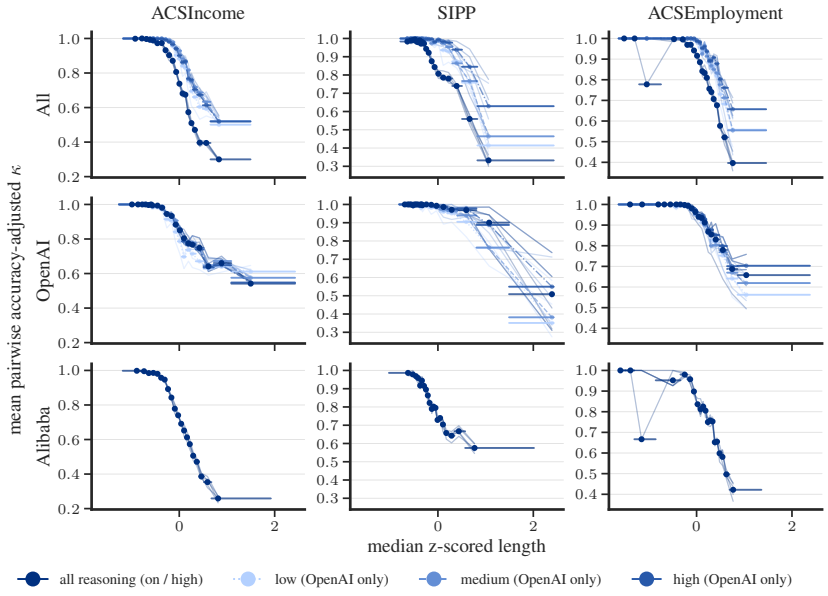
(b) Per model pair.

Figure A10: Accuracy-adjusted pairwise agreement κ by relative reasoning length, with percentile-rank bins formed (a) separately per developer and (b) separately per model pair. Rows in (a) show all models, OpenAI models, and Alibaba models. Each line in (b) represents one model pair at one reasoning setting (green: OpenAI; purple: Alibaba); dotted, dashed, and solid green lines denote `low`, `medium`, and `high` effort, respectively. Bins need not contain the same individuals across developers in (a) or across pairs in (b); the panels compare trends rather than identical groups of individuals.

Binning by developer. Figure A10a tests whether the length–agreement association persists when each developer defines its own length bins. Agreement generally decreases toward longer traces for both developers across the



(a) Raw reasoning length.



(b) Z-scored reasoning length.

Figure A11: Accuracy-adjusted pairwise agreement κ by reasoning length, measured (a) in raw tokens (logarithmic x-axis) and (b) as z-scores within each model. Rows show all models, OpenAI models, and Alibaba models, with bins formed separately using the models in each row. Dark blue denotes reasoning enabled (Alibaba `on`, OpenAI `high`); lighter curves show OpenAI-only comparisons at `low`, `medium`, and `high` effort.

three tasks. This shows that the overall trend is also visible within each model family. Because bins are constructed separately, comparisons across rows concern the shape of the relationship rather than the same groups of individuals.

Binning by model pair. Figure A10b goes further by constructing length bins separately for each model pair. Most pairs show lower agreement at longer relative lengths on ACSIncome and ACSEmployment. On SIPP, OpenAI pairs decline near the longest traces, while Alibaba pairs show substantial variation without a consistent downward trend. The association is therefore common across pairs but not universal; each curve also refers to its own groups of individuals.

Measuring reasoning length. Figure A11 checks whether the result depends on measuring length by percentile rank. The decline generally remains when using raw token counts or within-model z-scores, although the Alibaba-only curve on SIPP is approximately flat in raw tokens. Absolute token counts differ substantially across models, so relative measures facilitate comparisons while preserving many of the observed trends.

A.6.3 Observed Agreement and Reasoning Length

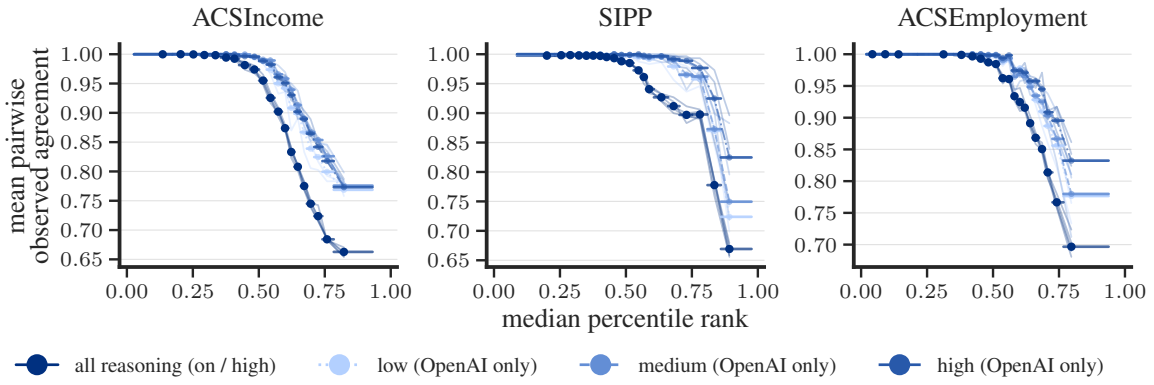


Figure A12: Mean pairwise observed agreement p_o by reasoning-length bin, using the same binning as in Figure 4: individuals are grouped by their median reasoning-length percentile rank across the nine models with reasoning enabled. Dark blue shows all nine models with reasoning enabled (Alibaba `on`, OpenAI `high`); lighter curves show the three OpenAI models at each indicated effort level. Unlike κ, p_o is not adjusted for agreement expected from model accuracy.

Observed agreement also falls toward longer reasoning traces (Figure A12), reaching approximately 0.66–0.70 in the longest bins across all nine models. The decline is therefore visible before adjusting for agreement expected from accuracy. OpenAI-only comparisons show a similar tail decline, with differences across effort levels, especially on SIPP. This supports interpreting the main result as a change in shared predictions rather than solely a consequence of the normalization in κ.

A.6.4 Recourse Levels and Reasoning Length

Figure A13 connects the pairwise results to individual recourse. Short traces are dominated by unanimous predictions, whereas longer traces more often accompany a mixture of correct and incorrect models. Relative to the independent-error baseline, full recourse has its largest excess around intermediate lengths and then declines; no recourse remains overrepresented in bins where models make errors. Greater disagreement at long lengths thus does not eliminate shared failure.

Figure A14 shows the corresponding distributions separately for each true label. Short traces are mostly associated with unanimous predictions in both groups, while intermediate recourse levels become more common toward longer traces. The composition of unanimous outcomes differs by label: for positive instances on ACSIncome and SIPP, shared errors are more prevalent in parts of the long-reasoning range. Empty short-length bins for positive ACSEmployment instances reflect the absence of those individuals in those bins.

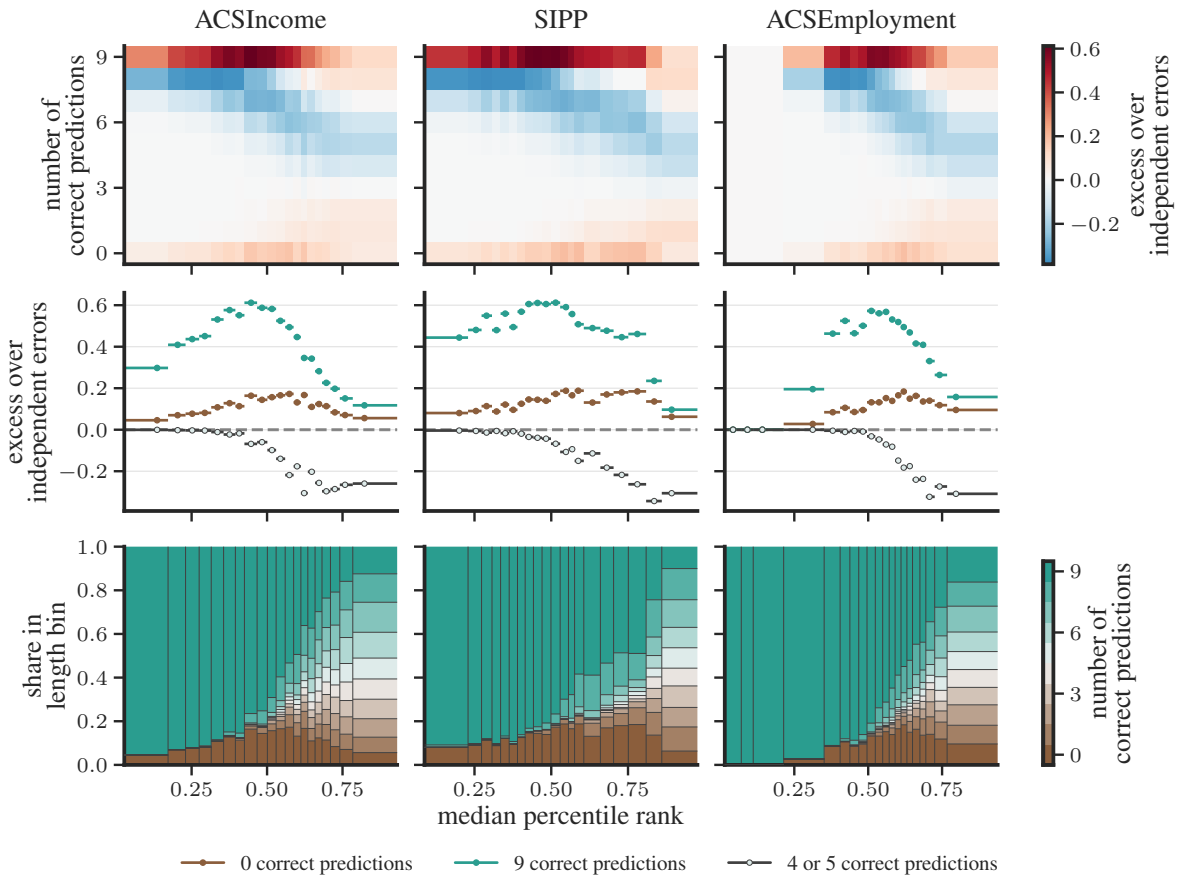


Figure A13: Recourse levels by reasoning-length bin across all nine models with reasoning enabled (Alibaba `on`, OpenAI `high`), using the binning of Figure 4. Recourse level $m/9$ is the fraction of models predicting an individual correctly. *Top:* excess over the independent-error baseline for each number m of correct models. *Middle:* excess at no recourse ($m = 0$), full recourse ($m = 9$), and the maximal split ($m \in \{4, 5\}$). *Bottom:* observed distribution of recourse levels. Baselines use accuracies within each bin; distributions and baselines are computed per resample and averaged over five resamples.

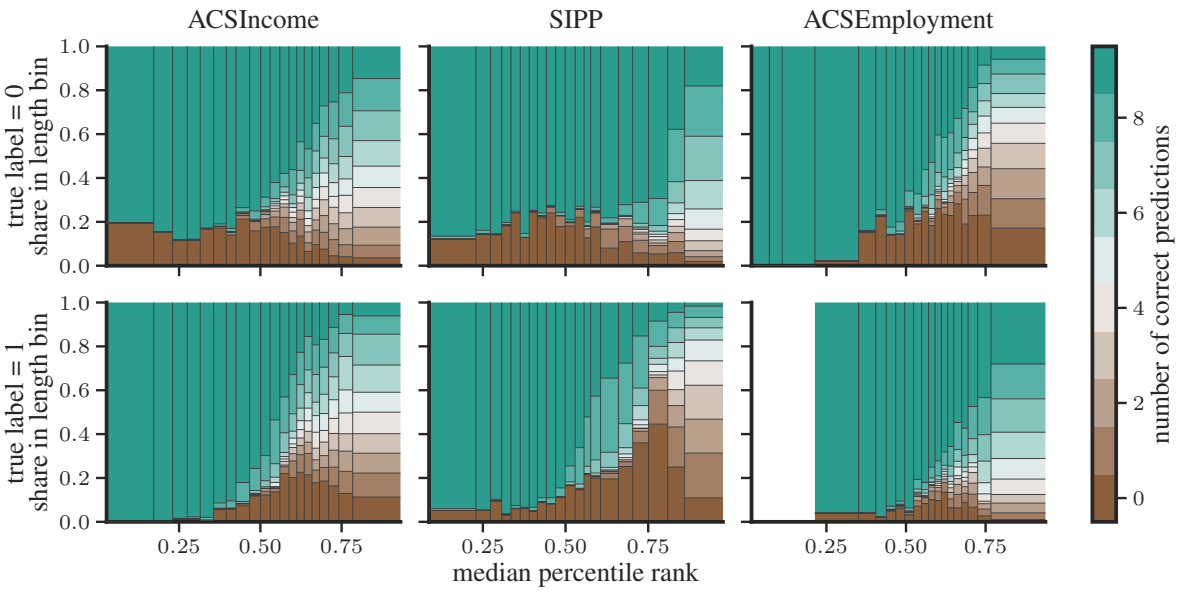


Figure A14: Observed distribution of recourse levels by reasoning-length bin, conditioned on the true label (*top:* $y = 0$; *bottom:* $y = 1$). Recourse level $m/9$ is the fraction of the nine models predicting an individual correctly, with reasoning enabled (Alibaba `on`, OpenAI `high`). Colors indicate the number m of correct models, and each bar gives shares within a length bin and label group. The x-axis is the median reasoning-length percentile rank, as in Figure 4.