Learning interpretable descriptions of human preferences

Rajiv Movva [1] **Emma Pierson** [1]

Abstract

Language model preference datasets are designed with desired goals (helpfulness, harmlessness, *etc.*), but it is unclear which attributes are ultimately encoded in the collected datasets. Given a preference dataset, we propose a general method to extract natural language concepts that raters tend to favor (e.g., "responses with numbered lists"). We use sparse autoencoders to map response text embeddings to an interpretable feature basis, and then perform feature selection to identify concepts that predict preferences. We apply our method to six widely-studied RLHF datasets: across datasets, just 5-10 natural language concepts account for about 30% of the preference signal that is predictable from blackbox embeddings. We find preferences—such as disfavoring uncertainty or follow-up questions—that may lead to undesirable downstream model behaviors. We discuss how our method enables intervening on undesirable preferences.

1. Introduction

Preference datasets are one of the main levers to encode values in language models. For example, in HH-RLHF (Bai et al., 2022), annotators are instructed to choose which of two possible chat replies is more helpful and harmless, and the LLM is fine-tuned via reinforcement learning to produce replies more like the chosen ones. Regardless of these stated design goals, it is critical to study *revealed preferences*: what are the characteristics of the responses humans ultimately choose? Answering this question will help us better understand which model behaviors, both desirable and undesirable, are traceable to preference data.

Prior approaches to studying human preferences have been motivated by specific observations about model outputs. For example, Singhal et al. (2024) observe that RLHF increases output length, and study the extent to which this is attributable to datasets; Sharma et al. (2023) perform a similar analysis for sycophancy, and Hosking et al. (2024) for assertiveness. However, in each case, researchers must pre-specify the concepts being studied. To understand preference learning more completely, it is critical to describe all such patterns rather than relying on intuition.

In this work, we provide a fully data-driven description of what distinguishes chosen and rejected responses in preference datasets. We build on prior work that trains sparse autoencoders to produce interpretable text embeddings, which are then used for *hypothesis generation* (Movva et al., 2025). Specifically, our method outputs natural language concepts—such as "the response expresses uncertainty"—that vary in prevalence between chosen and rejected responses. Importantly, our method requires no prior specification, so we are able to produce concepts which may be difficult to foresee. We are also able to control for features which are already known, such as length (Singhal et al., 2024).

We apply our method to six datasets used for RLHF, including datasets with LLM responses evaluated by humans (HH-RLHF, Chatbot Arena, PRISM); LLM responses evaluated by an LLM (UltraFeedback); and human responses evaluated by humans (Reddit, StackExchange). We learn shared preferences across datasets, such as favoring numbered lists and disfavoring uncertainty. Other preferences are dataset-specific, such as favoring both-sides responses to moral questions in PRISM or disfavoring illegal activites in HH-RLHF. Some preferences flip across datasets, such as personal experiences with first-person pronouns being favored on Reddit and disfavored in PRISM. Importantly, some preferences we observe may be undesirable: for example, disfavoring uncertainty may reward LLM overconfidence (Zhou et al., 2024).

We show that human preferences on each dataset are relatively well-described by a small number of these natural language descriptions. We annotate the presence or absence of each concept on a large number of response pairs to validate that they predict the preferred response. With just $\sim$10 binary variables per dataset, we achieve about one-third of the gain in AUC contributed by blackbox text embeddings relative to a length baseline. We discuss the possible applications of these findings towards more interpretable control over how models are tuned using preference data.

[1]UC Berkeley. Correspondence to: Rajiv Movva <rmovva@berkeley.edu>.

ICML 2025 Workshop on Models of Human Feedback for AI Alignment, Vancouver, Canada. 2025.

2. Method

We study *pairwise preference datasets*

$$\left\{\left(P, R_A^{(i)}, R_B^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N},$$

where a single datapoint includes a prompt P, two possible responses R_A, R_B, and a label y, where $y = 1$ if R_A is preferred over R_B and $y = 0$ if R_B is preferred. Our goal is to identify *natural language concepts* that, when present in one response but not the other, are predictive of y.

To generate such concepts, we would like representations of response text that both (a) contain features with clear interpretations and (b) predict preferences as well as possible. These two properties usually trade off: either we can compute interpretable features (like n-grams or topics), or we can use blackbox text embeddings. A recent method, HypotheSAEs (Movva et al., 2025), helps resolve this challenge by using sparse autoencoders (SAEs). The method proceeds as follows:

1. Learn interpretable representations by training an SAE on response text embeddings.
2. Select a sparse set of features in the SAE representation which predict the target, y.
3. Automatically interpret these features using an LLM.

The first step of training the SAE is critical, because individual features in dense neural embeddings are generally not interpretable. In the SAE, individual features correspond to individual concepts (Bricken et al., 2023). Therefore, step 1 enables the downstream feature explanation in step 3.

The procedure starts from text embeddings and outputs a list of features which predict y, along with their coefficients and natural language interpretations. For implementation, we largely follow Movva et al. (2025): for step 1, we train a top-k SAE with $M = 512$ total neurons and $k = 32$ active neurons per input on 1536-dim embeddings of the responses computed by OpenAI `text-embedding-3-small`. Then, for responses $R_A^{(i)}, R_B^{(i)}$ with dense embeddings $x_A^{(i)}, x_B^{(i)}$, the SAE yields length-2048 sparse representations $z_A^{(i)}, z_B^{(i)}$, each with 32 nonzero values. For step 2, note that we are predicting a single outcome given a pair of arbitrarily-ordered responses. We therefore predict y using the delta $z_\Delta = z_A - z_B$, where each dimension of z_Δ captures how much more strongly R_A contains the feature than R_B. We fit an L1-regularized logistic regression (Lasso) to predict y from z_Δ, yielding a sparse set of coefficients for each dimension in z. In step 3, we interpret the dimensions in z with nonzero coefficients by prompting GPT-4.1 with a sample of texts that strongly activate that dimension. Besides those specified, we use default hyperparameters.

Improving feature selection. To select features of interest, we modify feature selection in two ways. First, many features of the responses relate to specific topics, such as "Unix/Linux command-line usage" or "describes visual art." We are more interested in stylistic or value-laden attributes, so we prompt GPT-4.1 to filter out features whose descriptions are about specific topics (prompt in Appendix, Figure 4). This step reduces the SAE feature count from 512 to 190. Second, prior work has shown that response length is highly predictive of preferences; we would like to focus on attributes which remain predictive after controlling for length, so we include the difference in word count as an unpenalized covariate in the Lasso.

Datasets. We study six preference datasets. HH-RLHF (Bai et al., 2022), Chatbot Arena (Zheng et al., 2023), and PRISM (Kirk et al., 2024) each contain human-written prompts, responses generated by various open- and closed-weight LLMs, and human ratings. UltraFeedback (Cui et al., 2023) contains a combination of human- and LLM-generated prompts, with LLM responses and GPT-4 ratings; we use the Argilla binarized version. Reddit and Stack, originally assembled for the Stanford Human Preferences dataset (Ethayarajh et al., 2022), consist of human-written posts ("prompts") and comments ("responses") from informational subreddits (e.g., /r/AskScience) and various StackExchange forums, respectively. Comment A is considered "preferred" over comment B if, despite being posted later, it received more upvotes. Each of these datasets has been used to fine-tune and/or evaluate LLMs, usually via RL.

3. Results

SAE features predict preferences. First, we evaluate how well the SAE features can predict human preferences. We use both the full SAE representation (filtered to the 190 non-topic-specific features), as well as the Lasso with various regularization strengths to reduce feature count. If the SAE features are not predictive, it's unlikely that the natural language descriptions of the features will be predictive either. As a baseline, we compare to predicting preferences using a blackbox representation of the prompt and the response. To do so, we compute the OpenAI embeddings E_A, E_B of the prompt concatenated to the response, and predict y from E_Δ. In each case, we standardize all features and include the difference in length as an unpenalized control.

Figure 1 reports results. Across datasets, using the SAE instead of the blackbox embedding retains substantial predictive performance on top of length. On average, using the SAE yields an AUC of 0.71, lower than the blackbox embeddings (0.76) but considerably higher than using length alone (0.66). Further, we find that even strongly L_1-regularized models—using a median of just 8 features with $\lambda = 0.02$—

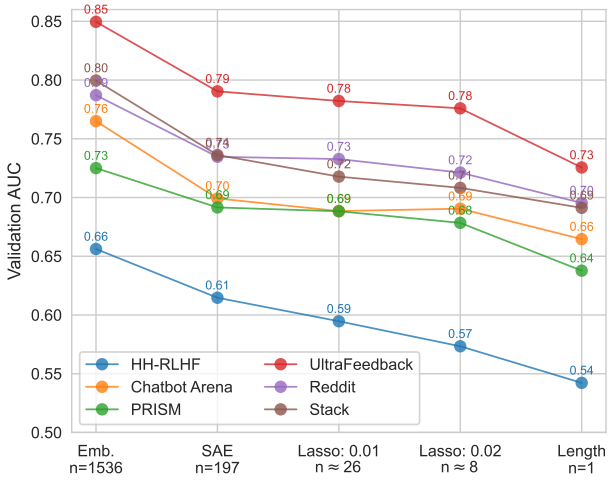

Figure 1. Validation AUC of a logistic regression trained to predict the preferred response using the difference in response length, and either: the text embedding of the prompt concatenated with the response ("Emb."); the SAE representation, filtered down to 190 non-topic-specific features; the SAE with 0.01 or 0.02 L_1 regularization; and length-only. For the L_1 models, we report the median number of nonzero features across the 6 datasets.

achieve an AUC of 0.69. These results show that a small number of SAE features contain substantial predictive value. Specifically, with about 8 interpretable features, we can explain about one-third of the gain in AUC contributed by a 1536-dimensional blackbox text embedding over length. Next, we examine these features' interpretations.

Explaining human preferences. In Figure 2, we display all features that are selected on more than one dataset by the dataset-specific Lasso classifiers with $\lambda = 0.02$. There are 11 such features. We display features using their natural language explanations, which are produced by prompting an LLM with examples of responses that strongly activate the feature. Because length is an unpenalized covariate in these regressions, these features are selected after controlling for length. Each cell is labeled with the percent change in the odds of a response being preferred given a 1-standard deviation increase in the feature. (Features that are selected on only a single dataset are in the Appendix, Figure 3.)

Features with the same direction of effect across datasets suggest general human preferences. We observe several such examples, revealing three findings about consistently favored and disfavored attributes:

Structured, in-depth responses are favored on Arena, PRISM, and UltraFeedback. One feature (row 7, from top) specifically captures multi-paragraph answers using headings or lists. A 1-SD increase in this feature increases the odds that response is preferred by ∼30%. This supports a prior analysis from the Chatbot Arena team (Li et al., 2024), who explicitly studied the effect of markdown tag count on Arena preferences. A second feature (row 6), with a large effect on PRISM, captures in-depth answers that include definitions. Notably, these strong effects exist after controlling for length.

Asking follow-up questions or expressing uncertainty is disfavored, especially in HH-RLHF, Reddit, and Stack (rows 1, 2, 5, and 11). For example, one feature is "asks the user for more information..." (row 1); on HH-RLHF, an example response which activates this feature is "I don't have any information about that. Can you explain more?" Such responses are perhaps disfavored because they are not immediately helpful to the user. However, disfavoring these attempted abstentions could reward overconfidence in model outputs. On Reddit and Stack, we observe that open-ended suggestions (row 2), asking for clarifications (row 5), and suggestions to seek professional advice due to uncertainty (row 11) are disfavored. If used as LLM training signals, these features pose similar risks of overconfidence (Zhou et al., 2024). We conclude that this dispreference for back-and-forth clarification holds across datasets both with LLM-written and human-written replies.

Negating or refusing to engage with the prompt is disfavored on multiple datasets, and manifests in multiple ways (rows 3, 8, 9). On Arena and PRISM, harmful prompts are often refused by apologizing first, while on UltraFeedback, refusals often begin with "As an AI assistant..." Interestingly, on HH-RLHF, refusals are not penalized. Instead, we see an explicit disfavoring when the model *does* provide details on illegal activities—perhaps both because HH-RLHF was collected using a more easily-jailbroken model, and because the hired annotators were explicitly told to prefer harmlessness (Ganguli et al., 2022).

Finally, we note an interesting example where the direction of a preference *shifts* across datasets (row 10). On Reddit, including personal opinions or subjective judgments with first-person pronouns is favored (+4.2%); in PRISM, it is strongly disfavored (-13%). In this case, it is likely that while first-person experiences help convince users on Reddit, humans disfavor overly anthropomorphized responses from LLMs. This example illustrates how human preferences can depend on context—underscoring the importance of explaining the patterns present in a dataset in an open-ended manner. (Other notable features that are specific to individual datasets are present in Figure 3, such a preference for "multiple perspectives on a controversial topic" in PRISM.)

Natural language descriptions of features predict preferences. Next, we show that *natural language descriptions* alone can predict human preferences. To do so, we ignore the actual feature values produced by the SAE and used

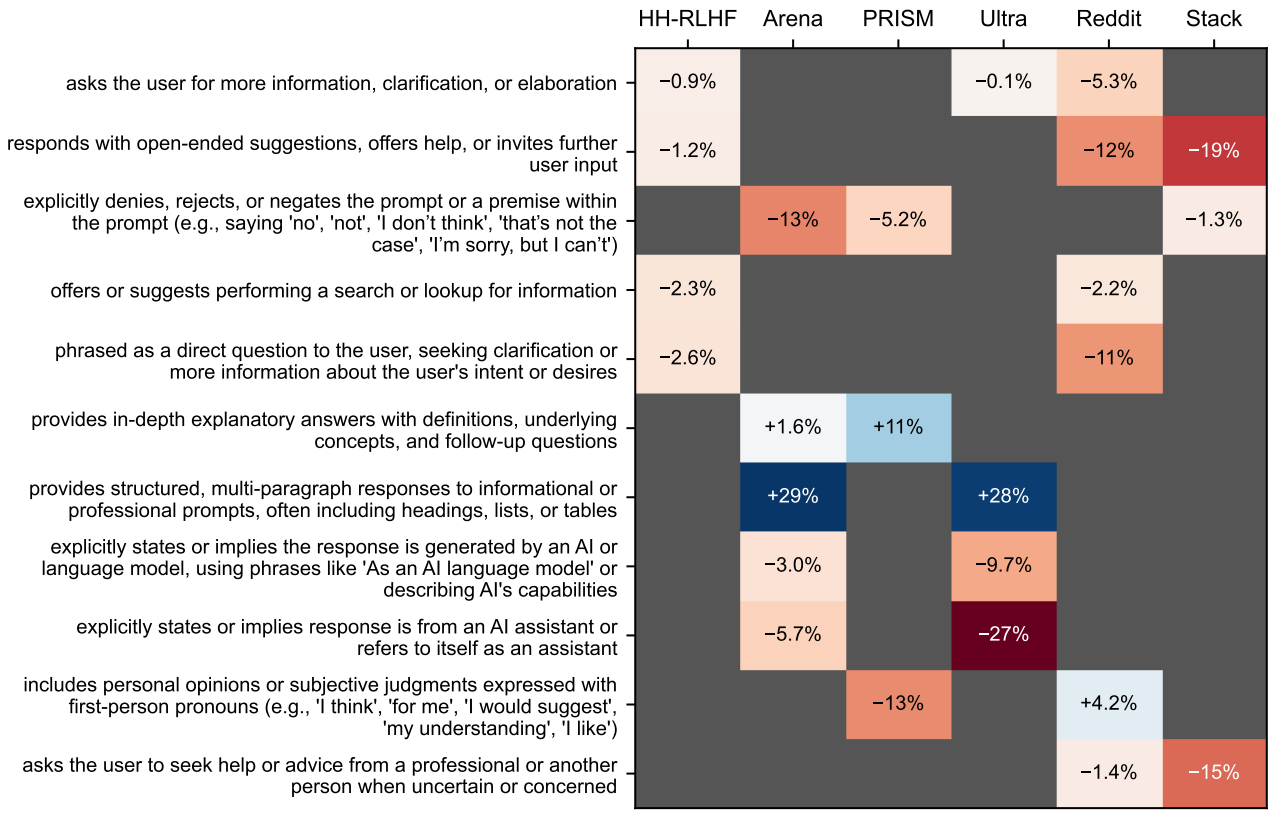

Figure 2. A heatmap of features and how strongly they affect preferences across six datasets. Blue features are favored, red disfavored. We include features selected by Lasso in at least two out of the six datasets. Feature descriptions are generated via autointerpretability, following Movva et al. (2025). The cell values are the percentage changes in the odds of a response being preferred given a 1-standard deviation increase in the feature, with all other features held constant. The effect sizes are computed while controlling for response length.

in the Lasso. We use only the generated concepts—i.e., the natural language explanations—of each selected feature, and we annotate a held-out set of response pairs with `GPT-4o-mini`. Each annotation is binary, and represents whether the response contains a concept or not, such as "asks the user for more information or clarification." For each dataset, we include up to 10 concepts sorted by Lasso coef. value, and we compute their annotations on the heldout split (at most 10% of the dataset, capped at 5000 samples).

Collectively, these concepts predict preferences as well as the SAE features that were used to produce them. The average AUC using the concept annotations is 0.69, equivalent to the average Lasso AUC with $\lambda = 0.02$. This result confirms that a small set of natural language descriptions can explain a substantial portion of predictable preference variation.

We validate our earlier findings regarding favored and disfavored attributes using the concept annotations. Since the concept annotations are binary, we define a simple metric, *preference rate*: when a concept is present in response A but not response B, how often is response A preferred to B? We replicate earlier findings: for example, requests for clarification are heavily dispreferred (e.g., 33% preference rate on Reddit, 29% on UltraFeedback), as are expressions of uncertainty (46% preference rate on HH-RLHF, 33% on Stack). Table 1 displays significant concepts and their preference rates for each dataset.

4. Discussion

We use SAEs to decompose preference datasets into succinct lists of concepts which raters prefer, and apply this method to six common datasets. We find several dataset-agnostic and dataset-specific preferences. Notably, responses requesting clarification or expressing uncertainty are disfavored, which may contribute to LLM overconfidence.

Our method allows practitioners to *intervene* in the design and use of preference datasets. First, dataset designers can explicitly encourage annotators to avoid certain pitfalls of standard preference data (e.g., favoring assertiveness and confidence). Second, for already-collected datasets, our method can reveal the datapoints responsible for each preference, allowing principled data curation (e.g., removing datapoints that disfavor uncertainty).

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2, 2023.

Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z., and Sun, M. UltraFeedback: Boosting Language Models with High-quality Feedback, October 2023.

Ethayarajh, K., Choi, Y., and Swayamdipta, S. Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 5988–6008. PMLR, June 2022.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., Hatfield-Dodds, Z., Henighan, T., Hernandez, D., Hume, T., Jacobson, J., Johnston, S., Kravec, S., Olsson, C., Ringer, S., Tran-Johnson, E., Amodei, D., Brown, T., Joseph, N., McCandlish, S., Olah, C., Kaplan, J., and Clark, J. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned, November 2022.

Hosking, T., Blunsom, P., and Bartolo, M. Human Feedback is not Gold Standard, January 2024.

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., and Hale, S. A. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models, April 2024.

Li, T., Angelopoulos, A. N., and Chiang, W.-L. Does style matter? Disentangling style and substance in Chatbot Arena. https://lmsys.org/blog/2024-08-28-style-control, August 2024.

Movva, R., Peng, K., Garg, N., Kleinberg, J., and Pierson, E. Sparse Autoencoders for Hypothesis Generation, March 2025.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards Understanding Sycophancy in Language Models. https://arxiv.org/abs/2310.13548v4, October 2023.

Singhal, P., Goyal, T., Xu, J., and Durrett, G. A Long Way to Go: Investigating Length Correlations in RLHF, July 2024.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, December 2023.

Zhou, K., Hwang, J. D., Ren, X., and Sap, M. Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty, July 2024.

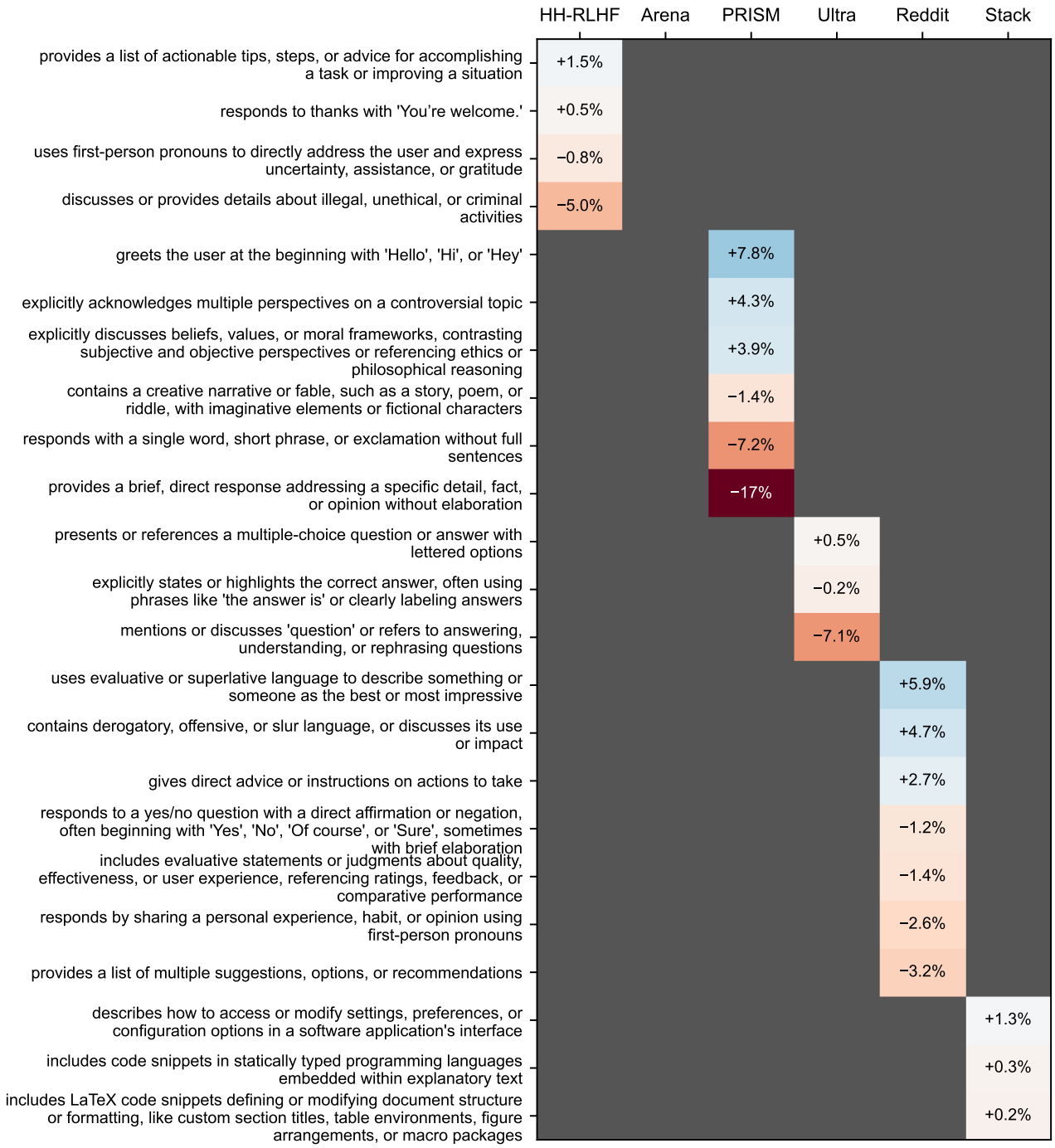

Figure 3. Counterpart to Figure 2, with features that are selected in a single dataset. Blue features are favored, red disfavored. Feature descriptions are generated via autointerpretability, following Movva et al. (2025). The cell values are the percentage changes in the odds of a response being preferred given a 1-standard deviation increase in the feature, with all other features held constant. The effect sizes are computed while controlling for response length.

Table 1. Concepts on each dataset with a preference rate that significantly differs from 50% ($p < 0.05$ after Bonferroni correction). When the concept is present in exactly one response, "Pref." is how often that response is preferred. All concepts are abbreviated for space.

Dataset	**Concept:** Red: ↓ dispreferred Green: ↑ preferred	**Pref. (%)**
HH-RLHF	Recipe with ingredients and cooking instructions	75
	Structured step-by-step advice or recommendations	62
	Open-ended suggestions and offers to help	46
	Uses first-person pronouns expressing uncertainty/gratitude	46
	Asks for more information or elaboration	45
	Direct questions seeking clarification from user	44
	Offers to help search for information or resources	41
	Discusses illegal, unethical, or criminal activities	29
Arena	Step-by-step mathematical solutions with formulas	81
	Structured responses with headings, lists, or tables	71
	Multi-paragraph in-depth explanatory answers	63
	Explicitly denies or rejects the prompt premise	39
	Refers to itself as AI, language model, or chatbot	34
	Explicitly expresses inability to answer using "I'm sorry"	30
PRISM	Multi-paragraph explanatory answers with definitions	70
	Acknowledges multiple perspectives on controversial topics	66
	Discusses beliefs, values, or moral frameworks	66
	Short, concise single-sentence responses (< 10 words)	21
UltraFeedback	Structured multi-paragraph responses with sections	80
	Instructions on writing or formatting content	70
	Explicitly states the correct answer	61
	Multiple-choice questions with letter options	61
	Single word or short phrase responses	41
	Mentions or discusses the concept of "question"	38
	Offers help or apologizes for inability	35
	Refers to itself as AI or discusses capabilities	34
	Asks for more information or elaboration	29
Reddit	Lists of examples (bulleted or numbered)	71
	Evaluative language describing things as "best" or superior	58
	Personal opinions or advice with subjective experiences	56
	Asks for more information or elaboration	33
	Direct questions seeking user clarification	31
Stack	Step-by-step software/OS instructions	67
	Factual/instructional information with technical guidance	66
	Technical hardware/electronics explanations	66
	Source code snippets in programming languages	64
	Factual information with context and elaboration	63
	Open-ended suggestions and offers for help	35
	Expresses uncertainty with "I'm not sure" or speculation	33

The following text describes a feature of a response from an AI chatbot or a human online forum user to a user's prompt. Please determine whether the feature is a general, stylistic, or value-laden attribute which may influence perception of the response's quality.

Examples of features that DO influence quality (output YES):

- 'includes numbered lists or sections to convey technical information' (a stylistic choice)
- 'explicitly mentions limitations or inability to provide certain types of assistance or information' (requires value judgements to decide what information cannot be provided)
- 'uses words that express uncertainty, like 'maybe' or 'perhaps" (hedging vs. assertiveness is a stylistic choice)

Examples of features that do NOT influence quality (output NO):

- 'mentions rocks, stones, or minerals in a factual or descriptive context'
- 'discusses acid, acidity, or acid-related topics in detail'
- 'uses words starting with the letter 'y' or 'Y"

These features on their own do not influence response quality.

Edge cases (topic-specific but still capture general attributes – output YES):

- 'provides a step-by-step list of instructions to solve a programming-related issue' (step-by-step lists are stylistic)
- 'mentions specific years or references to time periods in history' (mentioning a specific date is stylistic)
- 'provides specific numerical or quantitative recommendations, such as dosages, servings, or proportions, often with units like mg, cups, or sprays' (providing specific numbers is stylistic)
- 'provides advice, examples, or descriptions related to committing crimes or illegal activities' (providing advice about committing crimes is value-laden)

Instructions: Output 'YES' if the feature describes a general or stylistic attribute that could apply to many situations. Output 'NO' if the feature focuses on a specific topic. Do not output anything else.

Feature: {feature}
YES or NO:

Figure 4. Prompt for evaluating whether features describe general, stylistic, or value-laden attributes that may influence response quality perception.