Little of What Preference Data Says About a Person Is Personal

ANONYMOUS AUTHOR(S)

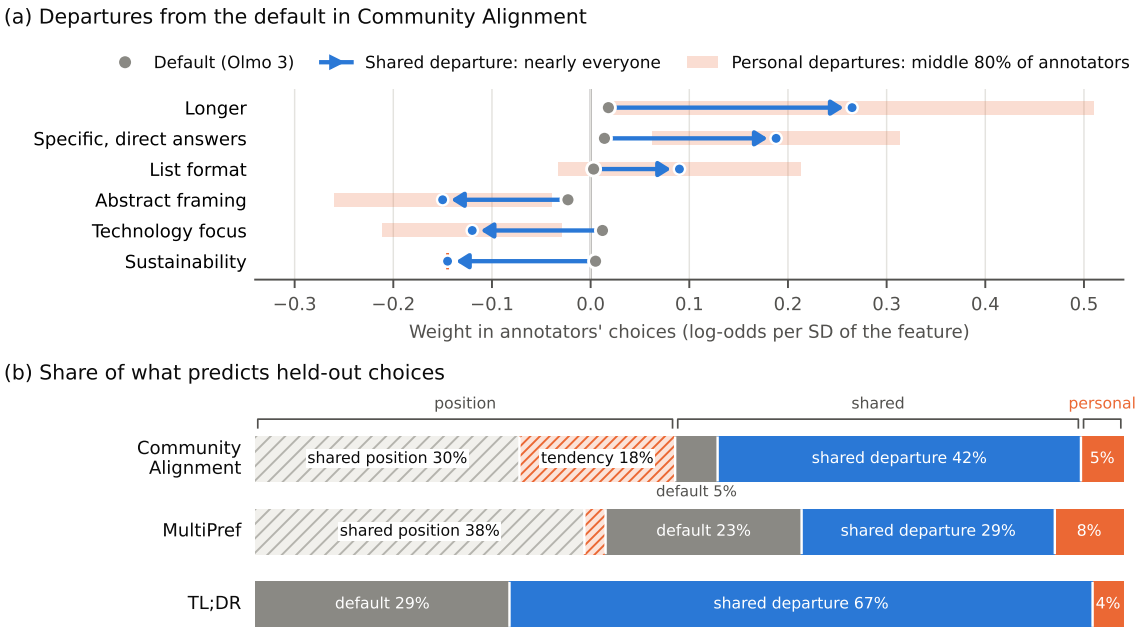


Fig. 1. In these datasets, little of what predicts a person's choices is specific to the person. (a) On named features, the default's lean (gray), the shared departure to annotators' overall weight (blue), and the estimated range of the middle 80% of annotators' own weights (orange); values are in Table 3 and Appendix A. (b) Each layer's share of the gain in predicting held-out choices over chance. The personal departure is fitted last; fitted before the position tendency, its share in Community Alignment is 10%. Gains are in Table 2 and Appendix A.

Preference tuning gives every user of a language model the same default preferences. Work that learns how people differ, to personalize a model or to ask whom tuning serves, often learns from preference data. In three public preference datasets, we measure how much of what a person's choices record is personal: specific to that person beyond what nearly everyone wants. Reading Olmo 3's default preferences from its checkpoints before and after tuning, we model each choice as layers: the shared effect of display position, each person's position tendency, the default, the departure from it that nearly everyone shares, and the person's own departure. Most of what predicts a choice is not personal. In Community Alignment, from about 60 training choices per annotator, the shared departure adds about eight times as much to predicting held-out choices as the personal departure, and position tendencies add more than the personal departure does. Without position tendencies, tuning appears to move the model away from 15% of annotators; with them, from 3%, and these data cannot determine how much of the difference reflects habit. The personal departure is specific to its annotator and replicates across defaults and datasets, but it concerns mostly the form of answers and depends on the kind of prompt. An estimated 5.5% of annotators want shorter answers while the default favors longer ones, yet the choices collected identify only 1 of 827; in simulation, 100 comparisons chosen for the purpose identify 70% of them. Studies that read individuals from preference data should model position tendencies, hold out whole prompts, and anchor at a model's default when asking what it gives people. Because no choice shows which wants a person stands behind, systems that act on what they learn should also ask; we propose four requirements for asking.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **Information systems** → **Personalization**; • **Human-centered computing** → *Human computer interaction (HCI).*

Additional Key Words and Phrases: preference data, personalization, pluralistic alignment, measurement, language models, preference tuning, position effects, endorsement

1 Introduction

One annotator in Community Alignment, a dataset collected to study how people differ in what they want from language models [82], chose among four responses 80 times. In 14 of the 21 conversations that opened with a prompt supplied by the dataset, the annotator picked the fourth response, which annotators overall picked 37% of the time. On these turns the fourth response is always the plain reply, generated separately with an instruction to be short, while the other three were written to represent differing values. A model fitted to the annotator's choices therefore infers a taste for shorter answers than most annotators prefer, although the same choices are what a habit of picking the last position, or a taste for the one reply not written to represent a value, would produce. A profile built from these choices would also list preferences that nearly every annotator shares, such as for specific, direct answers. One preference in these choices is probably the annotator's own: a stronger dislike than most for answers that focus on technology, whose direction is about 90% certain. Even that preference records only what the annotator picked, and nothing in the choices shows whether the annotator would stand behind it. Preference data records people through choices like these, and work that reads individuals from it learns from them. How much of what such choices record is specific to the person?

Preference tuning gives every user of a language model the same default preferences, the preferences the model acquires in tuning [3, 56, 61]. People differ in what they want from a model, and pluralistic alignment aims to serve those differences, by personalizing a model or by aggregating people's preferences [38, 72]. Each strategy depends on learning how people's preferences differ, and a main source is preference data, which records each person's choices between responses [39, 82]. Personalization also learns from what people write, state and correct [27, 83]; methods that learn a person from choices are mostly evaluated on simulated users or on the held-out choices of people in public preference datasets [9, 64], and studies that compare them on real people find modest gains over pooled models [17, 37]. Choices also record more than preferences: people construct answers on the spot [26] and miss swaps of their own choices [80]. Checks for position effects pool across annotators [77], and methods that learn individuals from choices between model responses do not account for a person's tendency to choose a position, so where position is tied to how responses were made, a habit is read as a preference. What these methods learn from pooled choices also mixes what the deployed default already gives with what nearly everyone wants beyond it. No study of preference data for language models has shown what the rest of a person's choices consists of, or how separating it changes what is concluded about people.

We measure how much of what a person's choices record is personal, meaning a preference about responses that is specific to that person beyond what nearly everyone wants, relative to the default of a deployed model, which every user of that model receives. We read the default preferences of Olmo 3 7B Instruct from its public checkpoints before and after preference tuning, as the reward that tuning implicitly learned [61, 75]. We describe responses with named features learned from how candidate responses differ [51], and we model each choice in three public datasets, Community Alignment, MultiPref and TL;DR [50, 74, 82], as a set of layers: the shared effect of where a response sits, each person's tendency to choose a position, the default, the departure from the default that nearly everyone shares, and the part of the person's departure specific to them. Every layer is evaluated on held-out choices.

Most of what predicts a choice is not personal (Figure 1). In Community Alignment, from about 60 training choices per annotator, the shared departure adds about eight times as much to predicting an annotator's held-out choices as the personal departure does. In this dataset, whose candidates appear to have been shown in the order they were generated, each annotator's position tendency predicts more than the personal departure, whichever of the two is fitted first. Modeling position tendencies changes a substantive conclusion: without them, preference tuning appears to move the model away from 15% of annotators, and with them, from 3%. The first position always holds the response written first, which the default favors on average, and these data cannot determine how much of the difference reflects a habit of picking that position and how much a taste for that response. In MultiPref and TL;DR, tendencies add almost nothing. The personal departure is specific to its annotator but mostly about the form of an answer and tied to the kind of prompt. An estimated 5.5% of annotators want shorter answers while the default favors longer ones, but from the choices available for fitting, only 1 of 827 can be identified as one of them. In simulation, 100 comparisons chosen for the purpose identify 70% of them.

These results serve two audiences. Studies that read individuals from preference data should model each person's position tendency, hold out whole prompts, and anchor at the deployed default when they ask what that model gives people. A system that acts on what it learns about a person faces a further limit: a choice records only what a person picked, and no choice shows which of a person's wants the person stands behind [24]. Such a system should therefore treat choices as evidence about what a person endorses, remain uncertain, and ask [29, 30].

Our contribution is a measurement with consequences for design. Methodologically, we apply to preference data for language models the layers that Xu et al. [81] fit to crowdsourced comparisons, and we anchor them at a released model's default preferences, read from its public checkpoints (Sections 3 and 4). Empirically, we show that in Community Alignment position tendencies predict more than personal preferences about content and change conclusions about whom preference tuning serves, and that the personal departure is too uncertain and too tied to its setting to identify individuals from the choices that dataset collects (Section 5). For design, we propose four requirements for systems that ask people to endorse what was learned from their choices (Section 6.2 and Figure 2).

2 Related Work

2.1 Default Preferences

Preference learning rests on an assumed model of how comparisons arise from preferences, and that model is a design choice. Reinforcement learning from human feedback fits a reward model to pairwise comparisons and optimizes a policy against it [3, 13, 56, 74, 85]. People's comparisons follow regret more closely than the partial return that standard methods assume [41], standard preference learning aggregates over hidden context by Borda count [71], and interfaces can steer people toward the preference model an algorithm assumes [31]. We keep the standard model of choice and give it separate terms for where each response sits, for the default, and for what nearly everyone wants.

Direct preference optimization makes a model's default preferences recoverable from its weights. Under its objective, the log-probability ratio between the tuned model and its checkpoint before tuning is the reward that tuning implicitly learned [61], and reward benchmarks evaluate models tuned this way as implicit reward models [46]. The data behind default preferences is increasingly synthetic. Olmo 3's preference data pairs responses from a large and a small model or labels them with a model judge [75], so its default preferences reflect the developer's tuning choices more than the judgments of any population of annotators. We read the default preferences of a model tuned this way from its public checkpoints and measure how people's choices depart from them.

2.2 Learning How People Differ

Pluralistic alignment asks whose preferences a model should serve [39, 72]. Its strategies range from personalization, within normative bounds [38], to aggregating individual preferences or matching their distribution, and each, the authors of Community Alignment argue, depends on learning how people's preferences differ, which is done mainly from preference datasets [82]. Personalization learns a person from many signals [27, 83], among them the person's writing and history, stated preferences [47, 84], edits [25] and answers to questions [48]. We study the signal that preference data records, choices between responses. Personalized reward models infer a user's preferences from few comparisons by treating the user as a latent variable [58], by placing users in a shared low-dimensional space [9, 12], or by meta-learning over synthetic users [70], and other methods turn a person's choices into a natural-language profile [23, 64]. Our results concern the data on which such methods are trained and evaluated, not any method's performance.

WIMHF describes preference data with sparse features of the differences between paired responses, names each feature in natural language, and estimates which features predict annotators' choices [51]. In Community Alignment, it also estimates how much annotators differ on each feature, and per-annotator weights on a few selected features with the population weights as a prior, which amounts to an offset from the population. Annotator Policy Models likewise learn each annotator's safety policy from their labels in an interpretable feature space and trace disagreements to their sources [55]. Weights to Words fits each person's weights over named preference dimensions and lets people correct the inferences; its studies with people use moral dilemmas and movies, and for model responses it derives dimensions without fitting anyone's choices [78]. We add to WIMHF's machinery each person's position tendency and an anchor at a model's default preferences.

Studies of real preference data find modest returns from learning individuals. Dong et al. [17] find that how much personalization helps depends on how much a dataset's annotators disagree, and that in the one real-human dataset they use, Reddit summarization comparisons, it helps little. Personalized fine-tuning on PRISM participants' earlier ratings, judged blind by 530 of them two years later, gave marginal gains over fine-tuning on pooled preferences [37]. When a language model predicted people's later choices, giving it their earlier choices, or rules inferred from those choices, predicted better than preferences the people wrote themselves [4, 20]. These studies score per-person models against pooled ones, or one signal against another, by how well each predicts held-out choices or by how people rate the resulting models. None anchors the pooled part at the default preferences of the model a person uses, separates what that default already gives from what nearly everyone wants beyond it, or models a person's tendency to choose a position.

2.3 What Choices Record Besides Preferences

Reading a person's preferences from their choices is a measurement, and whether it measures what it claims to is an empirical question [28, 33]. D'Alonzo et al. [16] read the collection of preference data for tuning as survey research that falls short of survey standards for construct validity and aggregation, and annotation practice more broadly discards what distinguishes one annotator from another [18, 52]. A growing literature questions whether comparisons measure preferences at all. Many annotation responses are non-attitudes, constructed on the spot, or artifacts of measurement [26], as are many preferences that language models express in surveys [36]; most annotators do not notice when their own choices are swapped [80]; and an annotator's state, such as fatigue or exposure to distressing content, may shift labels in ways that survive aggregation [42]. Coyne [15] distinguishes three models of annotation, as extension,

evidence and authority, and methods that learn individuals read a person's choices under the second, as evidence of that person's preferences.

Display position is a known influence on choices. The authors of Community Alignment cite survey research showing that which candidates are offered shapes conclusions about preferences [82]. Other survey research shows that respondents who satisfice, settling for an acceptable answer when an optimal one takes effort, favor options by their order [43, 44], and that respondents differ in response styles, such as a tendency to agree, which bias comparisons between people unless they are modeled [6]. In preference data, human annotators in HelpSteer2-Preference chose the second response more often than the first, which its authors attribute to how responses were sampled or to position bias [77]. This measure pools across annotators, and a small pooled lean is also what individual annotators would produce if they favored positions in opposite directions. Xu et al. [81] fit crowdsourced paired comparisons with the same layers as ours apart from the default: a common preference, each annotator's deviation from it, and each annotator's position bias, which they read as careless clicking of one side. Sachdeva et al. [65] separate item, annotator and instrument effects on hate speech labels with item response theory. Per-person position terms also appear in learning to rank from clicks [54], but not in methods that learn individuals from choices between model responses. We measure such tendencies person by person.

2.4 Endorsement and Assistance

Assistance games treat a person's behavior as evidence about what they want, which the system remains uncertain about [30]. Inverse reward design treats even a designer's stated objective as evidence that holds in the setting where it was written [29], and reward-rational choice unifies feedback of many kinds as evidence of this sort [35]. Our results bear on what such evidence contains when it comes from choices between model responses.

Philosophy of action holds that a choice does not settle which of a person's wants the person stands behind. Frankfurt [24] calls a want about which of one's wants should move one to act a second-order volition, and Christman [14] counts a preference as the person's own when the person would not feel alienated from it on reflection. Methods in AI that target a more reflective preference let the designer choose what counts as reflection [2, 49, 79], and recommender systems research models the gap between behavior and endorsement as one between an impulsive and a deliberative self [1, 40] and argues for listening to what users say as well as what they click [19]. Stenseke et al. [73] leave what counts as reflection to users: an interview elicits the better self a person would endorse, and the person confirms the resulting value profile. We also leave endorsement to the person and ask what a system must separate from a person's choices before it asks.

Systems already let people inspect and edit what has been inferred about them: natural-language user profiles [60, 62], concept-based profiles [53], rubrics inferred from browsing [5], and inferred preference dimensions, whose edits can improve prediction [78]. Each of these profiles is absolute: it states what a person wants without separating what everyone wants, what the default already gives, or what is a habit of choice. Systems are also trained and tested to follow preferences that people state [47, 84], turn critiques into editable principles [57], derive principles from public input or annotators' reasons [7, 32], and ask informative questions or comparisons [8, 48, 69].

3 The Layers of a Choice

We call the preferences that a model acquires in preference tuning, which every user receives by default, the model's default preferences. When a developer tunes a model with direct preference optimization and releases the checkpoint from before tuning, the default preferences can be read from the two checkpoints as the reward that tuning implicitly

Table 1. The layers of a choice, illustrated with response length in Community Alignment (numbers in Section 5).

Layer	What it is	Example: length	Who should act on it
Position	Where a response is displayed. People as a group favor some positions, or the responses that always fill them (the shared position effect); each person may also favor one (the position tendency).	On first turns, the short plain reply always sits fourth: a shared lean toward that position may be a preference for the reply, and a person's habit of picking it looks like a taste for short answers.	Whoever collects the data, by randomizing order
Default	The preferences a model acquires in preference tuning, which every user receives.	Olmo 3's default leans slightly toward longer answers.	The developer
Shared departure	How people's choices depart from the default on average; on most features nearly everyone departs in the same direction.	Annotators favor longer answers far more strongly than the default does.	The developer, who decides whether to follow it
Personal departure	How one person's choices depart from the default beyond the shared departure.	An estimated 5.5% of annotators want shorter answers while the default favors longer ones.	The person, by endorsing or rejecting it

learned (Section 4.2). The developer chose the preference data, the judges and the training objective, and every user receives the result. We therefore treat the default preferences as the developer's.

We read a person's choices in four layers, which Table 1 defines and illustrates with response length in Community Alignment. The default is a reference point, not a cause: our annotators never saw the model whose default we read. The layers can coincide in a single choice. An annotator who picks the short fourth response on a first turn with a supplied prompt may be expressing a habit, a personal taste for short answers, or both, while the default and nearly every other annotator favor longer answers. A system that learns from a person's choices without separating these layers lists what nearly everyone wants alongside what is specific to that person without marking which is which, and where position is tied to content it can list a habit as a preference.

Each layer calls for a different response from a system. A habit of choosing a position is not a preference about content, and a system should not act on it. A shared departure describes the default as much as the person: it marks where the default and nearly everyone's choices part ways, which the default's developer can address. We call only the last layer personal, meaning a preference about responses that is specific to the person beyond what nearly everyone wants. A preference that nearly everyone holds is still each person's preference, and a system that serves a person should honor it; the split matters for what a system shows a person and asks about, for whom tuning serves, and for who should change the default. Even the personal departure records what the person picked, and whether the person stands behind it on reflection is a further question [24], to which we return in Section 6.2.

Methods that learn individuals from choices fit each person within structure learned from pooled choices, such as a latent space, prototypes or basis rewards [9, 12, 58], or, in WIMHF, population weights from which each person is an offset [51]. The layers change what such a method attributes to a person in two ways. A profile that states a person's preferences in absolute terms presents what is shared with what is the person's, and the shared part mixes what the deployed default already gives with what nearly everyone wants beyond it; anchoring at the default changes not prediction but how the shared part is read. These methods also score each response by its content and have no

per-person term for where it was displayed, so a habit is noise when order is randomized but enters what is learned about the person when position is tied to content, as throughout Community Alignment. Section 5 measures both.

4 Data and Method

4.1 Data

We use three public preference datasets, each built for a different purpose, and use each for what its design can show.

Community Alignment, our primary dataset, was collected for research on pluralistic alignment, and its authors offer it for benchmarking personalization because it records more conversations per annotator than PRISM [82]. Each annotator holds conversations of up to four turns and at every turn picks one of four candidate responses, after being told to rank responses by their own opinion. All candidates come from Llama-3.3-70B-Instruct. To avoid a monoculture of similar answers, three first-turn candidates are written in a single instruction that asks for responses "that represent diverse values"; the fourth is what the dataset's authors call the default response, generated separately with the instruction to keep responses short while answering fully, which we call the plain reply to avoid confusion with default preferences. On later turns, all four come from the diverse-values instruction. The released data label candidates in the order in which they were generated, with the plain reply always fourth. The dataset's paper does not report randomizing the display order or describe the interface. Annotators' explanations refer to responses by the released letters (Appendix A), so we assume that the released order is the order shown. The first prompt of a conversation is either one of 1,146 prompts that the dataset's authors generated from everyday use cases—excluding politics, health and other sensitive topics—or the annotator's own. Annotators answered the supplied prompts first, and later prompts are always the annotator's own. We keep English conversations from annotators with at least 50 choices: 827 annotators, 449 in India and 378 in the United States, and 69,101 choices.

MultiPref was built as seed data for training reward models [50]. It pairs two responses from models of differing strength, and four annotators judged each pair against a rubric of helpfulness, truthfulness and harmlessness. Its display order follows which model wrote each response. Its candidates are ordinary responses, which nobody steered toward particular values, and it serves as a replication. After dropping ties and keeping annotators with at least 50 choices, it has 179 annotators and 29,674 choices.

The Reddit summarization comparisons (TL;DR) were collected to train a summarization reward model [74]. Its 70 annotators were contractors trained toward the researchers' judgments, so differences between them likely understate how much people differ, but they made a median of 1,346 choices each. It serves as a check with dense data per person: 176,492 choices in total.

For each annotator, a random 25% of their conversations (for MultiPref, of their choices) is held out for evaluation. In TL;DR, most posts appear in several comparisons, so we hold out a random 25% of posts for all annotators, and the autoencoder and every fit see only the remaining posts.

4.2 The Default Score

We take default preferences from Olmo 3 7B Instruct, whose checkpoints before and after preference tuning are both public, and which was tuned with length-normalized direct preference optimization with $\beta = 5$ [75]. Under this objective, the tuned policy π_{post} implicitly scores a response y in context x by its per-token log-probability ratio against the

pre-tuning policy π_{pre} [61]:

$$r(x, y) = \frac{\beta}{|y|} \sum_{t=1}^{|y|} \log \frac{\pi_{\text{post}}(y_t \mid x, y_{<t})}{\pi_{\text{pre}}(y_t \mid x, y_{<t})}. \tag{1}$$

We call r the default score. Both checkpoints read the same fixed conversation and candidate, and nothing is generated (Appendix A). Olmo 3 7B Instruct is trained further with reinforcement learning on verifiable rewards after preference tuning; we take its default preferences from the preference-tuning stage, which encodes preferences between responses. The annotators never interacted with Olmo 3, so their departures estimate how each would depart from its default preferences. We replicate the main analysis with OLMo 2 7B, whose post-training follows the Tülu 3 recipe with the same objective and β [45, 76].

4.3 Response Features

Following WIMHF [51], we describe candidates by named features learned from how they differ. Each candidate's embedding is the tuned checkpoint's final-layer state averaged over its tokens, minus the mean embedding of the candidates in its turn. A sparse autoencoder with a Matryoshka BatchTopK objective [10, 11], trained on training turns, maps these differences to 32 signed features, about four of which are active for any candidate. Log length enters as a separate feature, and all features are centered within each turn and standardized.

Following WIMHF's procedure, Qwen3.8-27B [59] names each feature. A name's fidelity is the correlation between the feature and the model's judgments of which of two candidates the name fits, and a name is valid if its fidelity on held-out conversations is significant at $p < 0.05/32$ (Appendix A). In Community Alignment, 26 of 32 names are valid, with a mean fidelity of 0.43. As shared features in the choice model, the named features keep 84% of the gain in held-out log-likelihood that the full 4,096-dimensional embedding achieves over position and default score. We also checked every name we interpret by hand, blind to the judge's answers (Appendix A). All 32 features enter the choice model, and our interpretation uses only names that passed both checks.

4.4 Choice Model

We model each choice as a conditional logit over the candidates in a turn. Annotator i chooses candidate k with probability proportional to $\exp(u_{ik})$, where

$$u_{ik} = b_k + h_{ik} + a\, r_k + (w + \delta_i)^\top \phi_k, \tag{2}$$

b_k is the shared preference for display position k, or for whatever response always fills it; h_{ik}, the position tendency, is annotator i's own propensity to choose position k, whether from habit or from a taste for the response that always fills it; a is an exchange rate that converts default scores into the scale of annotators' choices; and ϕ_k are the features of candidate k. Because the default score is in the model, $w + \delta_i$ is annotator i's departure from the default: w is the shared departure and δ_i the personal departure. We fit b, a and w by maximum likelihood on all annotators' training turns, then fit each annotator's position tendencies and δ_i on that annotator's training turns alone, under normal priors centered on zero. Following WIMHF, each feature's prior variance is estimated by a random-effects meta-analysis across annotators, so features on which annotators differ more are shrunk less; we estimate these variances beyond each annotator's position tendency. In the main specification, position terms, shared and personal, are fitted separately for first and later turns, because Community Alignment's first-turn positions differ in how their responses were generated.

We evaluate on held-out turns by the log-probability that the model assigns to the annotator's actual choice, averaged over choices, and report gains in nats per choice with bootstrap confidence intervals. A gain of 0.01 nats per choice

raises the geometric-mean probability of the actual choices by about 1%. A personal departure is person-specific if, with the annotator's own position tendency held fixed, it predicts the annotator's held-out choices better than the average of other annotators' departures; we test this against a null distribution from randomly reassigning departures among annotators. The default's lean on a named feature is $a\,d_j$, where d_j is the coefficient from regressing the default scores of a turn's candidates on their features, and annotators' overall weight is $a\,d_j + w_j$.

4.5 Whom the Default Works Against

We ask, person by person, whether the default works against a person in two ways. Directly, we pool each annotator's own weight on the default score, fitted together with their position tendency, in a random-effects meta-analysis; the estimated share of negative weights is the share whom preference tuning moved the model away from. By feature, the default works against a person when, on a validly named feature on which the default leans significantly, the person's total weight $a\,d_j + w_j + \delta_{ij}$ has the opposite sign to the default's lean; we estimate this share from the pooled distribution of personal weights and count individuals whose posterior probability of an opposite sign is at least 0.8, with the expected number of false identifications. To ask how many choices it takes to find them, we simulate annotators from the fitted population and contrast random comparisons with ones drawn from the 10% of turns in which the feature differs most, shown in random order (Appendix A).

5 Results

We report five findings (Sections 5.1–5.5) and repeat the main analyses with a second default model, OLMo 2, and on MultiPref and TL;DR, reporting each repetition next to the finding it tests.

5.1 Most of What Predicts a Choice Is Not Personal

Most of what predicts an annotator's choice is shared with other annotators or comes from where the responses sit (Table 2). The shared position effect alone raises accuracy from the 25% expected by chance to 38.5%, the default adds 0.3 points, and the shared departure adds 6.1 points. Beyond these shared layers, each annotator's position tendency adds 0.033 nats per choice, and the personal departure adds 0.009 beyond the tendency. Of the gain over the default, the shared departure (0.077) thus contributes about eight times as much as the personal departure; without position tendencies in the model, the personal departure adds 0.019, and the shared departure still predicts four times as much. Together, the person-specific layers are 23% of the gain over chance (Figure 1b), so most of what predicts a choice is shared however the position tendency is read. In TL;DR, the default carries far more weight, and the personal departure is again far smaller than the shared departure: the default raises accuracy by 7.4 points, the shared departure by 6.0 and the personal departure by 0.2.

5.2 People Depart from the Default Mostly in the Same Direction

Annotators lean the default's way on style, more strongly than the default does, and penalize value angles that Community Alignment's candidates introduced unprompted, toward which the default leans only slightly (Figure 1a; Table 3 in Appendix A places the default's lean and annotators' overall weight on the same scale). The default favors specific, direct answers and longer answers, and annotators favor both more strongly. The default barely favors responses that raise sustainability and slightly favors those that focus on technology, about as much as it favors specific, direct answers, while annotators as a group penalize both; both penalties hold in all of 100 bootstrap resamples of annotators. Once position tendencies are modeled, an estimated 95% of annotators penalize technology focus, and on sustainability

Table 2. Held-out choices in Community Alignment (827 annotators, 16,928 held-out choices; chance is 25% and −1.386 nats per choice). Each row adds one layer to the model above it, and the gain is in log-likelihood per choice over that model; the first row's gain is over chance. The shared rows use position terms common to all turns and intervals from resampling held-out turns. The person-specific rows split position terms by first and later turns, measure the personal departure beyond the annotator's own position tendency, and take intervals from resampling annotators. Accuracy is reported for the shared rows.

Layer	Accuracy	Gain per choice in nats [95% CI]
Shared layers		
Shared position effect	38.5%	+0.056 [0.050, 0.061]
+ default	38.8%	+0.009 [0.007, 0.012]
+ shared departure	44.9%	+0.077 [0.071, 0.083]
Person-specific layers		
+ position tendency	–	+0.033 [0.027, 0.039]
+ personal departure	–	+0.009 [0.007, 0.012]

the spread between annotators is not distinguishable from zero. These angles were added without being asked for, and each annotator gets one weight per feature across topics, so the result concerns the angles these candidates carried, not people's values, on which the dataset's authors find that people differ [82]. The penalties hold even on annotators' own prompts, which they were asked to write about their values. MultiPref, whose candidates are ordinary responses, shows the same pattern: the default leans toward detailed, comprehensive explanations with structured formatting, and an estimated 74% of annotators [65, 85] lean against them. Where nearly everyone departs from the default in the same direction, the departure describes the default.

5.3 Position Tendencies Predict More Than the Personal Departure

Position tendencies predict held-out choices better than the personal departure does, whichever of the two is fitted first. Fitted alone over the shared model, the position tendency adds 0.033 nats per choice and the personal departure 0.019; fitted beyond each other, they add 0.023 and 0.009. The contrast is cleanest on annotators' own prompts, where the candidates in each position differ little in length. Fitted on those prompts alone, the tendency adds 0.022 nats per choice and the personal departure, beyond it, 0.006 [0.003, 0.008]. On supplied first-turn prompts, where the fourth position always holds the plain reply, the figures are 0.067 and 0.009.

Because the released order follows how responses were generated, position terms could instead record tastes for kinds of response. The term for the fourth position behaves mostly like a habit: that position holds the plain reply on first turns and an ordinary reply on later turns, yet a person's tendency to pick it carries across turns nearly as well as their tendencies for the second and third positions (correlations corrected for reliability: 0.80 against 0.87 and 0.87; Appendix A). The first position always holds the reply the model wrote first, so for it a habit and a taste cannot be separated. Tendencies also depend on how the data were collected. In MultiPref, where display order follows which model wrote each response and annotators judged against a rubric, the position tendency adds 0.003 nats per choice, with an interval that includes zero, and in TL;DR it adds nothing.

Modeling position tendencies changes a substantive conclusion about whom preference tuning serves. Without them, tuning appears to move the model away from an estimated 15% of Community Alignment annotators. With them, the share falls to 3%, and to 2% on annotators' own prompts; in MultiPref it is 0%. The default score is highest, on average, for the response the model wrote first, which it ranks above the other three on 34% of first turns and 49% of later turns,

so the share depends on whether a person's tendency to pick or avoid the first position is a habit or a taste for that response, which these data cannot separate. Annotators still differ reliably in how far tuning moves the model toward them (mean +0.19, standard deviation 0.10, heterogeneity $p < 10^{-27}$), and country explains 0.6% of that variation.

5.4 The Personal Departure Is Specific, Mostly About Form and Tied to Context

The personal departure is specific to its annotator. With the annotator's own position tendency held fixed, an annotator's personal departure predicts their held-out choices better than the average of other annotators' departures for 74.8% of annotators, well above the 58.6% that random reassignment of departures reaches at its 95th percentile (Appendix A). The result replicates with OLMo 2's default (+0.011 nats per choice, specific for 77% of annotators) and in MultiPref (+0.008, 73%). In TL;DR, where position terms add nothing, the personal departure fitted without them, with whole posts held out, adds 0.002 nats per choice [0.001, 0.003] and is specific for 80% of annotators, against a null 95th percentile of 64%. The result also holds when the shared model is fitted on a disjoint half of the annotators, and it is smaller but still specific under a single ridge penalty, across autoencoder seeds and across 16 to 128 features (Table 6 in Appendix A).

The personal departure grows with data and shows no plateau within the choices collected. Fitted together with the position tendency on 5, 10, 20 and 40 choices per annotator, it adds 0.002, 0.004, 0.006 and 0.009 nats per choice, and for the 48 annotators with 100 training choices it reaches 0.010 [0.001, 0.019].

The personal departure is mostly about form. Beyond position tendencies, 5 of the 33 features (the 32 from the autoencoder, and length) vary significantly across annotators, with person-to-person standard deviations of 0.19 for length, 0.10 for specific and direct answers, 0.10 for list format, 0.09 for abstract framing and 0.07 for technology focus. On the five features the spreads are large next to annotators' overall weights (length 0.19 against 0.265; specific and direct answers 0.10 against 0.188), so annotators differ in form more than about 60 choices each reveal. In MultiPref, length is again the most personal feature (0.18), followed by whether the response refuses (0.11).

The personal departure depends on the kind of prompt (Table 4 in Appendix A). Fitted once on all of an annotator's training choices, it is largest on supplied first-turn prompts (+0.019 nats per choice), smaller on later turns (+0.007), and not distinguishable from zero on first turns with the annotator's own prompt (+0.004 [−0.001, 0.010]). It transfers weakly between kinds of prompts. A departure fitted only on supplied prompts predicts later turns by 0.002 [0.001, 0.004], and one fitted only on own prompts predicts supplied prompts by 0.006 [0.001, 0.011]; each is about a third of what the departure fitted on all training choices achieves on the same turns. Our estimator fits one departure per annotator across all prompts, and choices on one kind of prompt say little about another. The personal part that these choices reveal is therefore small, about the form of an answer, and partly specific to the setting in which the choices were made.

5.5 Too Few Choices Are Collected to Find Whom the Default Works Against

In Community Alignment, annotators whom the default works against are estimated to exist, but the data cannot identify them individually. Length is the one feature on which the default works against a minority that holds in both random halves of the annotators and on own prompts: under a normal model of personal weights, an estimated 5.5% of annotators [4.0, 7.2] want shorter answers while the default favors longer ones. Fitted on each annotator's training choices, about 60 on average, only 1 of 827 annotators can be identified as one of them with a posterior probability of at least 0.8. In MultiPref, where annotators made about twice as many choices and most of them lean against the default's taste for detailed explanations, 86 of 179 can be identified, against 6.4 expected false identifications.

Choosing the comparisons and showing them in random order finds such annotators with several times fewer choices (Table 5 in Appendix A); the simulation does not separate the two. In a simulation from the fitted Community Alignment population, which assumes that the fitted model is correct and draws from the dataset's existing candidates, 100 comparisons drawn from the turns in which length differs most identify 70% of the annotators who want shorter answers, at 95% precision, and 100 drawn at random identify 13%. Ignoring position tendencies costs precision: at 100 random comparisons, precision for length falls from 88% to 64%, because the tendencies are read as preferences.

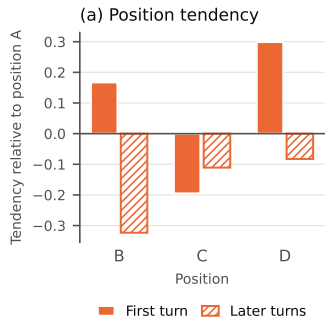
6 Discussion

Most of what predicts a person's choices among model responses is shared with other people or, in Community Alignment, is the person's position tendency. Personal preferences exist in these data: the personal departure is specific to its annotator, replicates across defaults and datasets, and grows with data. Our claim concerns what about 60 training choices per person can show once the layers that are not specific to the person are set aside. The personal departure stays small across the priors, penalties and feature counts we tested (Table 6), and in TL;DR, with a median of 1,346 choices per annotator, it is 4% of the gain, though those annotators were trained toward the researchers' judgments. These results do not show that people agree. People want different things from a model when asked in their own words [68], and suppressing such disagreement has ethical and epistemic costs [22]. Choices between responses record little of how people differ, which is one reason to ask people directly (Section 6.2).

6.1 Measuring a Person

Studies that read individuals from preference data risk attributing a person's position tendency to their preferences about content unless they model it wherever display position is tied to content. In Community Alignment, the position tendency carries more person-specific signal than content does, and modeling it cut the share of annotators whom preference tuning appeared to work against from 15% to 3%. Because position there also marks the order in which responses were written, these data cannot determine how much of that change reflects habit, and an analysis without per-person position terms reports 15% as if every tendency to pick a position were a preference. Position tendencies depend on how data are collected: they add almost nothing in MultiPref, where two responses of differing quality are judged against a rubric, and they are the largest person-specific signal in Community Alignment, where four responses written to differ compete for an annotator's own preference. Our results add a measurable case to the argument that many annotation responses are not preferences at all [26], since part of what looks like a person's preference behaves, at least at the fourth position, like a habit of choosing a position, the counterpart in preference data of a survey response style [6]. Data collected to learn individuals should randomize the order of responses for every choice, record the order shown, and model each person's position terms before reading anything into their content weights [16].

Anchoring at the default separates what the default already gives from what people want beyond it. Methods that learn individuals from choices fit each person within structure learned from pooled choices [9, 12, 51, 58], and that shared structure mixes the two. Anchoring does not change prediction, but with both checkpoints of a model, the shared departure becomes a measurable description of where that model's default and people's choices part ways. The clearest case in our data is MultiPref, whose candidates are ordinary responses: Olmo 3's default leans toward detailed, comprehensive explanations with structured formatting, and an estimated 74% of annotators lean against them. Such a description is an audit of the default that its developer could act on. In Community Alignment the default barely orders the candidates, so the shared departure there mostly records what people want on dimensions on which the default



(b) What a system that asks would show this annotator

Where nearly everyone departs from the default (for the developer to address; you can reject any line for yourself): compared with the default, nearly everyone prefers longer, more specific and direct answers, less abstract framing, and less unprompted discussion of sustainability or technology.

Where you differ from most people (for you to keep, drop or revise): you penalize answers that focus on technology, digital tools or data analytics more than most people do.

Not yet clear (the system asks with chosen comparisons): whether you prefer shorter answers than most people.

Not yet treated as a preference (tested with responses in random order): on first turns you often pick the fourth response, which always held the plain reply. You picked it on supplied prompts even when it was not the shortest, and never on your own first prompts, so your choices point to something tied to the supplied prompts and cannot show whether it is a habit or a taste for that reply.

Fig. 2. The annotator from the introduction, with layers fitted on the annotator's 63 training choices and selected by a rule fixed in advance (Appendix A). (a) The annotator's position tendency, relative to the first position. (b) The layers as a system that asks would present them.

leans little, including reactions to the candidates the dataset supplied. Extending the audit to closed models would require describing their default from behavior.

Evaluating personal preferences requires holding out whole items whenever the same person judged an item more than once. TL;DR is used to evaluate personalization [9, 17], and most of its posts appear in several comparisons. An earlier version of our analysis held out single comparisons, whose posts had also appeared in training; it raised the share of annotators whose own departure predicted their choices better than the average of others' departures from 80% to 96%, and doubled the personal gain. More generally, accuracy on held-out choices does not show that a per-person model measures the person's preference, because position tendencies and the shared departure also predict choices; gains beyond a shared model with per-person position terms, and person-specificity, separate them. For the same reason, a preference a person states can describe them faithfully and still predict their choices poorly, which may explain part of the advantage of preferences inferred from choices [4, 20].

6.2 Asking the Person

For a system that acts on what it learns about a person, the personal departure that remains after position tendencies and the shared departure is too uncertain, from about 60 choices, to act on without asking: it identifies only 1 of the estimated 45 annotators who want shorter answers while the default favors longer ones, and it transfers between kinds of prompts at about a third of its strength.

Even an accurate estimate of a person's departure records what the person picked. People hold wants about their own wants, such as a student who wants hints even when tempted to ask for the answer [24, 34], and no choice between two responses shows which of a person's wants the person stands behind. Defining the reflective want as what an improved version of the person would choose transfers that decision to whoever specifies the improvement [21, 63]. Assistance games offer a different stance: a person's behavior is evidence about what they want, the system remains uncertain, and it asks when the evidence does not settle the question [29, 30, 35]. Systems that take stated preferences, critiques and answers to questions already ask in this sense (Section 2). Our results specify what such a system must

separate from a person's choices before it asks, and what to ask about; we propose four requirements, the first three of which rest on our results.

The first two requirements separate what is not the personal departure. First, the system models a person's position tendency before it shows anything, so that a habit is not presented as a preference, and tests the tendency with responses in random order, since it may be a taste for the response in that position. Second, it separates the shared departure from the personal one. Profiles that people can inspect and edit state preferences in absolute terms [5, 60, 78], and our results imply that a profile inferred from choices like these would mostly describe what nearly everyone wants; a person shown that they prefer specific, direct answers learns something true of almost all annotators. The system should present the shared part as where nearly everyone departs from the default, which the developer can address.

The other two requirements govern how the system asks. Third, the system chooses the comparisons it asks for, aimed at the features on which a person may depart from the default. In simulation, 100 chosen comparisons identified 70% of annotators who want shorter answers, against 13% for random ones; drawing them from a fixed pool is a non-adaptive version of the query selection that active preference learning makes adaptive [66]. Fourth, the system presents each line of the personal departure for the person to endorse, reject or revise, and lets the person reject a shared line for themselves, since a want that nearly everyone shares may still be one a person does not stand behind. Weights to Words uses people's edits to inferred preferences to predict their choices better, and in its movie study people preferred profiles built from their edits although those profiles predicted no better [78]. Endorsement serves a different purpose: it decides which of a person's wants should govern the model, and a choice cannot supply it. Answers to an endorsement prompt are themselves subject to response styles, such as a tendency to agree [6], which the system's design must also account for.

Figure 2b applies the four requirements to the annotator from the introduction. Showing a person how they depart from a default could also pull them toward it, as feedback relative to a norm does [67]. Whether people recognize what such a system shows them, what they endorse, and how showing it changes them are questions for a study with participants.

6.3 Limitations

Annotators are not users. No person in our data saw or judged what we inferred about them, and the annotators chose among responses written by models other than the one whose default we read. The results therefore describe what such choices can reveal about a person. Whether people recognize and endorse their personal departure, and what they want from a model they use, requires a study in which people review what was inferred from their own choices.

One dataset carries the main results. Community Alignment's design serves its aim of eliciting differing values: one model wrote every candidate, most candidates were written to represent differing values, and the released order follows how responses were generated. For our question, this means that the default barely orders the candidates, that position terms may partly record tastes for kinds of response, and that the supplied prompts, generated by a model and answered first, confound the kind of prompt with time on the task. MultiPref and TL;DR replicate the small, person-specific departure but not the position tendencies, which are negligible there; their annotators also followed rubrics, which narrow personal differences, and MultiPref's experts saw prompts from one field each. Data that randomize the order of responses and collect many choices per person under instructions to follow one's own preference would remove these confounds. Beyond the data, the named features keep 84% of the full embedding's gain, so some personal signal lies outside them; one model family supplies both defaults; and the simulation assumes the fitted model is correct.

References

[1] Arpit Agarwal, Nicolas Usunier, Alessandro Lazaric, and Maximilian Nickel. 2024. System-2 Recommenders: Disentangling Utility and Engagement in Recommendation Systems via Temporal Point-Processes. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. doi:10.1145/3630106.3659004

[2] Anthropic. 2026. Claude's Constitution. https://www.anthropic.com/constitution.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, et al. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862

[4] Nishant Balepur, Malachi Hamada, Varsha Kishore, Sergey Feldman, Amanpreet Singh, Pao Siangliulue, Joseph Chee Chang, Rachel Rudinger, Eunsol Choi, Jordan Lee Boyd-Graber, Doug Downey, and Aakanksha Naik. 2026. DRACULA: Hunting for the Actions Users Want Deep Research Agents to Execute. arXiv:2604.23815

[5] Aadit Barua, Leijie Wang, and Amy X. Zhang. 2026. Compass: Continuously Aligning Social Media Feeds via In-Situ Reflections. In *Proceedings of the 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26)*. arXiv:2608.04274

[6] Hans Baumgartner and Jan-Benedict E. M. Steenkamp. 2001. Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research* 38, 2 (2001), 143–156. doi:10.1509/jmkr.38.2.143.18840

[7] Henry Bell, Lara Neubauer da Costa Schertel, Bochu Ding, and Brandon Fain. 2026. Learning Alignment Principles Grounded in Human Reasons and Values. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 6956–6997. doi:10.1145/3805689.3806453

[8] Avinandan Bose, Shuyue Stella Li, Faeze Brahman, Pang Wei Koh, Simon Shaolei Du, Yulia Tsvetkov, Maryam Fazel, Lin Xiao, and Asli Celikyilmaz. 2026. Cold-Start Personalization via Bayesian Adaptive Questioning. In *Proceedings of the 43rd International Conference on Machine Learning (PMLR, Vol. 306)*. 9295–9321.

[9] Avinandan Bose, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, and Maryam Fazel. 2025. LoRe: Personalizing LLMs via Low-Rank Reward Modeling. arXiv:2504.14439

[10] Bart Bussmann, Patrick Leask, and Neel Nanda. 2024. BatchTopK Sparse Autoencoders. arXiv:2412.06410

[11] Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. 2025. Learning Multi-Level Features with Matryoshka Sparse Autoencoders. arXiv:2503.17547

[12] Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. 2025. PAL: Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment. In *International Conference on Learning Representations (ICLR)*. 15755–15799. https://openreview.net/forum?id=1kFDrYCuSu

[13] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, Vol. 30. arXiv:1706.03741

[14] John Christman. 1991. Autonomy and Personal History. *Canadian Journal of Philosophy* 21, 1 (1991), 1–24. doi:10.1080/00455091.1991.10717234

[15] Steve Coyne. 2026. Three Models of RLHF Annotation: Extension, Evidence, and Authority. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 775–791. arXiv:2604.25895 doi:10.1145/3805689.3812225

[16] Samantha D'Alonzo, Frauke Kreuter, and Serena Booth. 2026. Helpful, Harmless, Honest? RLHF as Survey Design and Content Moderation. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 7587–7603. doi:10.1145/3805689.3806444

[17] Yijiang River Dong, Tiancheng Hu, Yinhong Liu, Ahmet Üstün, and Nigel Collier. 2025. When Personalization Meets Reality: A Multi-Faceted Analysis of Personalized Preference Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, Suzhou, China, 16880–16894. arXiv:2502.19158 doi:10.18653/v1/2025.findings-emnlp.916

[18] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Remi Denton. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2342–2351. doi:10.1145/3531146.3534647

[19] Michael D. Ekstrand and Martijn C. Willemsen. 2016. Behaviorism Is Not Enough: Better Recommendations through Listening to Users. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, 221–224. doi:10.1145/2959100.2959179

[20] Keaton Ellis and Wanying Huang. 2026. Should I State or Should I Show? Aligning AI with Human Preferences. arXiv:2603.29317

[21] David Enoch. 2005. Why Idealize? *Ethics* 115, 4 (2005), 759–787. doi:10.1086/430490

[22] Sina Fazelpour and Will Fleisher. 2025. The Value of Disagreement in AI Design, Evaluation, and Alignment. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2138–2150. doi:10.1145/3715275.3732146

[23] Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert Mullins. 2025. Inverse Constitutional AI: Compressing Preferences into Principles. In *International Conference on Learning Representations (ICLR)*. arXiv:2406.06560

[24] Harry G. Frankfurt. 1971. Freedom of the Will and the Concept of a Person. *The Journal of Philosophy* 68, 1 (1971), 5–20. doi:10.2307/2024717

[25] Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. 2024. Aligning LLM Agents by Learning Latent Preference from User Edits. In *Advances in Neural Information Processing Systems*, Vol. 37. 136873–136896. arXiv:2404.15269

[26] Bijean Ghafouri, Eun Cheol Choi, Priyanka Dey, and Emilio Ferrara. 2026. Position: RLHF May Not Reflect Genuine Preferences. In *Proceedings of the 43rd International Conference on Machine Learning (ICML), Position Paper Track*. arXiv:2604.03238

[27] Jian Guan, Junfei Wu, Jia-Nan Li, Chuanqi Cheng, and Wei Wu. 2025. A Survey on Personalized Alignment—The Missing Piece for Large Language Models in Real-World Applications. In *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria, 5313–5333. arXiv:2503.17003 doi:10.18653/v1/2025.findings-acl.277

[28] Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. 2023. Ground(less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 688–704. arXiv:2302.06503 doi:10.1145/3593013.3594036

[29] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. 2017. Inverse Reward Design. In *Advances in Neural Information Processing Systems*, Vol. 30. arXiv:1711.02827.

[30] Dylan Hadfield-Menell, Stuart J. Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Vol. 29. arXiv:1606.03137

[31] Stephane Hatgis-Kessell, W. Bradley Knox, Serena Booth, and Peter Stone. 2026. Influencing Humans to Conform to Preference Models for RLHF. *Transactions on Machine Learning Research* (2026). arXiv:2501.06416

[32] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 1395–1417. doi:10.1145/3630106.3658979

[33] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 375–385. arXiv:1912.05511 doi:10.1145/3442188.3445901

[34] Richard C. Jeffrey. 1974. Preference Among Preferences. *The Journal of Philosophy* 71, 13 (1974), 377–391. doi:10.2307/2025160

[35] Hong Jun Jeon, Smitha Milli, and Anca D. Dragan. 2020. Reward-Rational (Implicit) Choice: A Unifying Formalism for Reward Learning. In *Advances in Neural Information Processing Systems*, Vol. 33. 4415–4426.

[36] Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2151–2165. arXiv:2503.08688 doi:10.1145/3715275.3732147

[37] Hannah Rose Kirk, Liu Leqi, Fanzhi Zeng, Henry Davidson, Bertie Vidgen, Christopher Summerfield, and Scott A. Hale. 2026. PRISM-X: Experiments on Personalised Fine-Tuning with Human and Simulated Users. arXiv:2605.13307

[38] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2024. The Benefits, Risks and Bounds of Personalizing the Alignment of Large Language Models to Individuals. *Nature Machine Intelligence* 6, 4 (2024), 383–392. doi:10.1038/s42256-024-00820-y

[39] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. In *Advances in Neural Information Processing Systems*, Vol. 37. 105236–105344.

[40] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2024. The Challenge of Understanding What Users Want: Inconsistent Preferences and Engagement Optimization. *Management Science* 70, 9 (2024), 6336–6355. doi:10.1287/mnsc.2022.03683

[41] W. Bradley Knox, Stephane Hatgis-Kessell, Serena Booth, Scott Niekum, Peter Stone, and Alessandro Allievi. 2024. Models of Human Preference for Learning Reward Functions. *Transactions on Machine Learning Research* (2024). arXiv:2206.02231

[42] Elena Kopteva and Vitaliy Hlynianyi-Zhuk. 2026. Rater State Bias in RLHF Preference Data: An Audit Framework. arXiv:2607.16195

[43] Jon A. Krosnick. 1991. Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology* 5, 3 (1991), 213–236. doi:10.1002/acp.2350050305

[44] Jon A. Krosnick and Duane F. Alwin. 1987. An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly* 51, 2 (1987), 201–219. doi:10.1086/269029

[45] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, et al. 2024. Tülu 3: Pushing Frontiers in Open Language Model Post-Training. arXiv:2411.15124

[46] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, L. J. Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. RewardBench: Evaluating Reward Models for Language Modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Albuquerque, NM, 1755–1797. arXiv:2403.13787

[47] Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. 2024. Aligning to Thousands of Preferences via System Message Generalization. In *Advances in Neural Information Processing Systems*, Vol. 37. 73783–73829. arXiv:2405.17977 doi:10.52202/079017-2347

[48] Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2025. Eliciting Human Preferences with Language Models. In *International Conference on Learning Representations (ICLR)*. arXiv:2310.11589

[49] Kaiqu Liang, Haimin Hu, Ryan Liu, Thomas L. Griffiths, and Jaime Fernández Fisac. 2026. RLHS: Mitigating Misalignment in RLHF with Hindsight Simulation. In *Findings of the Association for Computational Linguistics: ACL 2026*. 11457–11483. arXiv:2501.08617

[50] Lester James V. Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. 2024. Hybrid Preferences: Learning to Route Instances for Human vs. AI Feedback. arXiv:2410.19133

[51] Rajiv Movva, Smitha Milli, Sewon Min, and Emma Pierson. 2026. What's In My Human Feedback? Learning Interpretable Descriptions of Preference Data. In *International Conference on Learning Representations (ICLR)*. 59420–59455. arXiv:2510.26202

[52] Sheza Munir, Benjamin Mah, Krisha Kalsi, Shivani Kapania, Julian Posada, Edith Law, Ding Wang, and Syed Ishtiaque Ahmed. 2026. The Consensus Trap: Dissecting Subjectivity and the "Ground Truth" Illusion in Data Annotation. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2643–2680. arXiv:2602.11318 doi:10.1145/3805689.3806510

[53] Sheshera Mysore, Mahmood Jasim, Andrew McCallum, and Hamed Zamani. 2023. Editable User Profiles for Controllable Text Recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. ACM, 993–1003. doi:10.1145/3539618.3591677

[54] Zechun Niu, Lang Mei, Liu Yang, Ziyuan Zhao, Qiang Yan, Jiaxin Mao, and Ji-Rong Wen. 2025. Addressing Personalized Bias for Unbiased Learning to Rank. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*. ACM, Seoul, Republic of Korea, 2222–2231. arXiv:2508.20798 doi:10.1145/3746252.3761377

[55] Alex Oesterling, Donghao Ren, Yannick Assogba, Dominik Moritz, Sunnie S. Y. Kim, Leon Gatys, and Fred Hohman. 2026. Understanding Annotator Safety Policy with Interpretability. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 5653–5690. arXiv:2605.05329 doi:10.1145/3805689.3806472

[56] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, Vol. 35. arXiv:2203.02155

[57] Savvas Petridis, Benjamin D. Wedin, James Wexler, Mahima Pushkarna, Aaron Donsbach, Nitesh Goyal, Carrie J. Cai, and Michael Terry. 2024. ConstitutionMaker: Interactively Critiquing Large Language Models by Converting Feedback into Principles. In *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24)*. ACM, New York, NY, 853–868. arXiv:2310.15428 doi:10.1145/3640543.3645144

[58] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning. In *Advances in Neural Information Processing Systems*, Vol. 37. 52516–52544. arXiv:2408.10075

[59] Qwen Team. 2026. Qwen3.8-Max: A New Bar for Coding and Cowork. Blog post. https://qwen.ai/blog?id=qwen3.8

[60] Filip Radlinski, Krisztian Balog, Fernando Diaz, Lucas Dixon, and Ben Wedin. 2022. On Natural Language User Profiles for Transparent and Scrutable Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. ACM, 2863–2874. arXiv:2205.09403 doi:10.1145/3477495.3531873

[61] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, Vol. 36. 53728–53741. arXiv:2305.18290

[62] Jerome Ramos, Hossein A. Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. 2024. Transparent and Scrutable Recommendations Using Natural Language User Profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand, 13971–13984. arXiv:2402.05810 doi:10.18653/v1/2024.acl-long.753

[63] Connie S. Rosati. 1995. Persons, Perspectives, and Full Information Accounts of the Good. *Ethics* 105, 2 (1995), 296–325. doi:10.1086/293702

[64] Michael J. Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. 2025. SynthesizeMe! Inducing Persona-Guided Prompts for Personalized Reward Models in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 8045–8078. arXiv:2506.05598 doi:10.18653/v1/2025.acl-long.397

[65] Pratik S. Sachdeva, Renata Barreto, Claudia von Vacano, and Chris J. Kennedy. 2022. Assessing Annotator Identity Sensitivity via Item Response Theory: A Case Study in a Hate Speech Corpus. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 1585–1603. doi:10.1145/3531146.3533216

[66] Dorsa Sadigh, Anca D. Dragan, Shankar Sastry, and Sanjit A. Seshia. 2017. Active Preference-Based Learning of Reward Functions. In *Robotics: Science and Systems*.

[67] P. Wesley Schultz, Jessica M. Nolan, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius. 2007. The Constructive, Destructive, and Reconstructive Power of Social Norms. *Psychological Science* 18, 5 (2007), 429–434. doi:10.1111/j.1467-9280.2007.01917.x

[68] Julia Sepúlveda Coelho and Scott A. Hale. 2026. What Do People Actually Want From AI? Mapping Preference Plurality. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 6305–6350. arXiv:2606.06674 doi:10.1145/3805689.3812398

[69] Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. 2025. Language Model Personalization via Reward Factorization. In *Second Conference on Language Modeling (COLM)*. arXiv:2503.06358

[70] Anikait Singh, Sheryl Hsu, Kyle Hsu, Eric Mitchell, Stefano Ermon, Tatsunori Hashimoto, Archit Sharma, and Chelsea Finn. 2026. FSPO: Few-Shot Optimization of Synthetic Preferences Personalizes to Real Users. In *International Conference on Learning Representations (ICLR)*. 109496–109524. arXiv:2502.19312

[71] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2024. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. In *International Conference on Learning Representations (ICLR)*. arXiv:2312.08358

[72] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. Position: A Roadmap to Pluralistic Alignment. In *Proceedings of the 41st International Conference on Machine Learning (PMLR, Vol. 235)*. 46280–46302. arXiv:2402.05070

[73] Jakob Stenseke, Jocelyn Zhu, and Dylan Hadfield-Menell. 2026. Personalizing AI to Align with the Better You. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 9. AAAI Press, 2435–2449. doi:10.1609/aies.v9i4.43318

[74] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to Summarize from Human Feedback. In *Advances in Neural Information Processing Systems*, Vol. 33. arXiv:2009.01325

[75] Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, et al. 2025. Olmo 3. arXiv:2512.13961

[76] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, et al. 2025. 2 OLMo 2 Furious. arXiv:2501.00656

[77] Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2025. HelpSteer2-Preference: Complementing Ratings with Preferences. In *International Conference on Learning Representations*. arXiv:2410.01257.

[78] Zachary Wojtowicz, Ayush Nayak, and Jacob Andreas. 2026. From Weights to Words: Expressing and Editing Preference Model Inferences in Natural Language. arXiv:2607.16232

[79] Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025. CollabLLM: From Passive Responders to Active Collaborators. In *Proceedings of the 42nd International Conference on Machine Learning (PMLR, Vol. 267)*. 67260–67283. arXiv:2502.00640

[80] Wenbin Wu. 2026. Aligning to Illusions: Choice Blindness in Human and AI Feedback. arXiv:2603.08412

[81] Qianqian Xu, Jiechao Xiong, Xiaochun Cao, Qingming Huang, and Yuan Yao. 2019. From Social to Individuals: A Parsimonious Path of Multi-Level Models for Crowdsourced Preference Aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 4 (2019), 844–856. arXiv:1804.11177 doi:10.1109/TPAMI.2018.2817205

[82] Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kris Rose, and Maximilian Nickel. 2026. Cultivating Pluralism In Algorithmic Monoculture: The Community Alignment Dataset. In *International Conference on Learning Representations (ICLR)*. 147076–147215. arXiv:2507.09650

[83] Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen K. Ahmed, and Yu Wang. 2025. Personalization of Large Language Models: A Survey. *Transactions on Machine Learning Research* (2025). arXiv:2411.00027

[84] Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs. In *International Conference on Learning Representations (ICLR)*. 15888–15931. arXiv:2502.09597

[85] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593

A Implementation Details

Data. Community Alignment's released candidates follow a fixed order: on the first turn the plain reply is always the fourth, and the others follow the order in which the model wrote them. We assume that this is the order shown. Annotators' explanations use the released letters: of 11,468 explanations that name exactly one response by letter ("response A") and mention no other letter, 97.2% name the response the annotator chose, against about 25% if the letters had been reassigned at random. This supports the released order only if the interface showed the letters in order down the page, which the dataset's paper does not describe. The 113 explanations that name a response by its position ("the first response") name the chosen one 30% of the time, too few and too loosely worded to settle the question. In MultiPref, 11% of an annotator's choices share a prompt with another of their choices, so a held-out choice can repeat a training item; this may raise MultiPref's personal gain slightly. No choice in the three datasets is part of Olmo 3's synthetic preference data [75], although 4.7% of MultiPref's response pairs share an identical prompt with it (6.8% counting near-identical prompts). In our data, later-turn candidates differ along the same feature directions as first-turn candidates, about half as strongly. Turns with fewer than four candidates, or whose prompt alone exceeds the token cap below, are dropped (91 candidate rows). Held-out conversations are drawn separately for each annotator. In TL;DR, each post is framed as a summarization request and each summary pair is scored once.

Default score. Each checkpoint reads the conversation in its own chat format with its default system prompt: earlier prompts, the responses the annotator chose at earlier turns, the current prompt, and the candidate, which ends with the model's end-of-text token. Sequences are capped at 4,096 tokens by dropping the oldest exchanges, and the shared history is encoded once per turn. Weights are in bfloat16.

Sparse autoencoder. Following WIMHF, the autoencoder has 32 features with on average 4 active per input, a Matryoshka prefix of 8, and an identity activation, so that a feature's sign records whether a candidate has more or less of it than its alternatives. Inputs are rescaled to median norm 1; training uses Adam (learning rate 5×10^{-4}, batch 1,024, 4,000 steps); at inference a feature is active when its pre-activation exceeds a threshold learned in training. Results

are reported for seed 0; seeds 1 and 2 and feature counts of 16, 64 and 128 are robustness checks. Each seed-0 feature is matched to the feature in another seed with which it correlates most strongly across candidates; the Community Alignment features on which our results rest reappear in seeds 1 and 2 with correlations between 0.63 and 0.99. Autoencoders are trained separately for each dataset.

Feature names. Conversations are split 50/25/25 for proposing, selecting and evaluating names. Each of five candidate names is proposed from five turns drawn from the 5% with the largest within-turn spread of the feature, showing the highest-activation candidate as response A and the lowest as response B, with WIMHF's naming prompt. Fidelity pairs take a candidate with a nonzero activation and another candidate from the same turn, in random order; the judge answers A, B or neither from its next-token probabilities. Names are selected on 100 pairs and evaluated on 300 fresh pairs. Over all 32 names, the mean fidelity is 0.36. As shared features, the named features keep 84% of the full embedding's gain over position and default score, matching the 84% of its embedding's gain in AUC over chance that WIMHF reports for its features.

Hand checks of the names. For the first check, 30 of the judge's evaluation calls were drawn from six features with high shared or personal weight, five per feature, and shown with the judge's answer hidden. Our labels agreed with the judge on 21 of 30 ($\kappa = 0.53$); we answered "neither" on 16 and the judge on 12, and when we saw a difference, it ran in the feature's direction 13 of 14 times. For two features, "does not focus on technology, digital tools or data analytics" and "uses a structured list format instead of continuous prose", we saw no difference in any of the five pairs. We restated the first name positively, as "focuses on technology, digital tools or data analytics" with the feature's sign reversed (judge fidelity 0.27, $p < 10^{-5}$), and labeled the ten evaluation pairs in which each of the two features differs most. We picked the feature's side in all 10 list-format pairs, and in all 6 technology pairs in which we picked a side. The same check on the other names we interpret gave 10 of 10 for elaborate literary analysis, 10 of 10 for MultiPref's substantive response instead of refusal, and 6 of 6 for MultiPref's detailed, comprehensive explanation, where we answered "neither" on the other four pairs. An independent second coder labeled all 80 pairs blind; the two of us agreed on 81% ($\kappa = 0.72$), and on the 50 pairs where both picked a side, we picked the same side every time.

Fitting. The shared parameters use a ridge penalty of 1. Per-feature prior variances come from per-annotator estimates of each feature's deviation from the shared weight (damped Newton steps with a weak $N(0, 1)$ prior), pooled by a Paule–Mandel random-effects estimator; variances are floored at 0.005^2. Personal position terms have a $N(0, 0.5^2)$ prior and are split by turn. In the main results they are fitted jointly with the personal feature weights; the prior variances, the learning curve, the analysis by kind of prompt and the minority analysis fit them first and hold them fixed, which is slightly conservative for the feature weights. The permutation null for specificity comes from this sequential fit, whose observed share is 73.7%. A few annotators' Newton steps oscillate when estimating prior variances; for those, the one-dimensional problem is solved exactly with a bounded scalar optimizer. The single-penalty robustness checks replace the per-feature priors with one ridge penalty of 300 on every feature. TL;DR's fit without position terms chooses its penalty, also 300, by cross-validation on 20% of training turns. Beyond the five features whose spread across annotators is significant, the next, concise lists of recommendations, has $p = 0.002$.

Evaluation. Intervals on gains for the shared layers come from 1,000 bootstrap resamples of held-out turns, and intervals on per-annotator gains from 2,000 resamples of annotators (1,000 for the learning curve). Intervals on annotators' weights in Table 3 use 100 resamples of annotators, those on the shares opposed 300, and those on the default's lean 200 resamples of training turns. Person-specificity compares each annotator's own departure with the

average of 200 randomly drawn other annotators' departures, among annotators with at least 10 held-out choices (at least two of the given kind in Table 4). Its permutation null uses 2,000 random derangements of the departures (1,000 in Table 4).

Whom the default works against. A feature qualifies when its name is valid and the 95% bootstrap interval of the default's lean d_j excludes zero. The share opposed is $\Phi\big(-\operatorname{sign}(d_j)\,(a\,d_j + w_j + \mu_j)/\tau_j\big)$, where μ_j and τ_j are the pooled mean and standard deviation of personal weights, with a 95% bootstrap interval over annotators. When τ_j is zero, everyone is on the same side as the crowd, and the feature is reported as shared opposition. The analysis was specified before it was run; robustness checks use own prompts only, no personal position terms, and two random halves of the annotators. For length, the share opposed is 5% and 6% in the two halves and 4.7% on own prompts. Annotators' own weights on the default score are pooled with a Paule–Mandel estimator.

The annotator in Figure 2. The rule, fixed before the figure was made, takes the annotator with the median gain from the personal departure among annotators with at least 80 choices, at least 10 held-out choices, and at least one named feature on which the direction of the personal departure is at least 80% certain.

Simulation. Simulated annotators are drawn from the fitted Community Alignment population (shared weights, pooled spreads of personal weights and position tendencies), with 1,000 per condition, and choose by the conditional logit on real Community Alignment turns. Random comparisons use turns as collected; chosen comparisons use the 10% of turns in which the feature differs most, with candidates in random order. A third condition fits random comparisons without position terms.

Position terms across turns. For the 822 annotators with at least four first and four later turns, we fit turn-split position terms (relative to the first position) on all of an annotator's turns, with the shared model as offset and the $N(0, 0.5^2)$ prior. Across annotators, the first-turn and later-turn terms correlate at 0.43 for the second position, 0.43 for the third and 0.59 for the fourth. Split-half reliabilities, from fitting each annotator's terms on two halves of their conversations and applying the Spearman–Brown correction, are 0.45, 0.44 and 0.77 on first turns and 0.54, 0.54 and 0.71 on later turns, which gives corrected correlations of 0.87, 0.87 and 0.80. Within turns, the default score, centered within each turn, averages +0.41, −0.21, −0.16 and −0.05 for the four positions on first turns with a supplied prompt, +0.28, −0.19, −0.11 and +0.01 on first turns with the annotator's own prompt, and +0.59, −0.12, −0.28 and −0.19 on later turns. With position terms split by turn in the shared layers as well, the default adds 0.008 nats per choice [0.006, 0.011] over position and the shared departure 0.073 [0.067, 0.079], 8.0 times the personal departure.

Figure 1. In panel (a), list format, which Table 3 omits, has a default's lean of +0.003 and an annotators' weight of +0.090; each band is annotators' weight plus or minus 1.28 times the estimated spread of personal weights, from the per-feature priors estimated beyond position tendencies, which is 0.191, 0.098, 0.096, 0.086, 0.071 and 0 for the six features from top to bottom. In panel (b), each share is a layer's gain in log-likelihood per held-out choice divided by the sum of all layers' gains. For Community Alignment the gains are those of Table 2; fitted in the other order, the personal departure adds 0.019 and the position tendency 0.023 beyond it (Section 5.3), which gives shares of 10% and 12%. In MultiPref they are 0.040 (shared position), 0.024 (default), 0.031 (shared departure), 0.003 (position tendency) and 0.008 (personal departure); in TL;DR, with whole posts held out and no position terms, 0.020 (default), 0.045 (shared departure) and 0.002 (personal departure).

Table 3. Where the default leans on named features in Community Alignment, and where annotators lean, both in the scale of annotators' choices. The default's lean is its within-turn regression weight on the feature, converted by the exchange rate, $a\,d$; annotators' weight adds the shared departure, $a\,d + w$. Brackets are 95% bootstrap intervals over annotators. The named features account for 14% of the variation in the default score within a turn, so the default also rewards qualities that the features do not describe.

Feature	Default's lean	Annotators' weight
is longer	+0.018	+0.265 [0.245, 0.287]
provides specific, actionable details and direct answers	+0.014	+0.188 [0.177, 0.201]
focuses on abstract, philosophical or thematic interpretations	−0.023	−0.150 [−0.162, −0.137]
discusses sustainability or the environment	+0.005	−0.145 [−0.163, −0.125]
focuses on technology, digital tools or data analytics	+0.012	−0.120 [−0.138, −0.108]
provides elaborate literary analysis instead of a concise answer	+0.002	−0.048 [−0.058, −0.037]

Table 4. Personal departure by the kind of held-out choice, measured beyond each annotator's position tendency, with one departure per annotator fitted on all training choices. Specificity is the share of annotators, among those with at least two held-out choices of the kind, whose own departure predicts these choices better than the average of others' departures.

Held-out choices	Choices	Personal departure [95% CI]	Specificity
First turn, prompt supplied by the dataset	4,509	+0.019 [0.012, 0.025]	71%
First turn, annotator's own prompt	1,829	+0.004 [−0.001, 0.010]	60%
Later turns, annotator's own prompt	10,590	+0.007 [0.005, 0.009]	64%

Table 5. Recall / precision in identifying simulated annotators whom the default works against, from comparisons drawn at random or from the 10% of turns in which the feature differs most. A simulated annotator is identified when the posterior probability that their weight has the opposite sign to the default's lean is at least 0.8. Percentages in the header are the simulated minorities. The abstract-framing minority does not hold on own prompts.

	Wants shorter answers (5.5%)		Wants abstract framing (4.8%)	
Comparisons	Random	Chosen	Random	Chosen
25	2% / 100%	37% / 92%	0% / –	2% / 100%
50	5% / 100%	39% / 94%	0% / –	4% / 100%
100	13% / 88%	70% / 95%	0% / –	9% / 80%
200	38% / 100%	77% / 95%	3% / 100%	29% / 100%
400	43% / 100%	74% / 98%	13% / 75%	64% / 91%

Table 6. Robustness of the personal departure. The gain is in log-likelihood per held-out choice beyond the shared model and the annotator's own position tendency, except in the TL;DR row without position terms. Specificity is the share of annotators whose own departure predicts their held-out choices better than the average of other annotators' departures. The permutation null's 95th percentile is 58.6% for the main specification and 64.3% for TL;DR without position terms.

Specification	Gain per choice [95% CI]	Specificity
Main: Olmo 3, per-feature priors	+0.0091 [0.0066, 0.0116]	74.8%
Shared model fitted on a disjoint half of the annotators	+0.0083 [0.0045, 0.0120]	76.4%
OLMo 2 default	+0.0106 [0.0080, 0.0134]	77.0%
MultiPref	+0.0084 [0.0041, 0.0130]	73.4%
TL;DR, whole posts held out	+0.0023 [0.0003, 0.0049]	78.6%
TL;DR, whole posts held out, without position terms	+0.0024 [0.0013, 0.0034]	80.0%
Single ridge penalty, autoencoder seed 0	+0.0036 [0.0026, 0.0046]	66.7%
Single ridge penalty, seeds 1 and 2	+0.0041, +0.0037	67.8%, 67.3%
Single ridge penalty, 16, 64 and 128 features	+0.0036, +0.0030, +0.0025	69.9%, 67.1%, 66.5%