COMPO: Community Preferences for Language Model Personalization

Sachin Kumar[♣*] **Chan Young Park**[♡*]
Yulia Tsvetkov[♡] **Noah A. Smith**[♡,♢] **Hannaneh Hajishirzi**[♡,♢]
[♣]The Ohio State University, Columbus OH
[♡]University of Washington, Seattle WA
[♢]Allen Institute for AI, Seattle WA
kumar.1145@osu.edu, chanpark@cs.washington.edu

Abstract

Conventional algorithms for training language models (LMs) with human feedback rely on preferences that are assumed to account for an "average" user, disregarding subjectivity and finer-grained variations. Recent studies have raised concerns that aggregating such diverse and often contradictory human feedback to finetune models results in generic models that generate outputs not preferred by many user groups, as they tend to average out styles and norms. To address this issue, we draw inspiration from recommendation systems, and propose COMPO, a method to personalize **p**reference **o**ptimization in LMs by contextualizing the probability distribution of model outputs with the *preference provider*. Focusing on group-level preferences rather than individuals, we collect and release COMPRED, a question answering dataset with **com**munity-level **p**references from **Red**dit. This dataset facilitates studying diversity in preferences without incurring privacy concerns associated with individual feedback. Our experiments reveal that conditioning language models on a community identifier (i.e., subreddit name) during preference tuning substantially enhances model performance. Conversely, replacing this context with random subreddit identifiers significantly diminishes performance, highlighting the effectiveness of our approach in tailoring responses to communities' preferences.[1]

1 Introduction

Language models have become ubiquitous in user-facing applications, offering an opportunity for refinement through user feedback (Ouyang et al., 2022). A common approach to this refinement process is preference tuning, where users indicate their preference between two model-generated outputs, which is then used to adjust the model weights with the aim of generating outputs generally preferred by humans (Christiano et al., 2017; Rafailov et al., 2024). However, annotating preferences is inherently *subjective* (Kirk et al., 2023). Current methodologies typically aggregate these (often contradicting) preferences without considering individual differences, resulting in models tuned towards a hypothetical "average" user (Bakker et al., 2022; Chakraborty et al., 2024).

Previous research has explored different methods to address this issue. Some approaches focus on value pluralism (Sorensen et al., 2024a,b; Bakker et al., 2022)—aiming to present balanced views for inputs where different users may hold different values, such as subjective or opinion-based questions. Work on system instructions (Achiam et al., 2023) allows users to verbalize their preferences as part of the input, such as preferred content and style. Additionally, there have been efforts to categorize human feedback into underlying factors such as demographics, culture, and stylistic preferences that advocate for factor-specific fine-tuning (Jang et al., 2023). However, preferences can be nuanced and often implicit. Verbalizing them or factorizing them into interpretable dimensions may not always be feasible or practical in user-facing applications (Cunningham and De Quidt, 2022).

In contrast to prior work which tends to separate the preference provider from the preferences, in this work, we propose to model the provider directly. Inspired by research in recommender systems (Zhang et al., 2019), we propose to contextualize LMs with information about the users providing the preferences. That is, instead of $p(\mathbf{y} \mid \mathbf{x})$ where $\mathbf{x}$ and $\mathbf{y}$ are user input and model output, respectively, we propose to model $p(\mathbf{y} \mid \mathbf{x}, \text{user})$ where user provides the preference. Given a preference dataset marked with this additional information, we make simple modifications to the supervised fine-tuning and preference tuning stages to incorporate

[*]Equal contribution
[1]Data: https://huggingface.co/datasets/allenai/compred, Code: https://github.com/allenai/compred

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8246–8279
April 29 - May 4, 2025

this context (§3).

Since gathering extensive user-specific preference data is infeasible for academic research, we validate our methodology by training models customized to Reddit communities (i.e., subreddits). Reddit's post-comment format naturally supports building question-answering models where posts can serve as model inputs and comments as outputs, respectively. Additionally, subreddit users vote on others' comments, allowing us to use collective upvotes as a proxy for community preference. Reddit consists of thousands of subreddits often discussing similar topics but different participants and different upvoting patterns. As a result, this data offers a readily available resource of diverse preferences to learn from. We refer to this approach as **Com**munity **P**reference **O**ptimization (COMPO).

Our aim is to study whether incorporating the community's context during preference tuning results in models generating responses tailored towards the communities' preferences. To that end, we collect and release COMPRED, a preference dataset comprising five different groups of subreddits (covering themes related to gender, politics, history, science, and finance). Each group constitutes subreddits that discuss similar topics but differ from each other in values or norms and thus diverge in preferences (e.g., `r/askliberal` vs. `r/conservative` under politics). With experiments using direct preference optimization (DPO; Rafailov et al., 2024), we find that, across all domains, adding subreddit context indeed leads to more tailored responses, which are preferred both by human annotators and automated metrics. Conversely, conditioning on the wrong subreddit yields inferior results, indicating that generating outputs preferred by a different user is detrimental to the current user's experience. Ultimately, our work introduces **a novel dataset** and **method to personalize LMs to fine-grained communities**, paving the way towards personalized and adaptive AI assistants.

2 Related Work

Training LMs to align with human preferences While human annotations have been used in NLP for decades, aligning LMs with human feedback (RLHF) marked a key paradigm shift in the field (Ouyang et al., 2022; Rafailov et al., 2024). Yet popular alignment techniques ignore the subjectivity inherent in human preferences (Davani et al., 2022; Kirk et al., 2023, 2024b). Aggregating such preferences risks dilution, leading to models optimized towards a "generic" human that may not satisfy any particular user group well (Kirk et al., 2024a). Prior work in addressing these issues includes the investigation of value pluralism (Sorensen et al., 2024b) that attempts to summarize different perspectives, particularly in contexts where users' values diverge (Bakker et al., 2022). Another line of work has focused on factorizing subjectivity into explainable dimensions (such as style and complexity, among others) (Jang et al., 2023) or proposed that users can personalize their models by simply verbalizing their preferences in the instructions, such as desired content and style (Achiam et al., 2023). It is, however, not always feasible or practical for users to articulate their preferences (Cunningham and De Quidt, 2022). Most closely related to our work is research on incorporating social and cultural context in NLP systems (Solaiman and Dennison, 2021; Qiu et al., 2022; Bang et al., 2023; Kumar et al., 2024). Li et al. (2024) in particular propose a similar paradigm to ours, but their experiments are limited to summarization tasks with a small set of annotators, while we conduct large-scale experiments on a broader range of communities.

Personalization in Recommender Systems Our work draws parallels to recommender systems where personalization serves to match the right types of services, products, or content to the right users intended to improve user engagement whereas showing every user the same recommendation is undesirable. Collaborative filtering, a widely adopted technique in recommender systems, matches different users with similar tastes (Koren et al., 2021). While early work in this space relied on learning linear separable or interpretable features for each user and product (Rendle, 2010), those approaches have been surpassed by neural network-based methods to learn user and product embeddings as a means to capture the underlying preferences and intrinsic characteristics of users and items (Mnih and Salakhutdinov, 2007; Zhao et al., 2023). We draw inspiration from this line of work to model preference providers directly in language models.

Preference Tuning Datasets High-quality datasets have been the primary driver of advances in NLP. While several human-annotated preference datasets are publicly available (Bai et al., 2022;

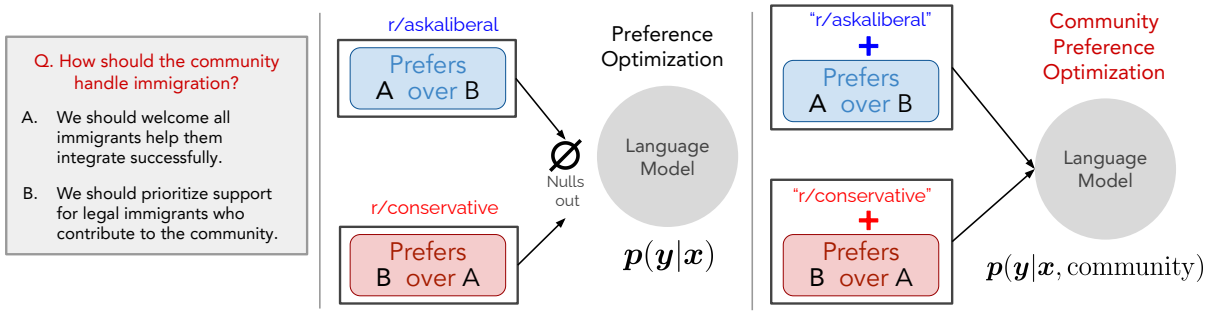

Figure 1: **Conceptual overview of community preference optimization**: When asked about immigration, two communities may prefer different answers (A and B). Conventional preference optimization aggregates these conflicting responses, often averaging them out or reflecting the majority view. Our proposed community preference optimization incorporates subreddit-specific contexts into the model, tailoring outputs to align with the distinct norms and values of individual communities.

Nakano et al., 2021; Kaufmann et al., 2023, *inter alia*), they have been primarily focused on aspects such as helpfulness and harmlessness and are collected with a goal of obtaining high agreement among annotators or between researchers and annotators (Ziegler et al., 2019; Stiennon et al., 2020) rather than embracing subjectivity. Our work is related to Jang et al. (2023), who collect a synthetic data of factor-based preferences using GPT-4 as a judge; we rely on natural existing community-based preferences. Most closely related to our work is Ethayarajh et al. (2022, SHP) who also collect a Reddit based preference dataset. They largely focus on knowledge-seeking subreddits aimed at improving general purpose alignment. Our goal is modeling diversity in preferences; hence, we curate a different set of subreddits where divergences are more prevalent.

3 Community Preference Optimization

To build conversational or instruction following language models $p_\theta(\mathbf{y} \mid \mathbf{x})$, a typical pipeline starts with pretraining, followed by supervised finetuning (SFT) to obtain p_{sft}, and finally preference tuning. SFT requires a dataset with instances of the form $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x}$ is the input query in natural language, and $\mathbf{y}$ is the expected output. Preference tuning is done with a dataset comprising prompts, preferred answers, and dispreferred answers, $(\mathbf{x}, \mathbf{y}_j, \mathbf{y}_k)$. While the standard recipe for preference tuning relies on reinforcement learning (RLHF), in recent work, Rafailov et al. (2024) proposed direct preference optimization (DPO), which implicitly optimizes the same objective as RLHF but offers higher stability. Hence, we adopt this framework for all experiments in this work.[2] DPO's training objective is as follows

$$-\mathbb{E}_{(\mathbf{x},\mathbf{y}_j,\mathbf{y}_k)}\left[\log\sigma\left(\beta\log\frac{p_\theta(\mathbf{y}_j|\mathbf{x})}{p_{\text{sft}}(\mathbf{y}_j|\mathbf{x})} - \beta\log\frac{p_\theta(\mathbf{y}_k|\mathbf{x})}{p_{\text{sft}}(\mathbf{y}_k|\mathbf{x})}\right)\right]$$

where σ is the sigmoid function and β is a hyperparameter. The loss is optimized with respect to θ where p_θ is initialized with p_{sft}, which is often referred to as the reference model, but the latter is kept frozen during this stage.

Our goal is to build a community-personalized model, $p_\theta(\mathbf{y} \mid \mathbf{x}, r)$, where r is the subreddit to which the model will be personalized towards. To train this model, we collect a preference dataset from Reddit along with the subreddit: tuples of the form $(r, \mathbf{x}, \mathbf{y}_j, \mathbf{y}_k)$. The subreddit r is represented by its name (such as `r/conservative`).[3] The same pair of answers $(\mathbf{y}_j, \mathbf{y}_k)$ may have different preferences depending on the subreddit r (see Figure 1). To incorporate r into the pipeline, we concatenate the subreddit name in front of the input text $\mathbf{x}$ for both the SFT and DPO stages. We refer to this approach as community preference optimization. In what follows, we detail our data collection strategy.

4 COMPRED: A Dataset of Diverse Community Preferences from Reddit

The Reddit platform is divided into community forums known as subreddits. The majority of

[2]Our goal in this work is to understand the effect of personalization, which is orthogonal to the choice of the preference tuning approach. Our method can, in principle, be applied to any preference tuning approach.

[3]While special tokens for each r may also be considered for this task, prior work in personalization in NLP has shown that choosing tokens from the existing vocabulary can perform just as well without blowing up the vocabulary size when the number of users increases (Mireshghallah et al., 2021).

subreddits are created based on discussion topics; however, new subreddits are often formed where a subset of the community diverges from the existing ones. These divergences stem from multiple factors, including topic specialization, user demographic differences, value disparities, and different community norms (Hessel et al., 2016; TeBlunthuis et al., 2022). As a result, the preferences of the communities for what discussions they encourage are different. For instance, as illustrated in Figure 1, the subreddits `r/askliberal` and `r/Conservative` have different values and the same two comments posted in both forums to the same post will receive vastly different upvotes. Note that while some subreddits explicitly encode their rules and norms, these preferences are implicit and more nuanced; the members generally do not verbalize their reasons for upvoting or downvoting a comment. This setup mirrors a chat-based LM where a user may not be able to or do not explicitly verbalize their preferences.

Our work capitalizes on these factors to collect diverging preference datasets across different subreddits. Each instance in our dataset consists of the subreddit name, a question, a preferred answer, and a dispreferred answer. We begin with a collection of threads. Each thread consists of a main post q posted in the subreddit r along with $N \geq 0$ comments made by other users: $\{(a_1, u_1, t_1), \ldots, (a_N, u_N, t_N)\}$ where a_i denotes the comment text posted at time t_i with u_i being the number of upvotes it receives.[4] For each pair of comments a_i and a_j in this thread, we add to our dataset a preference instance (r, q, a_i, a_j) if $t_i > t_j$ and $u_i > u_j$. That is, the comment a_i is preferred over a_j if it was posted after a_j and still received more upvotes than the latter. We adopt this strategy following the Stanford Human Preference dataset (Ethayarajh et al., 2022) which argues that using the post time along with the upvote counts provides a more robust signal of preference than just using upvotes as it avoids recency bias. Early comments receive higher upvotes on average due to more exposure time, but if a comment has received more upvotes despite being posted later, it is a clearer signal of preference.

Notably, preferences are conditional on the subreddit in which the post was made. This means that answering the same prompts with same answers may elicit different preferences in distinct subreddits. Our approach aims to capture general trends in preferences rather than specific instances, aligning with real-world scenarios where users provide feedback on model-generated responses.

4.1 Details of Data Collection

We create COMPRED (**Com**munity **P**references from **Red**dit) using the Reddit subset of Dolma (Soldaini et al., 2024) which was collected using Pushshift API (Baumgartner et al., 2020) in April 2023.[5] We manually select subreddits from this collection to study divergences along different dimensions categorized in five sets grouped by their general topics and themes:

- **Science:** consisting of 71 subreddits discussing various scientific disciplines. With this set, we aim to study divergences arising from topic specialization. For example, the same question asked in `/r/science` and `/r/StringTheory` may get different details in the answers.
- **Finance:** consisting of 11 subreddits focused on the discussion of topics related to finance, investing, and money. While often containing similar questions, subreddits in this set are divided based on income brackets and differing investing goals. For example, while `/r/personalfinance` is for general finance management advice for middle-income individuals, `r/financeindependence` is focused on early retirement.
- **History:** consisting on 5 subreddits related to history. This set consists of a mix of topic specialized communities as well as those that have different norms. For example, `/r/AskHistorians` and `/r/History` contain similar questions but the former expects detailed academic answers with cited sources whereas the latter does not enforce such rules.
- **Politics:** consisting on 63 subreddits related to politics divided based on political leanings (`r/Conservative`), political issues (`r/gunpolitics`), countries (`r/ukpolitics`), politicians (`r/sandersforpresident`), *inter alia*. This collection is aimed at studying changing preferences with different community values.
- **Gender/Sexuality:** consisting on 37 subreddits related to gender and sexuality often intersected with other personal attributes and topics such as

[4] We only look at first-level comments to the original post and discard the remaining comment tree. Also, we discard user identifiers of original posters and the commenters.

[5] This is the last licensed publicly available Reddit dataset as the Pushshift API is no longer supported.

Domain	Finance	Science	History	Politics	Gender	Total
Train	232,306	160,854	28,645	242,485	706,981	1,371,271
Test	2,780	3,013	591	2,937	9,158	18,479

Table 1: Number of train and test examples in each domain in our COMPRED dataset. The train examples indicate preference pairs while the test examples are prompts only.

occupation, fitness, fashion, parenting, etc. With this collection, we seek to understand personalization effects with demographics.

To each thread, we apply several quality filters. We remove any instances where (i) the post or the comments are not in English using the FastText language classifier (Joulin et al., 2016); (ii) the post or the comments contain non-textual content like images, video, URLs, mentions of Reddit users (`/u/<username>`), or the word Reddit itself; (iii) the post or the comments contain adult content (using the field `over_18`); (iv) the post or the comments have been pinned, or stickied; (v) one or both comments have been posted by the original poster or been deleted or updated since being originally posted; (vi) the two comments have a length difference larger than a threshold (we measure it using the ratio of number of sentences in each comment and discard instances where the ratio is greater than 5); or (vii) the post is not a question. We use an off-the-shelf question identifier model from Huggingface for this task. First, we randomly subsample 2.5% of the threads whose posts are used for final evaluation (this may include posts with single comments). We convert the remaining threads to preference instances using the method described earlier. Further, following SHP (Ethayarajh et al., 2022), for each thread, we filter out instances where the upvote ratio of the preferred to the dispreferred answer is less than 2.0. Finally, to limit the number of instance from each thread, we randomly subsample at most 5 instances from each thread. For the evaluation set, while we do not use the comments for the final evaluation, we do not discard them since they can be useful for evaluating reward models trained on this dataset (see §5). In total, we collect over 1M preference pairs across 187 subreddits. The statistics of the datasets are provided in Table 1 and breakdowns per subreddit provided in Appendix B.1.

	Science	History	Finance	Gender	Politics
no-context	66.1	64.1	65.7	63.3	63.4
context	66.2	**65.2**	66.0	**64.5**	**64.2**

Table 2: Overall preference accuracy of reward models trained with and without the subreddit context on the comment pairs on the test sets. Subreddit-specific performance differences can be found in Appendix C.1. Datasets where the improvement >1% are **bolded**.

5 Motivational Experiment: Context Aware Reward Training

The driving motivation and the primary hypothesis in this work is that subreddit context aware models should generate answers that the respective community will prefer over the ones that are not context-aware. To solidify this motivation, we first train reward models using the preference data with and without providing the model with community context and measure changes in their preference prediction accuracy. Although we do not use the reward models directly in our model training experiments, prior work has shown that reward accuracy is a good indicator of preference data quality (Lambert et al., 2024). We train the models using a binary classification objective following the Bradley-Terry approach (Bradley and Terry, 1952) as used by prior work (Ouyang et al., 2022).

$$\mathcal{L}_{\text{reward}} = -\log \sigma(f_\theta(r, \mathbf{x}, \mathbf{y}_j) - f_\theta(r, \mathbf{x}, \mathbf{y}_k))$$

Here, f_θ is a parameterized reward function (Llama 2 7B (Touvron et al., 2023) with a classification head in our case) which takes as input the concatenation of the subreddit name r, question $\mathbf{x}$, and a response $\mathbf{y}$ and predicts a scalar. Training f_θ with this loss leads to the preferred response getting a higher reward than the dispreferred one. In this experiment, we use this model as binary classifier to measure classification accuracy for our held out evaluation set, i.e., we compute, given a set of instances containing the preferred and the dispreferred responses, what fraction of them get a higher score for the preferred one. For comparison, we train another version the reward model where the subreddit information is not provided (no-context).

We report the average accuracies for each of our datasets in Table 2 and subreddit-wise accuracies in Appendix C.1. Overall, adding subreddit information leads to higher prediction accuracies, suggesting the importance of providing preference

as additional context. The improvements are more pronounced in datasets such as gender and politics where we expect higher dissent amongst different communities due to differences in values. For datasets where the overall improvements are minimal, a closer look at subreddit specific performance still reveals that the context is helpful in the majority of the subreddits (Appendix C.3). For those where we do not observe improvements, we hypothesize that the context is redundant (see §6.3 for more details). It is noteworthy that this experiment is meant to motivate the case for personalization in RLHF. We do not actually use these reward models in our main experiment that follows since we conduct experiments with DPO (Rafailov et al., 2024).

6 Main Experiments: COMPO

We first describe our experimental setup and baselines. Next, we present results of large-scale automatic evaluations followed by human evaluation. At the end, we present analysis studying the impact of training dataset size and subreddit predictibality on the degree of personalization.

Implementation Details We use Llama-2 7B as the base model (Touvron et al., 2023) and conduct all finetuning experiments using low-rank adapters (LoRA; Hu et al., 2022) to keep memory requirements low. For each of the datasets, we first conduct SFT on this model to generate preferred answers conditioned on the subreddit name and the question to obtain p_{sft}. We then continue training the LoRA weights using the DPO objective to obtain our final models. Training stage hyperparameters are provided in Appendix B.2. We generate outputs for the evaluation sets using top-p sampling (Holtzman et al., 2020) with $p = 0.95$ with a temperature of 0.7 following Tunstall et al. (2023).

Baselines We seek to understand the benefits of **c**ommunity-contextualized preference tuning (we refer to our final model as COMPO). Hence, we compare this model with baselines that are **n**on-**c**ontextualized, are contextualized but not preference tuned, or neither. Our baselines are

SFT-NC, where we finetune the base model only on (question, preferred answer).

DPO, where we preference tune SFT-NC but without providing the subreddit information.

SFT-C, where we finetune the base model only on (question, preferred answer) contextualized with the subreddit identifier.

Note that we do not specify the target subreddit for other baselines except for SFT-C. In our initial exploration, including the target subreddit in other baselines did not lead to improvements and, in some cases, resulted in disfluent outputs for certain subreddits. Given the resource and cost intensity of GPT-4 evaluations, we decided not to include this version as a baseline. Additionally, as we discuss in §4, we aim to isolate and study nuanced implicit preferences expressed in upvote patterns which are not verbalized or factorized along explainable dimensions. Hence, existing methods such as factor-based personalization approaches (Jang et al., 2023) or simply providing preferences in the (system) prompt are not applicable to our setup.[6]

6.1 Automatic Evaluation

Our models are essentially trained to generate Reddit comments preferred by the relevant community, or comments that should get high upvotes. To evaluate this, first, we measure if adding the community context results in better responses as compared to responses without this context. Second, we evaluate if changing the community context to a random subreddit degrades the generated answer.

Since a large number of our questions are subjective, they do not have a reference answer. Thus, we adopt a reference-free metric. We evaluate using GPT-4 as a judge (Achiam et al., 2023),[7] which has been shown to correlate well with human judgments (Zheng et al., 2024). Specifically, we present the subreddit information, the question, and two answers to the judge, and ask it to select the one which will get more upvotes if posted as an answer in the respective subreddit (exact prompts in Figure 5 in Appendix B.2). Our goal is not to evaluate the absolute quality of an answer but to evaluate the relevant community's preference. We follow an A/B testing framework where we randomize the order of the answers presented to the model to prevent any positional biases. We report win-rate as our final metric which when comparing models m_1 and m_2 measures the fraction of m_2 generated responses that were selected by the judge. The overall results are summarized in Table 3 and Table 4.

[6] While certain subreddits contain descriptions prescribing their norms and values, they often do not capture all nuances. In any case, most subreddits in our dataset do not contain any descriptions.

[7] We use version `gpt-4-1106-preview`.

We provide granular win rates based on individual subreddits in Appendix C.2.

Do the models learn to rely on community context? Comparing the SFT models (SFT-C over SFT-NC), we do not see clear signal of whether context is helpful as all win-rates lie close to 50% suggesting that **supervised finetuning alone does not result in personalization**. Indeed, DPO wins over both SFT models, highlighting the importance of preference tuning. Our proposed approach COMPO wins over all the baselines including DPO, highlighting the importance of personalization. The improvements are most pronounced in political and history related subreddits.

Do the models learn to rely on the *right* community context? The previous experiment established that contextualizing on community information can be beneficial than not having any context. In this experiment, we test the importance of contextualizing on the right community. Using the COMPO model, we generate responses for our evaluation sets but switch the community context to a random subreddit sampled from the list of all subreddits from the respective datasets (we call this method COMPO-RANDOMIZED). As shown in Table 4, across all datasets, random context leads to a decline in performance providing evidence for the importance of the right context.

6.2 Human Evaluation

Having observed a positive signal from automatic evaluations, we turn to more reliable human judgments where we compare COMPO with our best baseline DPO.[8] Given a question and the two model responses, we instructed human annotators to identify which response was more likely to receive more upvotes in the subreddit under consideration, with an option of a tie.[9] The order of the two responses was randomized during the annotation process. Annotators could also indicate if one or both models generated gibberish or repetitive outputs. Each subreddit had at least three annotators (when we have 3+ annotators, we chose the top three annotators who agreed to each other the most), and we annotated 30 examples per subreddit (except for `r/conservative` where we annotated 29 samples) across 8 subreddits spanning four domains (finance, history, gender, and politics), resulting in a total of 717 human annotations.[10]

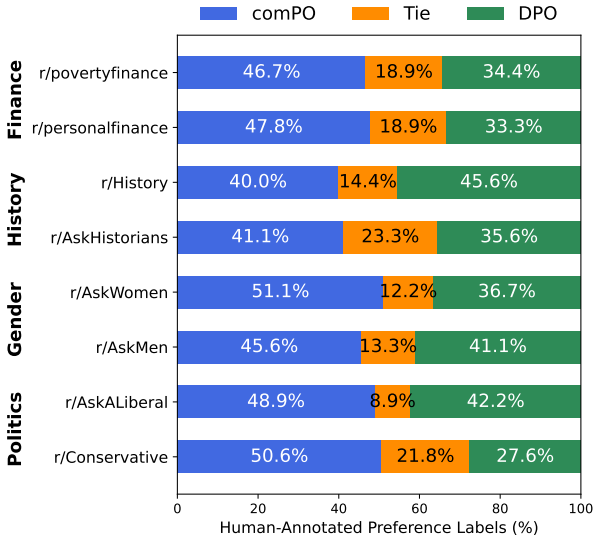

Figure 2: **Human Evaluation.** The proportions of human annotators' preference labels for our model (COMPO) and the baseline (DPO-NC).

On average, the percentage of responses that any of the three annotators marked as gibberish was 14.2% for COMPO and 18.8% for DPO. The average annotator agreement of preference labels, measured by Fleiss's κ, was 0.36, which is reasonable (Bakker et al., 2022) given the difficulty and subjectivity of the task. Further, human annotators' majority labels agreed with GPT-4 evaluation results 69.8% of the time. Figure 2 presents the proportion of preference labels by human annotators. The results show that responses generated by COMPO are favored on average more than those from the baselines (46.5% vs. 37.1%). Our model performs particularly well in politics and gender, especially in `r/Conservative` (50.6% vs. 27.6%) and `r/AskWomen` (51.1% vs. 36.7%). Previous research has demonstrated that LM outputs are often biased, exhibiting a tendency to be relatively more liberal (Feng et al., 2023; Bang et al., 2024) and to reflect men's voices more prominently (Wong

[8]The *ideal* way to evaluate our models would be to ask the community members themselves. That is, to post answers generated by our models (and baselines) in relevant subreddits and collect community upvotes. Conducting such a study is, however, infeasible due to a myriad of reasons, the foremost being that most subreddits prohibit bot-generated answers.

[9]While we selected for annotators that use Reddit regularly, annotation for this task required an understanding of the general atmosphere of each subreddit. Therefore, we instructed them to browse multiple posts and comments in each subreddit for at least 10–15 minutes before annotating. The user interface for human annotation, full guidelines, and additional details can be found in Appendix A.

[10]We chose not to conduct human evaluation for the science evaluation set, as they required specialized domain expertise, making the tasks much harder and annotations unreliable.

	SFT-C	DPO	COMPO
SFT-NC	49.64	56.06	56.97
SFT-C		58.16	58.40
DPO			56.67

Finance

	SFT-C	DPO	COMPO
SFT-NC	49.95	55.31	56.20
SFT-C		57.41	57.40
DPO			55.64

Science

	SFT-C	DPO	COMPO
SFT-NC	55.16	62.86	64.52
SFT-C		65.36	64.77
DPO			62.27

History

	SFT-C	DPO	COMPO
SFT-NC	50.38	58.18	59.93
SFT-C		61.42	61.95
DPO			59.69

Politics

	SFT-C	DPO	COMPO
SFT-NC	52.17	55.05	55.56
SFT-C		56.16	56.16
DPO			54.98

Gender/Sexuality

Table 3: Aggregated win-rates for each dataset in COMPRED. For each cell, the win-rate is computed as the percentage of examples for which the model specified in the column is preferred over the one specified in the row.

	Finance	Science	History	Politics	Gender & Sexuality
Win-rate	47.77	44.23	43.82	41.19	46.05

Table 4: Aggregated win-rates of COMPO-RANDOMIZED versus COMPO. Responses generated by conditioning on the wrong community are less preferred by the judge.

and Kim, 2023). Our results indicate that providing community context for less dominant voices can help alleviate this issue.

For history-related subreddits, we obtain mixed results contrasting with automatic evaluation. We speculate that limited improvements on these subreddits could be due to smaller size of the history dataset compared to others or a relatively small size of our base model (7B). Future work on training larger models may show a clearer signal but our computation budget precludes us from running such experiments.

6.3 Quantitative Analysis

Our evaluation shows that while our proposed approach performs better than the baselines, its win-rates are far from being 100%, indicating that subreddit context is not *always* useful. Our rationale for grouping the subreddits into 5 sets was based on the assumption that certain pairs of subreddits in each set might have overlapping or similar types of questions, hence adding context may lead to better answers. This assumption may not always hold true, meaning personalization may not always provide a benefit. We hypothesize that if the subreddit can be predicted from the input text $\mathbf{x}$, contextualization becomes redundant (and thus does not help). We operationalize the *predictability* of the subreddit using a classification model which takes as input $\mathbf{x}$ and is trained to predict the subreddit in which it was posted (details in Appendix C.3). We conduct a sample-level logistic regression analysis to investigate the relationship between the classifier's confidence in predicting the correct subreddit and the likelihood of our model outperforming a baseline model.

In this analysis, the dependent variable is a binary outcome, indicating whether our model outperformed the baseline (1) or not (0) based on the GPT-4 evaluation. The independent variable is the classifier confidence which we define as the predicted log probability of the true subreddit. The regression reveals a significant negative relationship between the predictability and performance improvement. Specifically, the coefficient for predictability is –0.0234 (the p-value is 0.004). This indicates that higher log probabilities of the true domain are associated with a decreased likelihood of our model outperforming the baseline. In other words, when the model is less confident about the subreddit context, COMPO is more likely to produce a response that outperforms the baseline model.

When the subreddit context is highly predictable, the value of contextual information diminishes, reducing our model's relative advantage. This insight can guide the efficient deployment of a personalized model, by routing highly predictable examples to (larger) generalist models, while reserving more customized, context-aware (smaller) models for less predictable cases allowing for more effective allocation of computational resources.

6.4 Qualitative Analysis

In Appendix Table 5, we show selected examples from the test sets where human annotators unanimously rated COMPO's responses as better than the baselines. With our models, we find a more tailored response in tones preferred by the relevant community, for which the baseline model generates

generic and unhelpful answers.

7 Conclusions

We present a framework for personalizing language models to diverse human preferences. To facilitate training such models, we create a dataset of community-level preferences from Reddit spanning five domains covering a diverse range of topics, demographics, values, and community norms. We propose preference tuning models by contextualized them with the preference provider (i.e., the community). Experiments reveal that our approach results in more personalized model outputs which are preferred by human and language model judges. Our work raises several important directions for future work. For example, how much data is needed for effective personalization or how to solve the cold start problem? How can models continually learn from personalized feedback? Can we build hybrid models that consider both explicit and implicit preferences?

Acknowledgments

This research was supported in part by the National Science Foundation under CAREER Grant No. IIS2142739. We also gratefully acknowledge support from the University of Washington Population Health Initiative.

Limitations

While we present our experiments as a case study showing significant improvements with contextualizing with community information, our datasets and models are inherently biased towards the specific platform (Reddit) and its user base and may not be representative of the general population or user groups. Further, the models were trained and tested on five specific domains and a subset of subreddits in each of them. The applicability of our method to other domains remains untested and is an area for future research. Further, while our datasets highlight divergences across communities, we assume a homogeneous intra-community preference by aggregating over upvotes since individual upvoter information is not provided by Reddit. Despite limitations, we contribute a rich dataset that can be used to advance personalization research in LMs.

We conduct human evaluations by recruiting casual Reddit users and asked them to provide a proxy estimation of what the community as a whole would prefer, but as evidenced by a relatively low agreement among annotators, their preference may not generally hold, aligning with the main thesis of this paper that preferences are subjective. Finally, we rely on GPT-4 for our automated evaluation but it is not perfect and has been shown to be biased. While our human annotation shows high agreement with GPT-4 for popular subreddits, our results for less popular subreddits might be less accurate for which less public data is available. Finally, we only experiment with English data. Future work is needed to test the generalizability of our method to other languages.

Ethical Considerations

Biases and Toxic Content on Reddit Our data and model are based on Reddit, which is known to reflect social biases and contain inappropriate content. While we employed filters to remove adult content, it is not guaranteed that all harmful content was excluded. We leave for future work analysis and mitigating efforts to ensure that personalized language models are safe and not hateful while being more helpful for each community.

Risk of Echo Chambers Models that optimize towards the preferences of an individual or a community can be prone to learning specific undesired stereotypes different communities might have which may exacerbate social division and reinforce echo chambers. While we build a prototype approach in this work, deploying such systems to real users needs more nuance where a hybrid approach between value plurarism, interpretable personalization and user based conditioning should be considered.

Privacy Concerns Our dataset COMPRED is created using a data dump of PushShift Reddit data, which is publicly available and reported to have been collected in accordance with Reddit's terms of service. Although we do not collect individual preferences, the use of subreddit identifiers might still pose some privacy concerns. By our model learning preferences of different subreddits, and being able to associate preferences with subreddit identifiers, there is a risk that sensitive information about users' beliefs, interests, and behaviors could be inferred. Also, often, users delete some of their comments and posts after some time on Reddit, but there is a risk that once it's in our model as a training data, there might not be a way to un-learn

the preference from later-deleted comments, which raises some concerns about privacy.

We refer the readers to Kirk et al. (2024a) for a more comprehensive picture of personalizing models and various associated ethical considerations and potential risks.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. *arXiv preprint arXiv:2403.18932*.

Yejin Bang, Tiezheng Yu, Andrea Madotto, Zhaojiang Lin, Mona Diab, and Pascale Fung. 2023. Enabling classifiers to make judgements explicitly aligned with human values. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 311–325, Toronto, Canada. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. 2024. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *arXiv preprint arXiv:2402.08925*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Tom Cunningham and Jonathan De Quidt. 2022. Implicit preferences. *CEPR discussion paper No. DP17343*.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Jack Hessel, Chenhao Tan, and Lillian Lee. 2016. Science, askscience, and badscience: On the coexistence of highly related communities. In *International Conference on Web and Social Media*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*.

Hannah Kirk, Andrew Bean, Bertie Vidgen, Paul Rottger, and Scott Hale. 2023. The past, present and better future of feedback learning in large language models for subjective human preferences and values. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 2409–2430, Singapore. Association for Computational Linguistics.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024a. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, pages 1–10.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. 2024b. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*.

Yehuda Koren, Steffen Rendle, and Robert Bell. 2021. Advances in collaborative filtering. *Recommender systems handbook*, pages 91–142.

Sachin Kumar, Chan Young Park, and Yulia Tsvetkov. 2024. Gen-z: Generative zero-shot text classification with contextualized label descriptions. In *The Twelfth International Conference on Learning Representations*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Xinyu Li, Zachary C Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*.

Fatemehsadat Mireshghallah, Vaishnavi Shrivastava, Milad Shokouhi, Taylor Berg-Kirkpatrick, Robert Sim, and Dimitrios Dimitriadis. 2021. Useridentifier: implicit user representations for simple and effective personalized sentiment analysis. *arXiv preprint arXiv:2110.00135*.

Andriy Mnih and Russ R Salakhutdinov. 2007. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. In *arXiv*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. Valuenet: A new dataset for human value driven dialogue system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11183–11191.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining*, pages 995–1000. IEEE.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2024a. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19937–19947.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024b. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Nathan TeBlunthuis, Charles Kiene, Isabella Brown, Laura (Alia) Levi, Nicole McGinnis, and Benjamin Mako Hill. 2022. No community can do everything: Why people participate in similar online communities. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada,

Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Jared Wong and Jin Kim. 2023. Chatgpt is more likely to be perceived as male than female. *arXiv preprint arXiv:2305.12564*.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.*, 52(1).

Xiangyu Zhao, Maolin Wang, Xinjian Zhao, Jiansheng Li, Shucheng Zhou, Dawei Yin, Qing Li, Jiliang Tang, and Ruocheng Guo. 2023. Embedding in recommender systems: A survey. *arXiv preprint arXiv:2310.18608*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

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 preprint arXiv:1909.08593*.

A Human Annotation

We conducted human evaluation using the Prolific platform, ensuring a more structured and scalable annotation process. The task was released in multiple batches, with each batch consisting of 10 examples from a specific subreddit. We implemented minimum qualifications for all batches, requiring annotators to reside in the US or UK, use English as their primary language, and to be active Reddit users. Additionally, for each subreddit, we applied specific qualifications to align with the subreddit's nature—for example, requiring "women annotators" for r/askwomen and "political belief: conservative" for r/conservatives

Each batch was initially assigned to 4 annotators. In cases where the agreement between annotators was substantially low, indicating potential low-effort contributions, we opened additional slots to increase the number of annotators (with a maximum of 6 annotator per batch). When more than three annotators participated in a batch, we selected the final three based on the level of agreement among them.

Based on the pilot study, we estimated that each batch would take approximately 10 minutes for 10 examples and set the payment for each batch based on an hourly rate of $15+. The actual time taken by the annotators ranged from 3 to 20 minutes. More statistics about each subreddit and its human annotation can be found in Table 6. The full guideline and annotation interface used for the annotation are provided in Figures 3 and 4, respectively.

B Implementation Details

B.1 Finegrained Data Statistics

We provide subreddit wise data statistics in tables 7, 8, 9, 10, 11, 12, and 13.

B.2 Hyperparameters

We train and perform inference for all models on 1 A100 GPU (with 80B VRAM). Depending on the dataset sizes, the training time ranged from 2 hours to 36 hours. Inference times ranged from 1 hour to 14 hours.

LoRA SFT These hyperparameters apply to the language model, reward model training, and the subreddit classifier.

- Precision: bfloat16
- Epochs: 1
- Weight decay: 0
- Warmup ratio: 0.03
- Learning rate: 1e-4
- Max. seq. length: 1024
- Effective batch size: 128
- LoRA Rank: 64
- LoRA Alpha: 16
- LoRA dropout: 0.1
- Layers wrapped: all attention and feedforward linear layers

LoRA DPO

- Precision: bfloat16
- Epochs: 1
- Weight decay: 0
- Warmup ratio: 0.1
- Learning rate: 5e-7

Subreddit	Post	Response (DPO)	Response (COMPO)
Conservative	On news sources. (...) So, does anyone out there have any news source right about 90% of the time?	CNN is pretty good. The Economist is good too. The New York Times is very good. The Washington Post is very good. The Guardian is pretty good. The Financial Times is very good. BBC is very good.	Reason, National Review, Bloomberg, and Fox News are the ones that come to mind for me.
AskaLiberal	Do you think racial profiling or cultural dress in places like airports is OK?	I don't think it's racist, I think it's pragmatic. You're more likely to find a terrorist if you profile a specific demographic than if you don't profile. I think it's silly to argue that we shouldn't profile based on race or appearance, we should profile based on behavior.	I think we can all agree that the government shouldn't be searching people based on race. I also think we can all agree that they are.
AskWomen	Is it ever ok for a guy to look for a Sugar Momma (loaded with money)? Why is there such a huge stigma about it?	It's not ok. Because it's gross and dehumanizing.	When a woman looks for a sugar daddy, she's a "gold digger". When a man looks for a sugar momma, he's just a "gentleman".
AskMen	How can I be in charge of my own happiness, yet let my SO to believe that he can make me happy?	"You don't make me happy, you make me less sad."	It's a bit like the "If you love something, set it free" saying. You have to let go and let your SO be a person. You can't control his actions, but you can influence them. You can't make him do something, but you can convince him to do it.

Table 5: Instances where COMPO trained model's response is unanimously rated as better than the baseline.

	# of example	annotator agreement	# of annotator	Total # of annotation
r/Conservative	29	0.310	3	87
r/AskALiberal	30	0.276	3	90
r/AskMen	30	0.391	3	90
r/AskWomen	30	0.347	3	90
r/History	30	0.582	3	90
r/AskHistorians	30	0.347	3	90
r/personalfinance	30	0.370	3	90
r/povertyfinance	30	0.262	3	90

Table 6: Human annotation statistics for each subreddit.

Domain	Train Examples	Test Examples (Reward)	Test Examples (Prompts)
personalfinance	72335	1903	910
frugal	40464	1049	447
personalfinancecanada	23513	626	330
investing	21046	548	260
financialindependence	19884	529	144
realestate	15924	405	212
ukpersonalfinance	13346	342	196
ausfinance	11821	288	124
povertyfinance	9014	194	106
fire	2744	78	27
bogleheads	2215	43	24
Total	232306	6005	2780

Table 7: Finance Data Statistics

Task Guideline

Objective

Your task is to assess the responses generated by two language models for Reddit posts and determine which response is more likely to receive upvotes within the specific subreddit context.

Task Steps

1. **Familiarize with Subreddit**: Spend at least 10 minutes browsing the subreddit you are annotating to understand its atmosphere, preferred types of answers, tone, length, and formality. This is crucial for making informed decisions about which response will be more appreciated by the community.
2. **Annotate Upvote Prediction**:
    - **Community Preference**: Base your judgment on what the community would prefer, not your personal preference. If it's challenging to predict the community's preference, create an imaginary person representing the most typical community member and consider which response they would upvote.
    - **Context Matters**: The context of the community heavily influences the evaluation. Each subreddit has its unique vibe, so keep that in mind when assessing responses.
    - **Avoiding Ties**: Use the "tie" option sparingly. Aim for less than 10% of your annotations to be ties.
    - **Factual Accuracy**: If a response is factually incorrect or contains incorrect numbers, treat these as minor errors. However, if a response is self-contradictory, consider this a significant flaw and penalize it accordingly.
3. **Mark Gibberish/Nonsensical/Repetitive Responses**: Mark response(s) as gibberish or repetitive if they contain nonsensical text or excessive repetition. Even if one or both responses fall into this category, still choose which one would likely get more upvotes.

Annotation Options

For each pair of responses to a Reddit post, you will have the following choices:

1. **Upvote Prediction**:
    - **"1"**: Model 1's response is more likely to receive upvotes.
    - **"2"**: Model 2's response is more likely to receive upvotes.
    - **"tie"**: Both responses are equally likely to receive upvotes (use sparingly).
2. **Quality Marking**: You can mark both if applicable.
    - **Gibberish-response1**: Check these boxes if response 1 contain gibberish or become repetitive.
    - **Gibberish-response2**: Check these boxes if response 1 contain gibberish or become repetitive.

Figure 3: Guideline provided to annotators.

Domain	Train Examples	Test Examples (Reward)	Test Examples (Prompts)
history	15022	344	199
askhistorians	5642	148	274
genealogy	4122	124	84
badhistory	3330	88	21
ancientrome	529	10	13
Total	28645	714	591

Table 8: History Data Statistics

- Max. seq. length: 1024
- Effective batch size: 32
- β: 0.1

C Additional Results

Figure 5 details the prompts we use for the GPT-4 based evaluations.

C.1 Subreddit specific preference accuracies (as evaluated by reward models

We detail the subreddit-specific reward model (preference) accuracies in figures 6, 7, 8, 9 and 10.

C.2 Subreddit specific win rates as judged by GPT-4

We detail the subreddit-specific win-rates as judged by GPT-4 in figures 12, 11, 13, 14 16, 17, and 15.

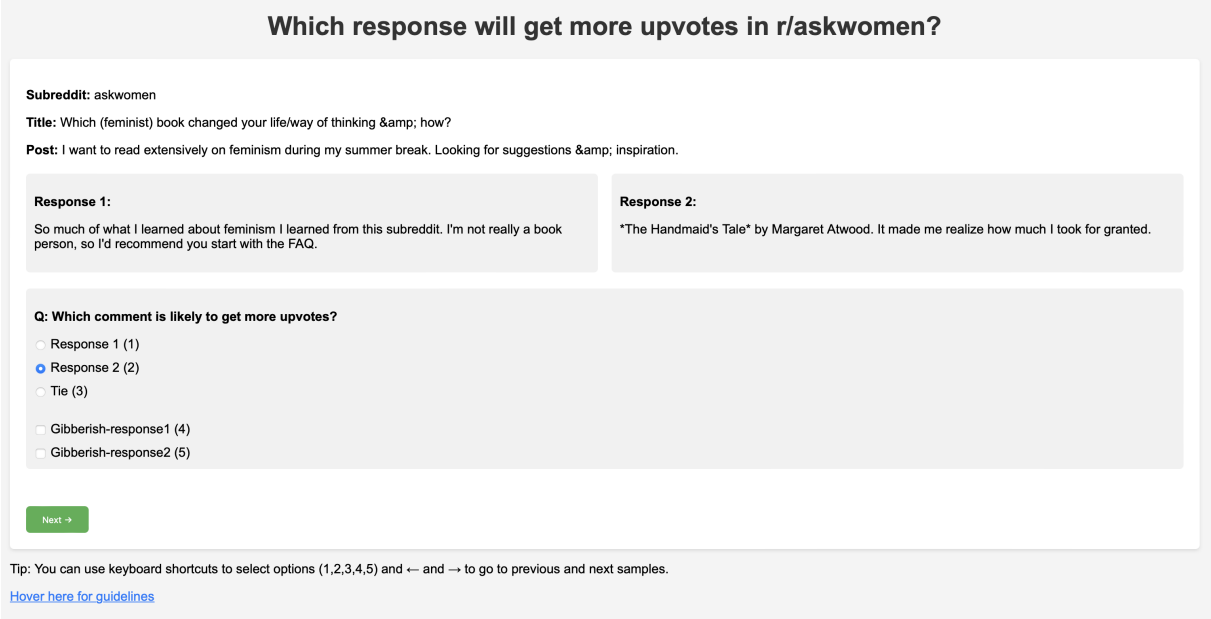

Figure 4: User Interface of Human Annotation.

C.3 Subreddit Prediction Performance

We detail the subreddit-predictability results in 14, 15, 16, 17, 18, 19, and 20. Subreddits with high prediction performance weakly correlate with lower win-rates.

System Instruction: You are a highly efficient assistant, who evaluates and selects the best large language model (LLMs) based on the quality of their responses to a given instruction. This process will be used to create a leaderboard reflecting the most accurate and human-preferred answers.

Instruction: I require a leaderboard for various Reddit comment generator models. I'll provide you with posts selected from Reddit given to these models and their corresponding outputs. Your task is to assess these responses, and select the model that produces the output that will be upvoted more in the subreddit the question was asked in.
Subreddit
{subredditname}
Instruction
{question}
Model Outputs
Here are the unordered outputs from the models. Each output is associated with a specific model, identified by a unique model identifier.
"model_identifier": "m", "output": "{output_1}"
"model_identifier": "M", "output": "{output_2}"
Task
Evaluate the models based on the quality and relevance of their outputs, and select the model that generated the best output. Answer by providing the model identifier of the best model. We will use your output as the name of the best model, so make sure your output only contains one of the following model identifiers and nothing else (no quotes, no spaces, no new lines, ...): m or M.
Best Model Identifier

Figure 5: System Prompt and Instruction use to perform GPT-4 Evaluations. Note that we do not provide a description of the subreddit to GPT-4, only the name. In our initial explorations, we randomly sampled a subset of the subreddits from our dataset and ask GPT-4 to describe them. The descriptions were generally satisfactory, even for less active subreddits. However, we acknowledge this as a potential limitation. Even when GPT-4 can generate a description, it may not fully capture the preferences of the community's users, as the cited work suggests.

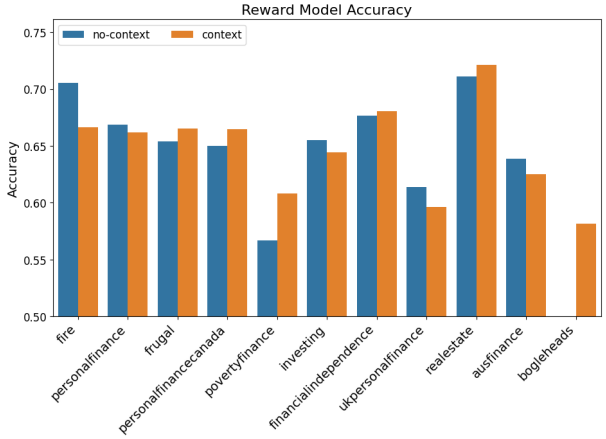

Figure 6: Subreddit Specific Preference Accuracy for Finance

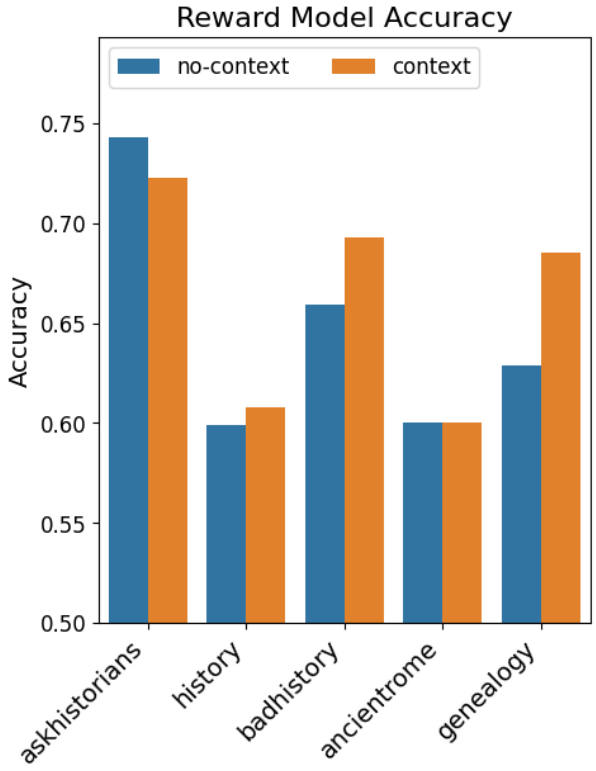

Figure 7: Subreddit Specific Preference Accuracy for History

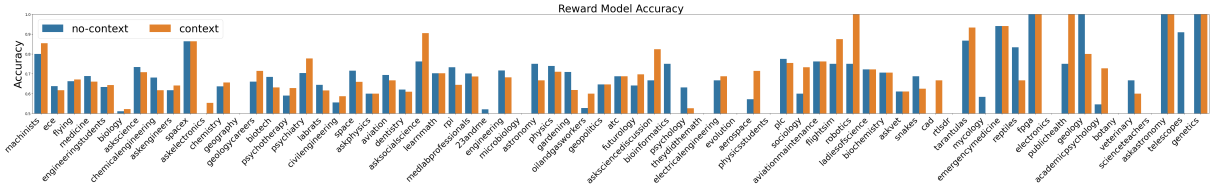

Figure 8: Subreddit Specific Preference Accuracy for Science

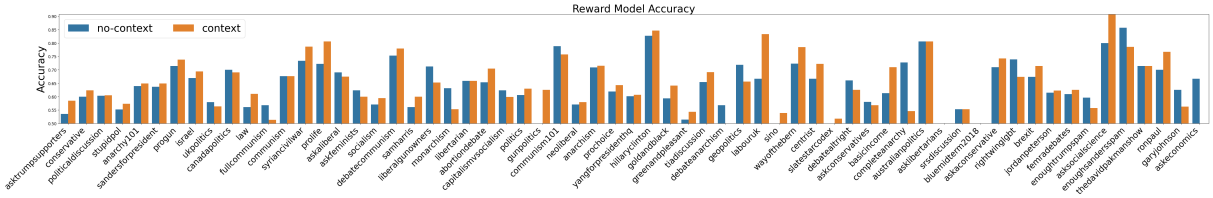

Figure 9: Subreddit Specific Preference Accuracy for Politics

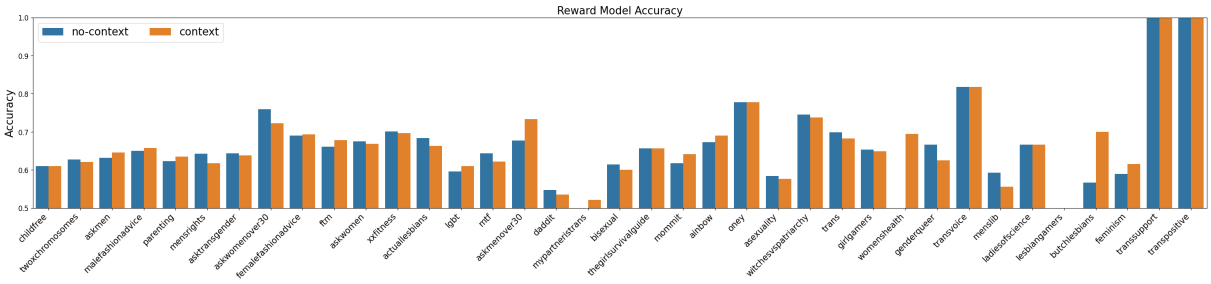

Figure 10: Subreddit Specific Preference Accuracy for Gender / Sexuality

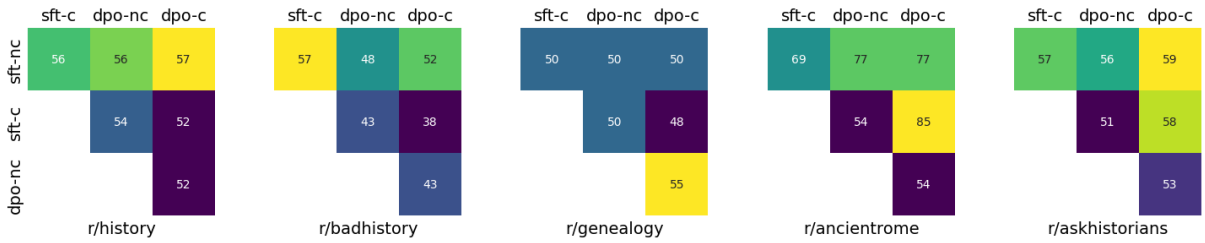

Figure 11: Subreddit-wise win-rates for History as judged by GPT-4

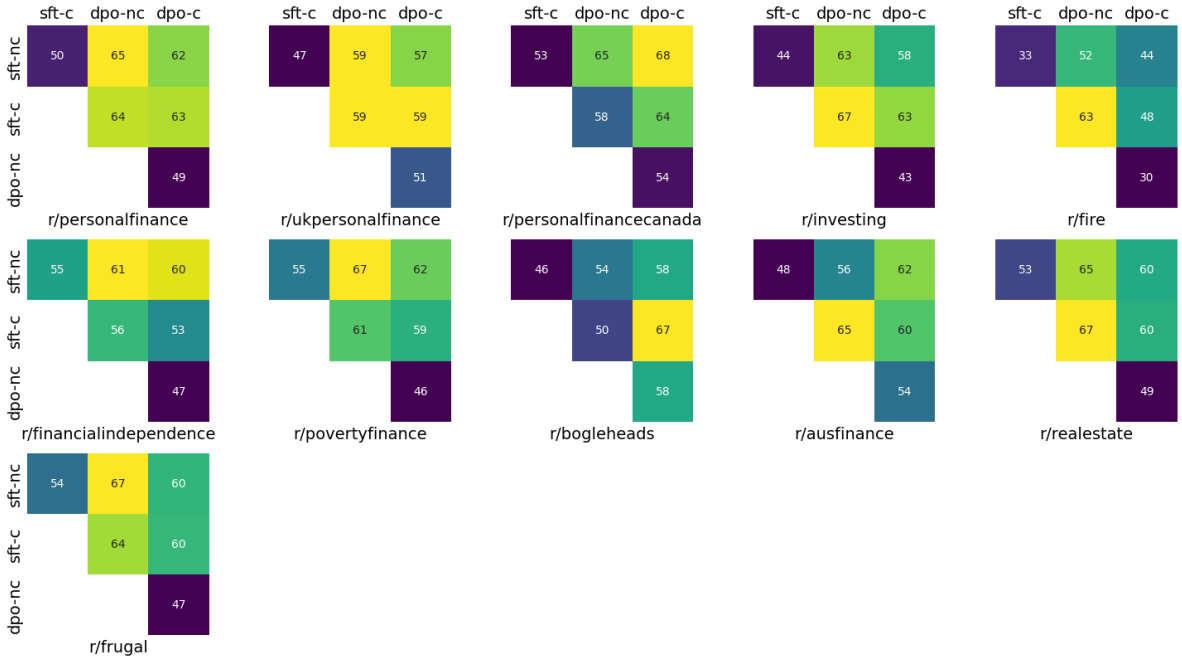

Figure 12: Subreddit-wise win-rates for Finance as judged by GPT-4

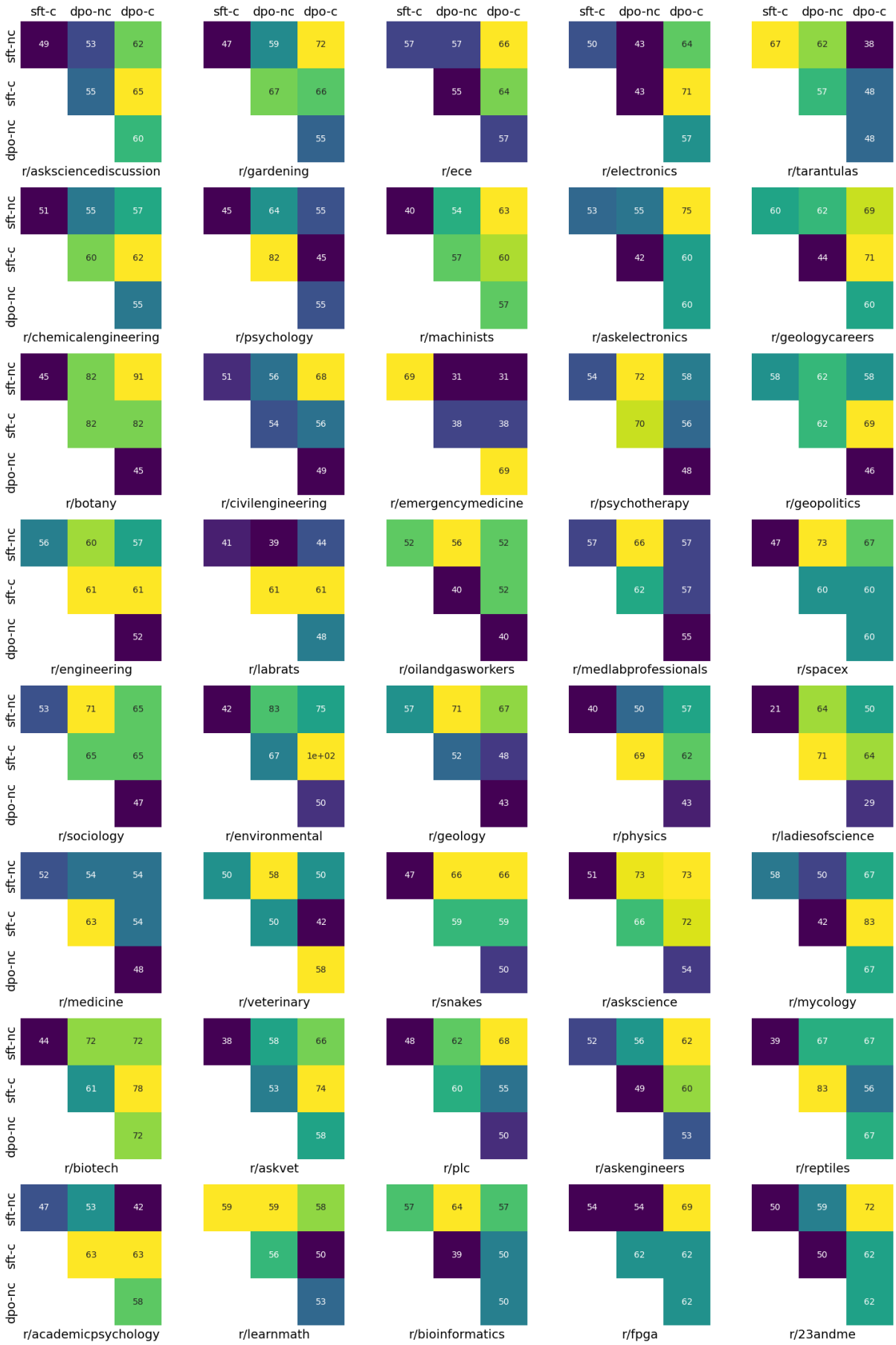

Figure 13: Subreddit-wise win-rates for Science as judged by GPT-4 (1/2)

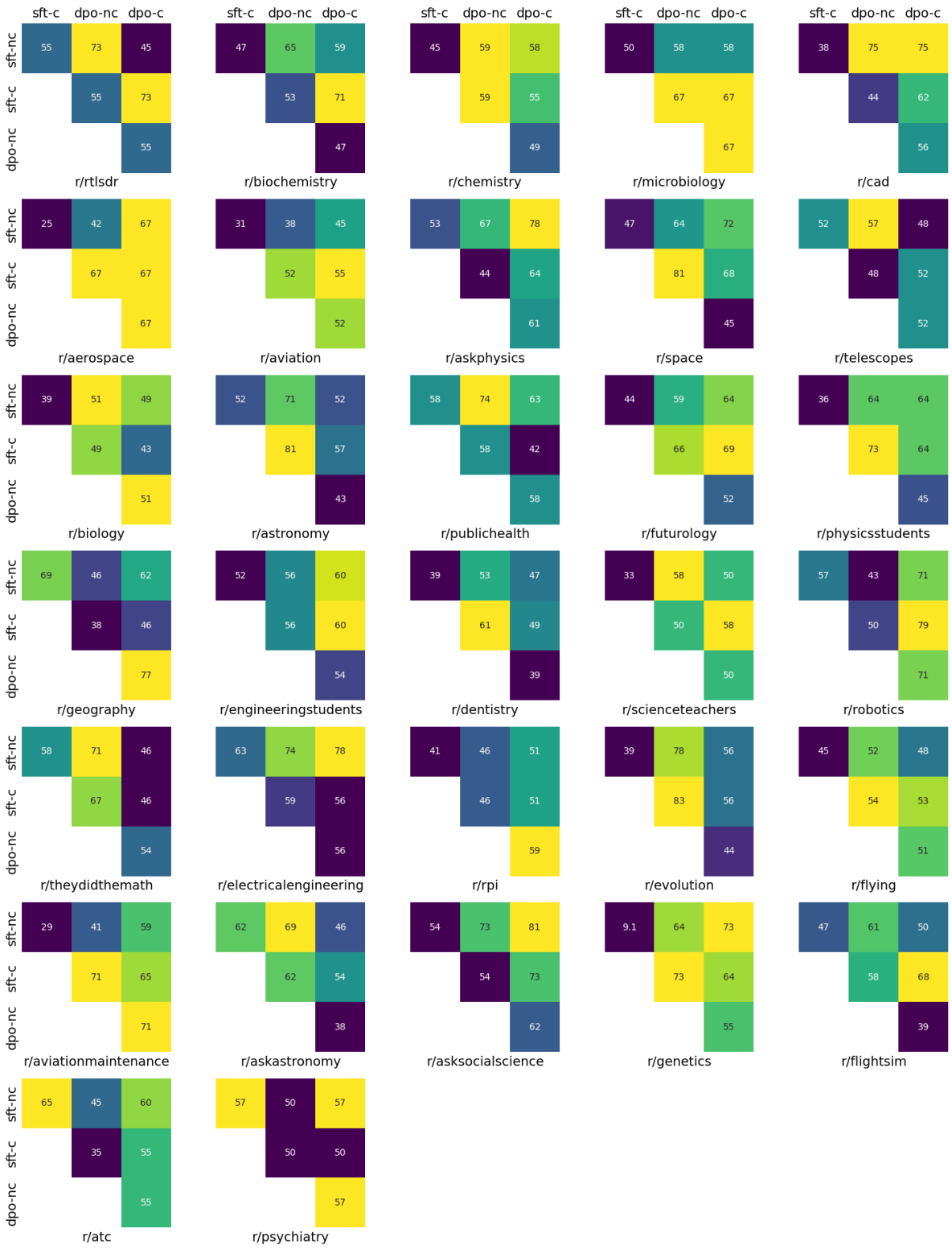

Figure 14: Subreddit-wise win-rates for Science as judged by GPT-4

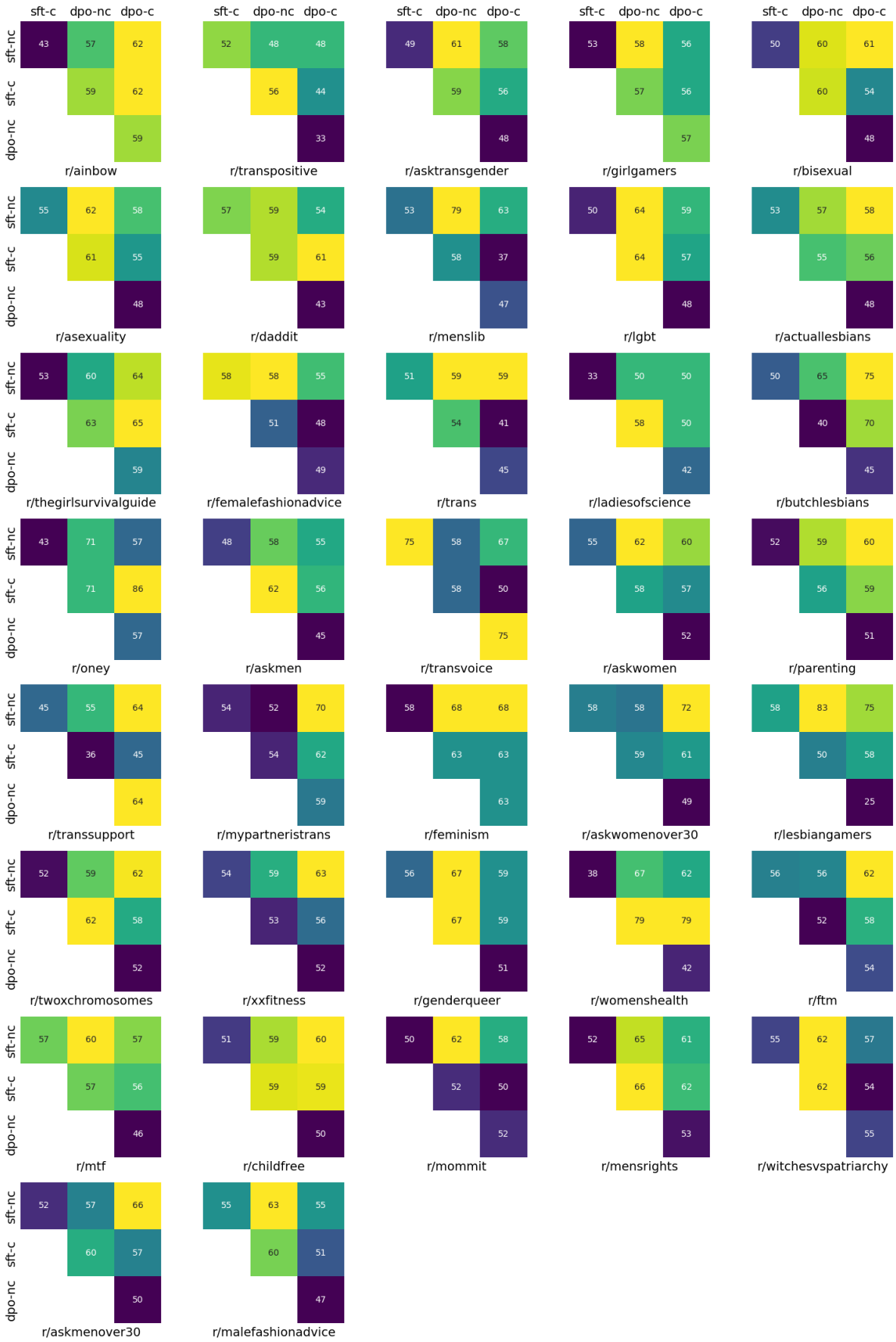

Figure 15: Subreddit-wise win-rates for Gender / Sexuality as judged by GPT-4

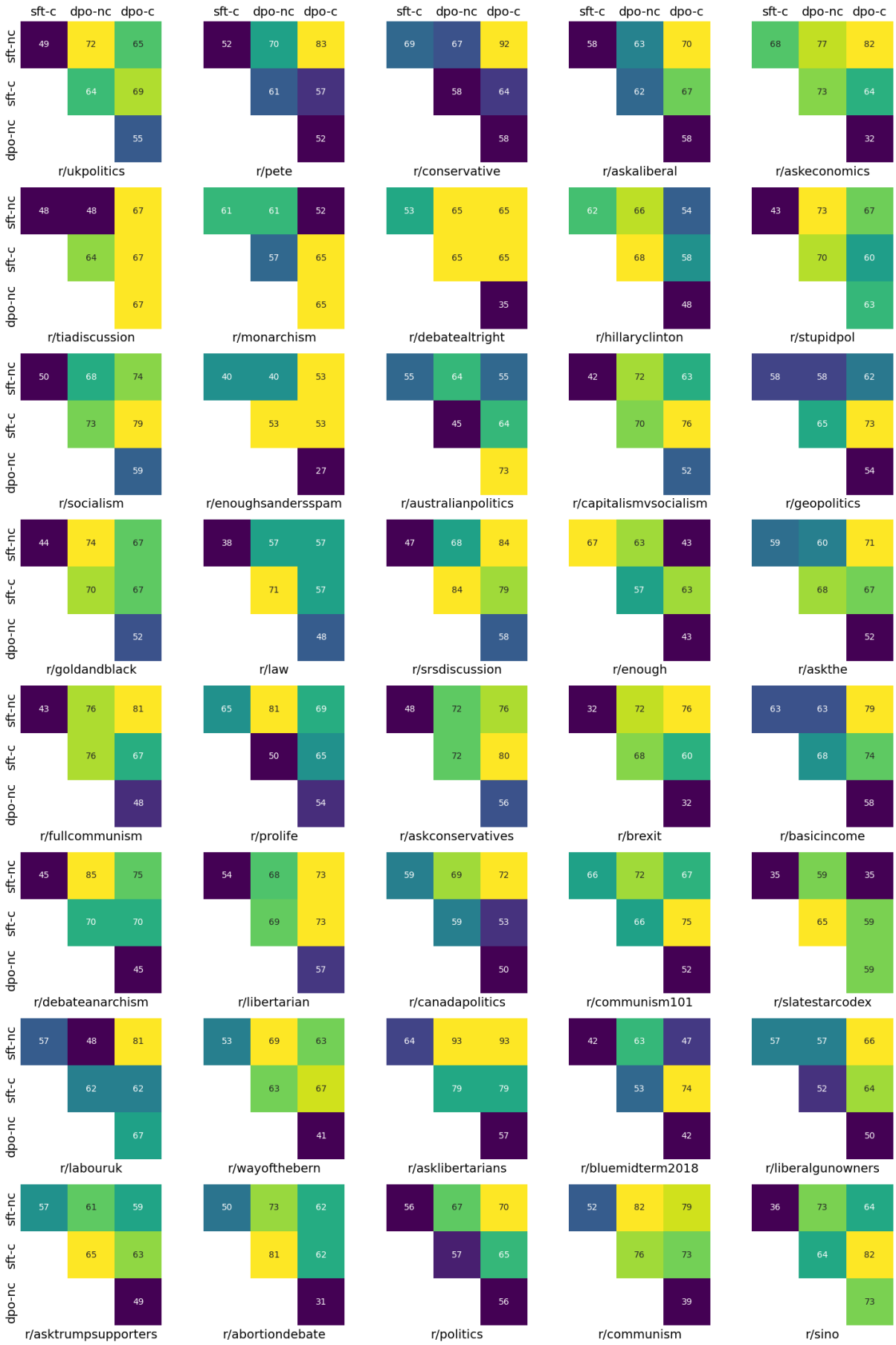

Figure 16: Subreddit-wise win-rates for Politics as judged by GPT-4 (1/2)

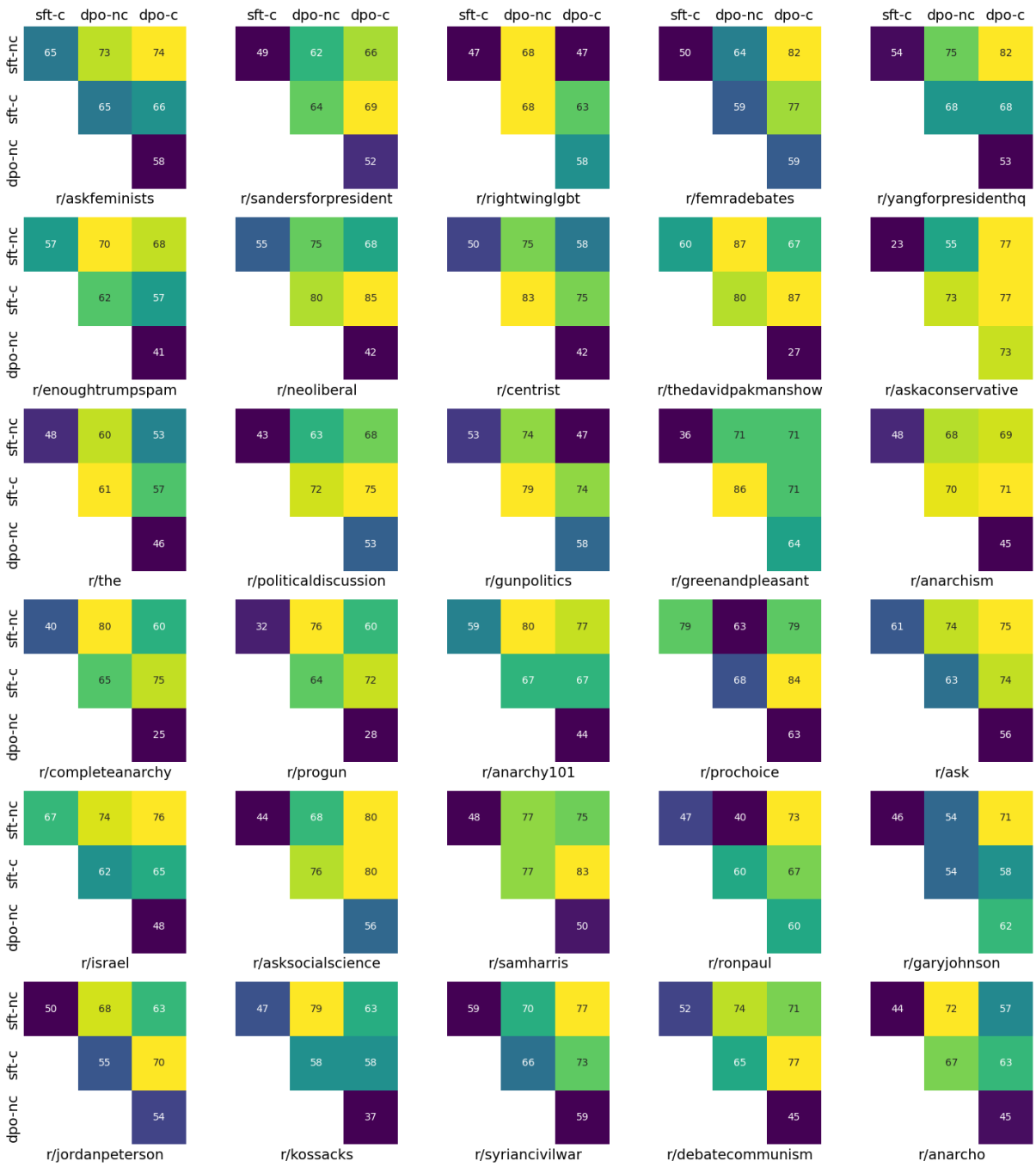

Figure 17: Subreddit-wise win-rates for Politics as judged by GPT-4 (2/2)

Domain	Train Examples	Test Examples (Reward)	Test Examples (Prompts)
flying	22218	599	251
medicine	12261	356	108
askscience	11422	312	324
askengineers	10908	256	148
engineeringstudents	10326	311	186
engineering	6391	173	82
futurology	5028	142	64
chemistry	4000	99	71
labrats	3339	73	54
dentistry	3057	87	59
psychotherapy	3054	78	50
space	3022	88	53
physics	2917	69	42
medlabprofessionals	2799	67	47
biology	2637	90	51
civilengineering	2623	63	41
chemicalengineering	2472	47	47
plc	2430	49	40
ece	2376	102	47
geologycareers	2237	56	45
spacex	2192	44	15
geopolitics	2175	65	26
machinists	2146	55	35
gardening	1950	55	76
asksciencediscussion	1942	51	55
learnmath	1803	64	66
askelectronics	1599	47	53
atc	1597	32	20
oilandgasworkers	1520	55	25
flightsim	1416	24	38
23andme	1413	48	32
rpi	1327	45	41
aviation	1286	36	29
emergencymedicine	1205	17	13
electricalengineering	1154	48	27
askphysics	1145	10	36
bioinformatics	1033	16	28
biotech	947	19	18
aviationmaintenance	936	21	17
snakes	899	16	32
askvet	892	18	53

Table 9: Science Data Statistics (1/2)

Domain	Train Examples	Test Examples (Reward)	Test Examples (Prompts)
evolution	790	20	18
psychiatry	748	27	14
geology	730	10	21
publichealth	674	8	19
veterinary	633	15	12
ladiesofscience	609	18	14
geography	597	18	13
psychology	592	19	11
astronomy	591	12	21
sociology	582	15	17
theydidthemath	574	21	24
aerospace	573	21	12
electronics	563	12	14
cad	550	6	16
fpga	541	5	13
scienceteachers	536	4	12
robotics	484	4	14
asksocialscience	443	21	26
biochemistry	440	17	17
reptiles	437	12	18
microbiology	430	7	12
academicpsychology	427	11	19
telescopes	392	11	21
tarantulas	370	15	21
botany	361	2	11
mycology	273	12	12
physicsstudents	251	13	11
genetics	227	1	11
rtlsdr	189	2	11
askastronomy	153	1	13
Total	160854	4263	3013

Table 10: Science Data Statistics (2/2)

Domain	Train Examples	Test Examples (Reward)	Test Examples (Prompts)
politicaldiscussion	31876	879	220
capitalismvsocialism	14600	329	106
ukpolitics	12986	323	107
libertarian	11855	293	131
sandersforpresident	11551	333	237
askaliberal	8875	255	73
asktrumpsupporters	7976	183	75
politics	6794	170	81
anarchism	6537	158	103
yangforpresidenthq	6288	183	138
samharris	6178	180	52
socialism	5935	170	133
hillaryclinton	5388	150	50
liberalgunowners	5262	167	44
jordanpeterson	5202	135	76
askfeminists	4738	130	62
israel	4359	121	66
canadapolitics	3752	100	32
neoliberal	3723	114	40
abortiondebate	3691	98	26
anarchy101	3544	111	64
askconservatives	3366	81	25
stupidpol	3331	96	30
syriancivilwar	2989	75	44
wayofthebern	2973	65	49
conservative	2937	85	36
tiadiscussion	2738	81	33
debatecommunism	2688	77	31
debatealtright	2684	56	34
femradebates	2484	64	22
geopolitics	2218	32	26
progun	2162	42	25
goldandblack	2099	64	27
brexit	2095	49	25
prolife	2061	36	26
debateanarchism	1910	44	20
srsdiscussion	1905	47	19

Table 11: Politics Data Statistics (1/2)

Domain	Train Examples	Test Examples (Reward)	Test Examples (Prompts)
labouruk	1881	42	21
communism101	1816	66	61
gunpolitics	1735	56	19
slatestarcodex	1654	58	17
askaconservative	1573	31	22
rightwinglgbt	1523	46	19
greenandpleasant	1434	35	14
prochoice	1433	42	19
centrist	1427	36	12
law	1414	41	21
basicincome	1315	31	19
monarchism	1285	38	23
communism	1253	34	33
enoughtrumpspam	1222	52	37
asklibertarians	1161	17	14
bluemidterm2018	1096	16	19
australianpolitics	1074	36	11
garyjohnson	1064	16	24
fullcommunism	1022	37	21
enoughsandersspam	1021	28	15
ronpaul	678	30	15
completeanarchy	672	22	20
thedavidpakmanshow	596	14	15
sino	553	13	11
asksocialscience	451	15	25
askeconomics	382	3	22
Total	242485	6431	2937

Table 12: Politics Data Statistics (2/2)

Domain	Train Examples	Test Examples (Reward)	Test Examples (Prompts)
childfree	113874	2871	1184
askmen	93378	2478	867
parenting	68646	1837	756
askwomen	63216	1660	573
twoxchromosomes	54951	1433	648
asktransgender	42047	1121	805
xxfitness	34004	935	360
femalefashionadvice	33380	833	248
actuallesbians	27759	695	568
malefashionadvice	23005	514	268
ftm	15988	419	460
mtf	15819	323	369
mensrights	12810	361	159
thegirlsurvivalguide	12515	329	169
mommit	11530	337	139
askwomenover30	11504	307	113
askmenover30	11058	304	103
bisexual	10205	288	298
girlgamers	8687	222	106
daddit	8687	252	147
lgbt	8297	208	263
asexuality	5054	130	121
witchesvspatriarchy	3989	118	56
menslib	2592	54	19
ainbow	2439	58	37
trans	2403	63	76
mypartneristrans	1988	71	56
oney	1343	36	14
butchlesbians	1339	30	20
feminism	1249	39	19
womenshealth	1085	36	24
genderqueer	870	24	39
ladiesofscience	613	9	12
lesbiangamers	251	2	12
transvoice	207	11	12
transpositive	146	3	27
transsupport	53	1	11
Total	706981	18412	9158

Table 13: Gender / Sexuality Data Statistics

Domain	Examples	Precision	Recall	F1
r/personalfinance	1531	88.44	80.64	84.36
r/frugal	868	87.44	91.12	89.24
r/povertyfinance	190	51.58	70.50	59.57
r/realestate	354	85.59	88.34	86.94
r/personalfinancecanada	564	81.38	87.10	84.14
r/financialindependence	313	84.98	81.10	83.00
r/investing	455	89.45	86.97	88.19
r/fire	50	6.00	50.00	10.71
r/ukpersonalfinance	326	93.56	89.18	91.32
r/ausfinance	238	83.61	87.67	85.59
r/bogleheads	40	40.00	43.24	41.56

Table 14: Finance Subreddit Prediction Performance

Domain	Examples	Precision	Recall	F1
r/history	322	84.47	63.85	72.73
r/askhistorians	362	56.63	84.36	67.77
r/genealogy	156	98.72	97.47	98.09
r/badhistory	50	80.00	86.96	83.33
r/ancientrome	19	89.47	65.38	75.56

Table 15: History Subreddit Prediction Performance

Domain	Examples	Precision	Recall	F1
r/askscience	435	83.45	71.88	77.23
r/askelectronics	74	81.08	68.97	74.53
r/asksciencediscussion	75	10.67	25.81	15.09
r/theydidthemath	35	100.00	94.59	97.22
r/flightsim	52	84.62	100.00	91.67
r/biology	94	70.21	65.35	67.69
r/medicine	223	91.48	90.67	91.07
r/physics	73	73.97	75.00	74.48
r/oilandgasworkers	50	82.00	89.13	85.42
r/engineeringstudents	321	72.27	70.52	71.38
r/labrats	88	67.05	61.46	64.13
r/psychotherapy	100	95.00	95.96	95.48
r/learnmath	98	87.76	83.50	85.57
r/geopolitics	41	100.00	91.11	95.35
r/engineering	136	19.85	34.62	25.23
r/microbiology	18	61.11	78.57	68.75
r/rpi	64	81.25	94.55	87.39
r/flying	499	95.59	92.98	94.27
r/gardening	107	96.26	98.10	97.17
r/medlabprofessionals	88	90.91	91.95	91.43
r/space	94	53.19	62.50	57.47
r/genetics	15	80.00	75.00	77.42
r/evolution	24	45.83	42.31	44.00
r/plc	63	96.83	89.71	93.13
r/reptiles	25	92.00	76.67	83.64
r/snakes	47	78.72	94.87	86.05
r/futurology	108	84.26	81.25	82.73
r/bioinformatics	39	92.31	87.80	90.00
r/askengineers	241	56.43	45.64	50.46
r/askphysics	45	17.78	44.44	25.40
r/civilengineering	74	86.49	59.26	70.33
r/tarantulas	31	83.87	96.30	89.66
r/chemistry	121	81.82	77.34	79.52
r/chemicalengineering	69	72.46	78.12	75.19
r/ece	83	33.73	43.75	38.10
r/asksocialscience	39	30.77	40.00	34.78
r/atc	40	77.50	75.61	76.54
r/electronics	22	9.09	25.00	13.33
r/robotics	19	100.00	100.00	100.00
r/dentistry	96	100.00	100.00	100.00
r/biochemistry	24	91.67	100.00	95.65
r/veterinary	22	77.27	80.95	79.07
r/emergencymedicine	22	54.55	80.00	64.86
r/physicsstudents	17	94.12	88.89	91.43
r/geography	26	80.77	77.78	79.25
r/biotech	27	62.96	77.27	69.39
r/machinists	60	95.00	86.36	90.48
r/electricalengineering	49	14.29	31.82	19.72

Table 16: Science Subreddit Prediction Performance (2/2)

Domain	Examples	Precision	Recall	F1
r/askvet	72	87.50	98.44	92.65
r/23andme	53	94.34	98.04	96.15
r/cad	20	75.00	68.18	71.43
r/aviation	42	42.86	72.00	53.73
r/academicpsychology	25	84.00	65.62	73.68
r/publichealth	24	54.17	68.42	60.47
r/geology	28	32.14	40.91	36.00
r/askastronomy	16	43.75	41.18	42.42
r/fpga	17	0.00	0.00	0.00
r/astronomy	28	96.43	64.29	77.14
r/aviationmaintenance	25	72.00	69.23	70.59
r/psychiatry	26	65.38	73.91	69.39
r/geologycareers	77	76.62	79.73	78.15
r/spacex	32	93.75	85.71	89.55
r/telescopes	28	60.71	80.95	69.39
r/psychology	16	37.50	85.71	52.17
r/scienceteachers	16	0.00	0.00	0.00
r/ladiesofscience	23	30.43	31.82	31.11
r/aerospace	21	66.67	93.33	77.78
r/mycology	15	73.33	73.33	73.33
r/botany	15	33.33	83.33	47.62
r/sociology	26	76.92	95.24	85.11
r/rtlsdr	13	69.23	75.00	72.00

Table 17: Science Subreddit Prediction Performance (2/2)

Domain	Examples	Precision	Recall	F1
r/ukpolitics	244	82.38	81.05	81.71
r/politics	131	58.78	67.54	62.86
r/bluemidterm2018	29	72.41	65.62	68.85
r/hillaryclinton	111	78.38	84.47	81.31
r/politicaldiscussion	410	80.24	66.46	72.71
r/askfeminists	105	85.71	78.95	82.19
r/capitalismvsocialism	239	79.50	81.20	80.34
r/anarchism	190	60.00	67.06	63.33
r/sandersforpresident	395	90.13	86.62	88.34
r/jordanpeterson	141	28.37	53.33	37.04
r/srsdiscussion	37	70.27	66.67	68.42
r/samharris	112	81.25	83.49	82.35
r/libertarian	248	65.73	75.46	70.26
r/liberalgunowners	97	81.44	71.17	75.96
r/asktrumpsupporters	148	58.11	58.90	58.50
r/stupidpol	68	42.65	54.72	47.93
r/debatealtright	66	59.09	78.00	67.24
r/askaliberal	136	47.06	40.51	43.54
r/abortiondebate	80	91.25	71.57	80.22
r/yangforpresidenthq	243	94.65	96.64	95.63
r/wayofthebern	96	55.21	76.81	64.24
r/askconservatives	63	69.84	88.00	77.88
r/communism	51	66.67	64.15	65.38
r/goldandblack	39	10.26	57.14	17.39
r/brexit	49	51.02	73.53	60.24
r/gunpolitics	31	25.81	29.63	27.59
r/progun	48	37.50	36.73	37.11
r/canadapolitics	68	82.35	91.80	86.82
r/basicincome	32	37.50	70.59	48.98
r/greenandpleasant	41	78.05	91.43	84.21
r/enoughsandersspam	28	21.43	75.00	33.33
r/debatecommunism	57	80.70	83.64	82.14

Table 18: Politics Subreddit Prediction Performance (1/2)

Domain	Examples	Precision	Recall	F1
r/law	33	84.85	93.33	88.89
r/geopolitics	51	31.37	53.33	39.51
r/askaconservative	45	24.44	55.00	33.85
r/labouruk	40	55.00	81.48	65.67
r/conservative	64	20.31	68.42	31.33
r/enoughtrumpspam	58	56.90	91.67	70.21
r/monarchism	37	78.38	100.00	87.88
r/socialism	157	64.97	59.65	62.20
r/israel	109	88.07	93.20	90.57
r/communism101	89	42.70	59.38	49.67
r/askeconomics	27	66.67	66.67	66.67
r/anarchy101	103	55.34	55.88	55.61
r/asksocialscience	35	37.14	76.47	50.00
r/syriancivilwar	67	91.04	87.14	89.05
r/slatestarcodex	26	80.77	77.78	79.25
r/neoliberal	77	31.17	64.86	42.11
r/tiadiscussion	57	70.18	74.07	72.07
r/centrist	23	60.87	66.67	63.64
r/asklibertarians	31	58.06	78.26	66.67
r/prolife	63	31.75	35.71	33.61
r/prochoice	43	67.44	78.38	72.50
r/debateanarchism	43	25.58	39.29	30.99
r/rightwinglgbt	40	52.50	51.22	51.85
r/fullcommunism	32	87.50	84.85	86.15
r/sino	21	66.67	100.00	80.00
r/femradebates	52	42.31	66.67	51.76
r/ronpaul	23	65.22	65.22	65.22
r/thedavidpakmanshow	19	31.58	75.00	44.44
r/garyjohnson	36	77.78	80.00	78.87
r/completeanarchy	32	9.38	21.43	13.04
r/australianpolitics	20	65.00	72.22	68.42

Table 19: Politics Subreddit Prediction Performance (2/2)

Domain	Examples	Precision	Recall	F1
r/childfree	2412	97.01	95.16	96.08
r/malefashionadvice	477	91.40	92.37	91.89
r/askmen	1839	83.36	76.80	79.95
r/twoxchromosomes	1270	72.68	69.29	70.95
r/askwomen	1352	75.15	69.16	72.03
r/xxfitness	788	94.54	93.12	93.83
r/parenting	1530	92.09	82.83	87.22
r/actuallesbians	951	80.44	75.89	78.10
r/ftm	716	73.04	77.48	75.20
r/daddit	263	45.63	75.47	56.87
r/askmenover30	219	35.62	50.65	41.82
r/bisexual	463	79.05	79.22	79.14
r/femalefashionadvice	618	91.75	86.30	88.94
r/girlgamers	210	99.05	87.39	92.86
r/asktransgender	1343	82.20	65.40	72.85
r/thegirlsurvivalguide	311	41.80	53.06	46.76
r/asexuality	193	89.64	96.11	92.76
r/mensrights	286	85.66	83.90	84.78
r/feminism	39	5.13	22.22	8.33
r/mtf	559	44.36	50.20	47.10
r/ainbow	71	0.00	0.00	0.00
r/lgbt	382	50.52	59.75	54.75
r/menslib	34	23.53	47.06	31.37
r/oney	31	3.23	100.00	6.25
r/genderqueer	56	33.93	65.52	44.71
r/mommit	270	47.04	68.65	55.82
r/mypartneristrans	95	74.74	83.53	78.89
r/askwomenover30	255	39.22	60.61	47.62
r/witchesvspatriarchy	99	63.64	94.03	75.90
r/trans	120	0.00	0.00	0.00
r/butchlesbians	34	52.94	81.82	64.29
r/transsupport	13	0.00	0.00	0.00
r/transvoice	19	5.26	100.00	10.00
r/lesbiangamers	16	0.00	0.00	0.00
r/transpositive	33	0.00	0.00	0.00
r/ladiesofscience	15	80.00	92.31	85.71
r/womenshealth	40	25.00	47.62	32.79

Table 20: Gender / Sexuality Subreddit Prediction Performance