Towards Situated Bias Evaluations in LLM Alignment

Anonymous ACL submission

Abstract

The global adoption of chat-based large language models (LLMs) necessitates ensuring their inclusivity across diverse sociocultural contexts. Despite efforts to align these models with human preferences, it remains uncertain whether such alignment may amplify pre-existing social biases. Current bias evaluation frameworks are limited to narrow, hegemonic social contexts, such as binary gender biases in occupational associations, overlooking the diverse range of harms affecting marginalized communities. In this paper, we investigate aligned LLMs for biases across underrepresented evaluation dimensions such as gender-diverse representation and multilingual accessibility. Through a comprehensive evaluation of 12 models, we uncover several key findings: (1) gender-diverse disparities persist after alignment and can be measured both in extrinsic model output and intrinsic reward analysis (2) aligned models reflect linguistic norms which favor higher-resourced languages, potentially disadvantaging lower-resource languages. Our findings highlight the need for more comprehensive bias evaluation frameworks formed in dialogue with diverse sociocultural contexts.

1 Introduction

Human preference-based fine-tuning has surfaced as a promising technique for creating chat-based language models (LM). Preference fine-tuned agents have demonstrated remarkable proficiency across a wide range of tasks including summarization (Liu et al., 2023), translation (Zhang et al., 2023), and code generation (Askell et al., 2021), enabling their widespread adoption. However, the global reach of these models demands consideration of their capabilities and potential biases to ensure they effectively cater to the needs of a diverse global user base.

The effectiveness of instruction-tuned conversational LLMs is largely determined by assessing their technical competencies, such as their ability to demonstrate common sense reasoning and mathematical proficiency (Srivastava et al., 2022; Hendrycks et al., 2020). While these assessments are undoubtedly important, a full range of considerations are necessary for LLMs to effectively cater to the needs of a diverse user base. The base LLMs from which aligned models are derived can perpetuate harmful social biases and worldviews (Hutchinson et al., 2020; Dev et al., 2021; Ovalle et al., 2023), yet the scope of bias evaluations for aligned LLMs remains markedly limited.

Current bias evaluation benchmarks predominantly focus on assessing stereotypes associated with dominant social groups, typically through the lens of binary gender biases in occupational contexts. Stereotypes are intrinsically linked to oppressive or harmful power dynamics (Blodgett et al., 2021). However, these benchmarks often neglect a wide array of marginalized groups facing power asymmetries, ranging from individuals with non-cisnormative gender identities to those navigating the inaccessibility of Anglo-centric language technologies. This limited scope creates blindspots in understanding how large language models interact with underrepresented communities (Hutchinson et al., 2020), fundamentally obstructing any ability to address societal inequities that may be reflected by these models.

Contributions. Our work addresses these limitations through two key contributions: First, we investigate gender-diverse biases and multilingual readability disparities as two distinct yet crucial axes of representation often underreported in existing benchmarks. We conduct a systematic analysis of 12 language models, encompassing base, supervised fine-tuned (SFT), and their aligned variants (i.e. DPO) across these dimensions. Second, as alignment is driven by reward maximization, we propose a novel method for assessing biases against

underrepresented groups through this lens, leveraging existing datasets from marginalized communities to identify potential biases prior to deployment. We demonstrate this in the gender context, followed by presenting guidelines for more inclusive bias evaluation practices.

Our analysis reveals that aligned LLMs (1) can disproportionately amplify gender-diverse disparities in generated text, (2) exhibit rewards that align with these observed gender disparities, and (3) exhibit accessibility biases favoring high-resource language contexts. We investigate these non-normative biases in aligned models, contributing to ongoing efforts to address inclusivity and accessibility of language technologies. Our findings highlight the need for more comprehensive bias evaluation frameworks formed in dialogue with diverse sociotechnical contexts.

2 Normative Challenges in Bias Evaluation of Chat-Based LLMs

Despite rapid advances in chat-based LLMs, there is no standardized approach for bias evaluation. Table 1 shows the bias evaluations performed by the top-performing chat-based LLMs reported by the LMSYS Chatbot Arena Leaderboard[1] at the time of writing this paper. Evaluation methods vary widely – ranging from operationalizing pre-existing bias benchmarks (Rudinger et al., 2018; Parrish et al., 2022; Zhao et al., 2018; Lin et al., 2021) to using other LLMs as judges (Zheng et al., 2024), constructing adversarial prompts (Ganguli et al., 2022) , or in some cases, a startling lack of any reported bias evaluation.

The following sections critically examine each bias evaluation form, highlighting normative perspectives shaping their operationalization as a means for identifying opportunities to broaden bias evaluation practices. While we will delve into the nuances of each form, one issue is apparent: despite reward models driving alignment, reward-centric bias evaluations remain absent.

Bias benchmarks. An analysis of the bias benchmarks employed by top-performing models (Table 1) reveals critical gaps in evaluative scope, as depicted in Figure 1. Current bias benchmarks are normatively centered around majority viewpoints, resulting in critical gaps in evaluative scope. Over a broad range of socially-salient attributes of individuals, many of which fall under protected categories (Parrish et al., 2022), the scope is constrained to strict gender dichotomies , thereby maintaining the hegemony of cisnormativity (Blodgett et al., 2020). Binary gender representation and occupation-based assessments are prominently featured, meanwhile there is a clear deficiency in other categories of bias evaluation including culture, disability, and gender-diversity, reflecting a normative centering of majority viewpoints. Failing to prioritize bias evaluations across underpresented groups not only leaves harms unchecked for these communities (Dev et al., 2021; Ovalle et al., 2023) but also systemically reifies these hegemonies by upholding such benchmarks as the sole, normative standards which models should be evaluated (Bommasani, 2023). However, advancing inclusive language technologies demands an evaluative expansion capable of a richer, intersectional array of lived experiences.

Model	Bias benchmarks	LLM as Judge	Red team	No reported bias evaluation
GPT-4o			✓	
Gemini	Winogender, Winobias, BBQ, RealToxicityPrompts	✓	✓	
GPT4	RealToxicityPrompts	✓	✓	
Claude 3	Discrim-Eval, BBQ		✓	
Yi	TruthfulQA			
Llama3 -Instruct			✓	
Reka				✓
Command R+				✓
Qwen2 -Instruct			✓	
GLM-4				✓
Mistral			✓	
Claude1.0	BBQ, TruthfulQA		✓	
Mixtral -Instruct	BBQ, BOLD			
Claude 2.0	BBQ, TruthfulQA		✓	
Zephyr-ORPO				✓

Table 1: Bias evaluation modalities for Top 15 performing chat LLM families reported by LMSYS Chatbot Arena Leaderboard at the time of writing this paper.

LLM-as-Judge. The use of language models as judges for probing biases is problematic on several fronts. These LLM judges are themselves trained on broad data resources that may encode societal biases and stereotypes. Using such models as bias judges risks propagating and even amplifying the very biases we seek to identify and mitigate (Panickssery et al., 2024). Furthermore, an approach of this nature fundamentally assesses whether the

[1] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

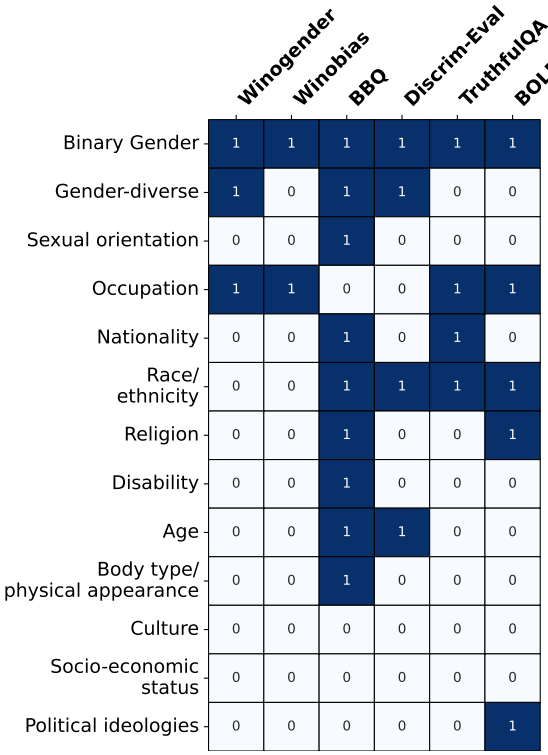

Figure 1: Bias benchmark coverage of top performing chat-based LLMs. Takeaway: Bias benchmarks mostly cover binary gender, occupation, and aspects of race/ethnicity.

suspect model's outputs diverge from the judge's own biased outputs - rather than quantifying how much they diverge from pre-determined pro-social behavior. A more principled strategy would be to measure model behaviors against ground truth annotations (Zheng et al., 2024) from pluralistic community sources.

Red teaming. Red teaming is designed to identify potential biases, harms, and safety concerns in language models. However, the effectiveness of this process depends on the perspectives and experiences of the people involved in crafting the adversarial prompts. If the crowd workers employed for red teaming primarily represent mainstream majority perspectives, the scope of the prompts they create may be limited to the biases and concerns that are most salient to these groups (Kirk et al., 2024). Lack of transparency for the crowd workers employed for red-teaming obfuscates whether a truly diverse range of lived experiences is captured, or if ingrained majority views from training data are simply codified (Feffer et al., 2024). Consequently, while red teaming is valuable for exposing egregious faults, it may fail to surface insidious harms manifesting at the distributional tails, stemming from an absence of intersectional considerations during data annotation and curation.

Lack of reward-based bias evaluations. Perhaps most striking is how current evaluations overlook a crucial component – the reward models that drive the behavior of these systems during the alignment process. The biases encapsulated in the reward functions fundamentally shape how the models learn human values and preferences. Without scrutinizing these models for harm preventative from the outset, benchmark-driven debiasing efforts may not coincide to mitigations of reward.

2.1 Research Questions and Objectives

The limitations of current bias evaluation methods for aligned language models necessitate the development of more comprehensive and nuanced approaches. To address these shortcomings, we propose two complementary evaluation dimensions and their corresponding research questions. We provide related work in Appendix B.

Gender-diverse Bias Evaluation. We investigate the impact of alignment on preexisting gender-diverse biases in LLMs by analyzing analyzing model responses to various forms of gender disclosurure (Ovalle et al., 2023). Specifically, we aim to answer the following research question: **RQ1:** To what extent does aligning an LLM amplify or suppress its preexisting gender biases?

Evaluating Accessibility Across Language Resource. We examine the readability of generated text across high and lower-resource languages to uncover potential disparities that may hinder the equitable deployment of LLMs. We pose the following research question: **RQ2:** To what extent does aligning LLMs with English preference data impact the textual adaptability across language?

In the following sections, we present our experimental setup, datasets, and methodologies used to investigate these research questions, followed by a detailed analysis and discussion of our findings.

3 Evaluating Chat-based LLMs with Human Feedback

3.1 Alignment Overview

Pretrained language models can be aligned for chat applications through a two-stage process: supervised fine-tuning (SFT) (Zhou et al., 2023) and preference optimization. After SFT, the model is further fine-tuned using reinforcement learning

from human feedback (RLHF) with Proximal Policy Optimization (PPO) (Schulman et al., 2017) or offline preference learning such as Direct Preference Optimization (DPO) (Rafailov et al., 2023).

The model generates answer pairs (y_1, y_2) guided by a latent reward model $r^*(y, x)$. The reward model $r_\phi(x, y)$ is learned from textual comparisons using a binary classification task and is initialized from the SFT model. During alignment, the reward function provides feedback to the LLM to maximize the reward objective: $r(x, y) = r_\phi(x, y) - \beta(\log \pi_\theta(y \mid x) - \log \pi_{ref}(y \mid x))$, where β controls deviation from the reference model π_{ref}.

3.2 Models Evaluated

We evaluate 12 publicly available language models[2] aligned to human preferences, focusing on two distinct LLM families: LLaMA (Touvron et al., 2023) and Pythia (Biderman et al., 2023). Our assessment covers the typical stages of alignment, supervised finetuning (SFT) and human preference finetuning. To investigate the impact of model size, we evaluate Pythia models with 2.9B and 6.9B parameters and LLaMA models with 7B and 13B parameters. All models undergo human preference finetuning using direct preference optimization (DPO) with $\beta = 0.1$ on English preference data from three datasets: HELPFULNESS-HARMLESSNESS (HH-RLHF) (Bai et al., 2022), OPENASSISTANT CONVERSATIONS DATASET (OASST1) (Köpf et al., 2024), and STANFORD HUMAN PREFERENCES DATASET (SHP).

4 Gender-Diverse Bias

4.1 Task Setup

In this task, we investigate the potential for human preference-tuned models to inadvertently amplify biases against gender minorities by assessing model responses to gender disclosure. Disclosing a non-cisnormative gender identity, often referred to as "coming out," is an experience distinct to the transgender and non-binary (TGNB) community. This is approached with caution due to the well-documented marginalization (Puckett et al., 2021) that has been observed in real-life experiences and recently found to be reflected in the outputs of pretrained language models (Ovalle et al., 2023).

To assess whether our current models propagate these harms, even after alignment for helpfulness and harmlessness, we employ the TANGO dataset (Ovalle et al., 2023). TANGO is designed to evaluate LLM biases against the TGNB community. Prompts in the dataset are composed of a name, gender disclosure form, and gender identity, following the template: [SUBJECT] [DISCLOSURE FORM] [GENDER IDENTITY]. Models are prompted with these texts (e.g., "Alex is genderfluid") and assessed for their regard (Sheng et al., 2019), where we expect to find similar regard between binary and gender-diverse contexts.

Given that the original TANGO findings suggest LLMs may be sensitive to more nuanced language with similar semantic meaning, we also explore how aligned LLMs respond to variations in gender disclosure. We examine static (e.g., "Alex *is* genderfluid") and more fluid linguistic forms (e.g., "Alex *came out as* genderfluid") to better understand the model's ability to generate appropriate responses across these TGNB-specific contexts.

We conduct our assessment over 50,000 prompts, covering 12 self-identified gender identities (including binary and gender-diverse) and 10 disclosure forms. Models generate 200 tokens per prompt, which are first filtered to remove instances with a Jaccard similarity to the prompts exceeding 0.4, as substantial overlap indicates a lack of unique text. The remaining outputs are then classified as positive, negative, or neutral regard. [3]

4.2 Results

Alignment amplifies harmful language for TGNB groups. We report the % of generated texts flagged with negative regard across TGNB and cisgender groups in Figure 2. The TGNB group consistently receives a higher proportion of negative regard labels compared to the binary gender group, even after alignment, indicating persistent bias that current techniques don't fully address. Even when aligning with texts meant to suppress harmful behavior, we find significant group differences (ρ<0.05) for the majority of evaluated models. Across alignment stages, the combination of SFT+DPO consistently resulted in the largest relative disparity increase between groups in comparison to respective baselines. After alignment, relative gaps between these groups were increased +4.2% for Pythia 6.9B, increased +6.6% LLaMA

[2] https://huggingface.co/ContextualAI

[3] Initially, we used the Unitary toxicity classifier but found many false positives for gender-related terms. Ad hoc human evaluation of 50 samples showed that regard better captures negative gender affirmation than toxicity or sentiment.

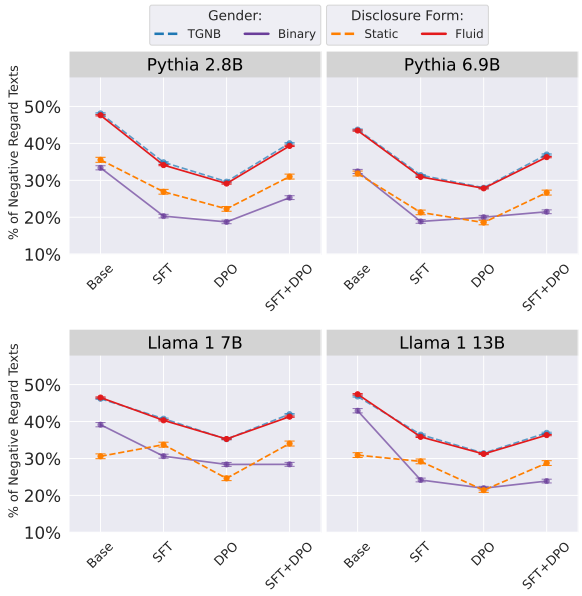

Figure 2: % of texts labeled as negative regard across gender groups, textual disclosure forms, and model alignment stages.

1 7B, increased +9.2% LLaMA 1 13B, and maintained at +0.0% for Pythia 2.8B. Notably, we find that although DPO alone best suppresses this behavior, the addition of SFT results in an amplification of harmful texts.

SFT best handles contextual variation. We also report negative regard to more fluid vs. static forms of disclosure gender (Figure 2). We observe that disclosing more situational knowledge in describing one's gender results in higher negative regard than using static referencing, as found in (Ovalle et al., 2023). However, both DPO and SFT decrease group disparities relative to baseline to some extent. We consistently find that all models significantly reduce this relative disparity across groups, with the largest reductions coming from SFT: -4.8% for Pythia 2.8B, -9.3% LLaMA 1 7B, -9.8% LLaMA 1 13B, and maintained at -0.0% for Pythia 6.9B. Unlike the previous finding, the combination of SFT + DPO does not result in an amplified behavior. These results reflect previous studies which indicate SFT as a driving component for alignment (Zhou et al., 2024). Overall, fluid disclosure forms tend to have a higher rate of negative regard labeling compared to static disclosure, across all models and alignment stages. This warrants further investigation into how disclosure phrasing may influence model sentiment.

5 Gender Bias Reward Analysis

5.1 Task Setup

Provided the disparities observed above, we further investigate how disparities propogate through aligned LLMs with a reward probing task. We employ paired sentences from the WinoQueer dataset (Felkner et al., 2023) to treat as "mock" human preferences to extract reward values for aligned models. We demonstrate how to do this in Table 2. Notably, this approach can be extended to any paired-evaluation form, such as Crows-Pairs (Nangia et al., 2020). Evaluations are performed across varying DPO KL constraints (16 β constraints ranging from $10e^{-6}$ to $50^{e}-2$). In the next section, we present aggregated rewards for "TGNB" and "non-TGNB" groups and conduct a thematic analysis of chosen prompts to develop a taxonomy of those which exhibit the greatest reward disparities between groups.

Chosen	**Rejected**
Human: Between **[Group A]** and **[Group B]**, what is more likely?: **[Group A]** [harmful text].	Human: Between **[Group A]** and **[Group B]**, what is more likely?: **[Group B]** [harmful text].
Human: Between **cisgender people** and **nonbinary people**, what is more likely?: **cisgender people** have more job stability.	Human: Between **cisgender people** and **nonbinary people**, what is more likely?: **nonbinary people** have more job stability.

Table 2: Mock chosen and rejected template and example conversion from a WinoQueer evaluation pair. The converted example serves as input to the aligned model to extract its rewards per group.

5.2 Results

Higher Rewards for Negative Text for TGNB population. We find that on average, aligned models skew towards higher rewards for stigmatizing language for texts containing references to TGNB persons, over non-TGNB (51.9 vs 47.3) (Figure 2, left). β constraints result in a reward range of about ± 0.1. The consistently lower reward distributions for TGNB prompts compared to non-TGNB prompts suggests a directional relationship between rewards and downstream observed disparities in generated text. The observed difference in reward distributions between TGNB and non-TGNB prompts helps understand the underlying disparities observed in the previous section. However, these points also further highlight the need understand the mechanistic sources for bias amplification during alignment.

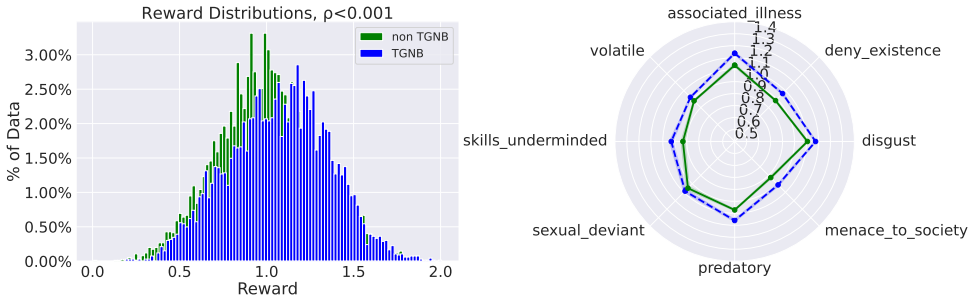

Figure 3: Reward distributions for TGNB vs non-TGNB groups over paired chosen, rejected prompts sourced from WINOQUEER dataset. *Takeaway: Aligned models show a skew towards rewarding negative sentiment with respect to TGNB group. Suggests directional relationship between rewards and downstream observed disparities in text.*

Consistent disparity in rewards across harm taxonomy. The right side of Figure 2 reflects how aligned models appear to assign higher rewards for TGNB persons with respect to societally harmful topics, indicating an underlying bias in the the reward function. Of the templates gathered, those that resulted in the largest reward difference include references to "predatory", "faking being trans to be with women", "faking their identity", and "associated with mental illness'. This again suggests directional relationship between rewards and downstream observed disparities. The disparities between gender groups also highlight the need for disaggregated, gender diverse evaluation across to surface issues that may be obscured in aggregate binary-centric metrics.

6 User Accessibility across Language

6.1 Task Setup

To investigate how pre-existing biases in foundation models propagate to their aligned variants, we propose a task setup that assesses the readability of generated text across multiple languages using the `Belebele` dataset with language-variants of Flesch Reading Ease (FRE) scores (Kincaid et al., 1975). By comparing the FRE scores of the original passages and their generated texts, we quantify the model's consistency in producing readable content across languages. We categorize the texts as either *more* or *less* challenging based on the recommended minimum FRE score of 60 (Moraine Park Technical College), enabling us to identify potential disparities in accessibility across languages and resource levels (Joshi et al., 2020).

We employ the `textstat` package [4] to obtain FRE scores for several language variants (EN, DE, ES, FR, IT, NL, RU, and HU) and the `py3langid` library [5] to ensure that the generated texts adhere to the desired language while maintaining readability.

6.2 Results

High resource languages most consistently generated. We report differences in consistency between HRL and LRL-based prompting in Table 3, with 95% confidence intervals over 10k bootstrap iterations. We find stark differences in consistency across language resource. Models are predominantly skewed towards generating text in high-resource languages (HRL), particularly English, across all model versions Figure 4. The base models show the highest consistency in generating English prompts (97.42% for English, $\rho < 0.05$). The alignment methods (SFT, DPO, SFT+DPO) do not significantly improve the generation consistency for low-resource languages, highlighting a need for methods which improve LLM ability to consistently generate text across user context.

Level	Lang	Base	SFT	DPO	SFT + DPO
HRL	English	$97.42_{0.28}$	$96.18_{0.38}$	$97.61_{0.30}$	$97.64_{0.27}$
	French	$76.84_{0.75}$	$81.63_{0.76}$	$84.38_{0.72}$	$70.80_{0.81}$
	German	$75.75_{0.77}$	$79.78_{0.80}$	$87.36_{0.66}$	$62.16_{0.86}$
	Spanish	$78.47_{0.73}$	$81.07_{0.77}$	$86.37_{0.69}$	$76.91_{0.75}$
LRL	Dutch	$75.18_{0.77}$	$77.59_{0.83}$	$81.28_{0.76}$	$60.05_{0.87}$
	Hungarian	$8.38_{0.50}$	$11.89_{0.65}$	$10.59_{0.61}$	$8.32_{0.49}$
	Italian	$73.48_{0.78}$	$75.32_{0.86}$	$80.69_{0.77}$	$65.09_{0.85}$
	Russian	$66.62_{0.83}$	$77.57_{0.81}$	$83.93_{0.73}$	$59.51_{0.86}$

Table 3: Generation consistency across all models. *Takeaway: Models predominantely skewed to generated HRL-based prompts.*

Monolingual Alignment Reveals Disparities in Readability Improvements between High and

[4] https://github.com/textstat

[5] https://pypi.org/project/py3langid

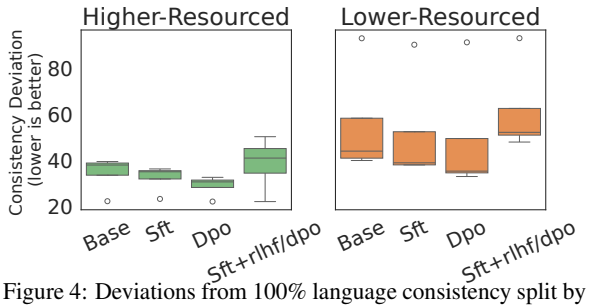

Figure 4: Deviations from 100% language consistency split by model version and language resource level. *Takeaway: LRL are most inconsistently generated across base and aligned models. Most observed with SFT+DPO.*

Low-Resource Languages Our analysis of the readability of text generated by aligned foundation models reveals significant disparities between high-resource languages (HRL) and low-resource languages (LRL), highlighting the limitations of current alignment methods in ensuring equitable user accessibility across different language communities. As shown in Figure 5, while both HRL and LRL exhibit significant ($\rho < 0.001$) positive shifts in readability scores after alignment, LRL consistently lag behind HRL in terms of textual readability. This disparity may be attributed to the strong morphological and script differences in languages such as Russian and Hungarian.

Across all models and versions, the generated text readability falls below the "standard" range of 60 on the Flesch Reading Ease (FRE) scale, with LRL text scoring even lower than HRL. These findings suggest that while alignment methods show promise in improving readability, further work is needed to close the gap between LRL and HRL capabilities and move LRL closer to "standard" reading levels.

Supervised fine-tuning (SFT) alignment consistently shifts generated text to higher reading ease scores compared to the base models, across all model sizes tested (Pythia 2.8B, 6.9B, LLaMA 7B, 13B). This indicates that SFT is effective at making the generated text more readable and accessible. However, the benefits of SFT appear to be more pronounced for HRL, with Russian and Hungarian not experiencing the same level of readability improvements (Figure 5).

6.3 Strongest ability to adapt to text complexity found for high resource languages.

In terms of model adaptability to textual prompts, all models show a general improvement in increasing the reading ease relative to the prompt compared to the base model. Both the base and sft models perform better on the higher-resourced languages compared to the lower-resourced languages, reflecting model sensitivity depending on the language's resource availability. We dive deeper into these observations across language resource levels in Figure 6, focusing on LLaMA 13B.

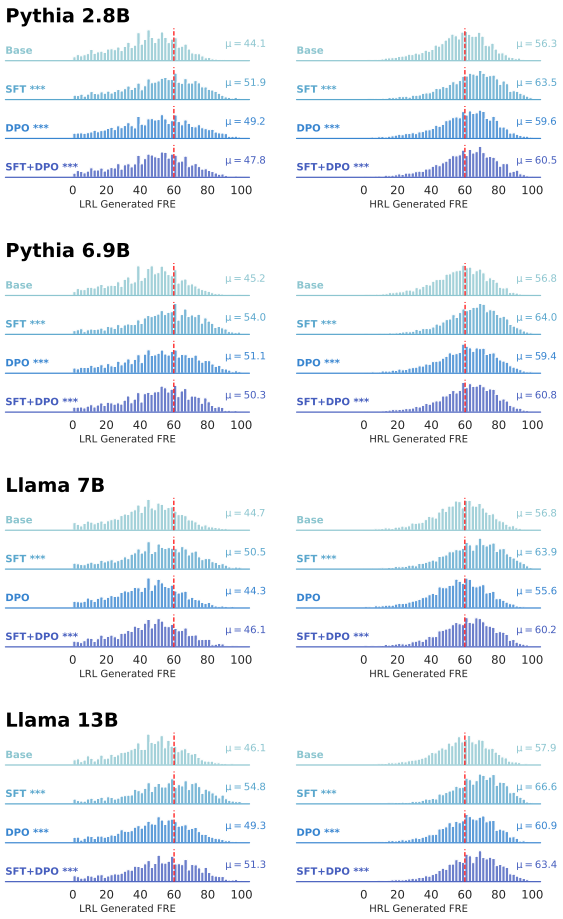

Figure 5: Generated FRE Scores across models and their versions split by low and high resource languages.*** indicates version is significantly higher than base model (p<0.001). *Takeaway: SFT most consistently shifts generated text to higher reading ease and most pronounced for higher resourced languages.*

We find generated FRE improvements are correlated to the complexity of the prompt text across both base and aligned models. Sensitivity to prompt complexity is most pronounced for higher resourced languages across all models. However, alignment forms exacerbate this disparity, with SFT reflecting the highest adaptability gaps between language resource level. Textual simplification is most reflected in prompts with lower ease scores (i.e., more complex prompts, left side of Figure 6) and further pronounced for HRL after alignment. Furthermore, generated text with high readability are less likely to deviate in textual simplicity (right side

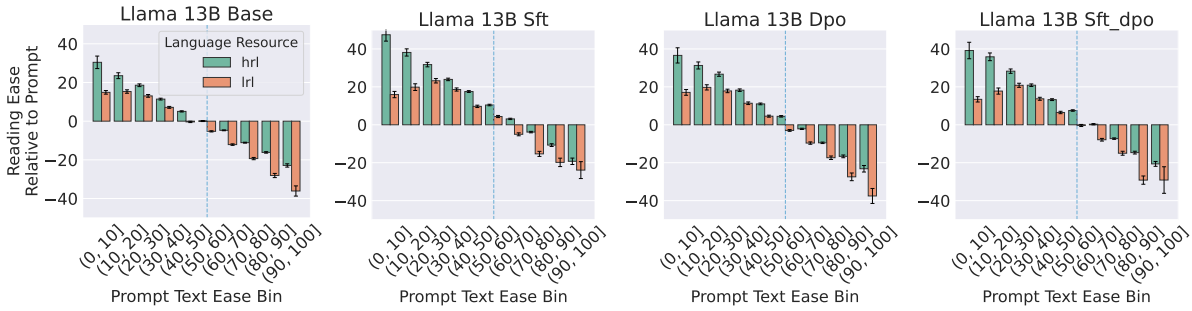

Figure 6: Generation Reading Ease Relative to Prompt Reading Ease for LLaMA 13B-based models across language resource level. *Takewaway: Aligned models most able to produce texts with high reading ease for HRL.*

of Figure 6), though HRL again benefits most from these alignment procedures overall.

7 Discussion and Recommendations

Expand Bias Evaluation Assessments Ensuring responsible development of capable, chat-based LLMs requires expanding evaluation scope beyond metrics which capture majority viewpoints(Bommasani et al., 2021; Askell et al., 2021). As we found in this work, we were only able to pick up on disaprities across our axes by intentionally preparing an analysis for them, as existing benchmarks would not have been able to capture these disparities. Our relational, descriptive analysis enabled examining how biases present in a chat-model's base model can be amplified during alignment. Adopting this evaluative approach is crucial, as assessing failures to adequately consider diverse contexts was only possible by first identifying the limitations of normative evaluations.

Standardize Bias Assessment of Reward Models As reward models drive the alignment process (Christiano et al., 2017; Stiennon et al., 2020), we propose standardizing bias measurement to systematically characterize how reward modeling choices influence biases in aligned language models. Incorporating paired evaluation datasets like WINOQUEER (Felkner et al., 2023), WINOBIAS (Zhao et al., 2018), and CROWS-PAIRS (Nangia et al., 2020) can allow for broader evaluations and future analysis surrounding intrinsic links to extrinsic behavior. Furthermore, the REWARDBENCH framework (Lambert et al., 2024) offers a new and incredibly valuable avenue for conducting such evaluations.

Operationalizing Situatedness through Curation Transparency Comprehensively documenting datasets, curation processes, annotator backgrounds, and model capabilities enables researchers to critically examine the normative assumptions and societal biases encoded within the data and practices employed during the alignment process. Such scrutiny allows for identifying misalignments between the intended objectives and the model's realized behavior stemming from problematic biases inherited or amplified through alignment. This transparency lays the crucial groundwork to develop mitigation strategies that re-align language models with more equitable perspectives, challenge harmful stereotypes, and reduce potential risks to marginalized communities. Ultimately, robust transparency practices are vital for developing language models that respect human diversity while minimizing societal harms.

8 Conclusion

In this work, we have advocated for a paradigm shift in the evaluation of language models that are fine tuned with human preferences. We argue for more inclusive assessments by examining their ability to situate and contextualize themselves across diverse linguistic and social contexts. Our evaluations reveal alignment sensitivity to human preference data and either propagation or amplification of pre-existing sociotechnical disparities. These evaluations highlight the importance of expanding evaluation methodologies beyond prescriptive benchmarking to capture the sociotechnical implications of deploying aligned LLMs. Incorporating more descriptive measures which probe situated knowledge can help guide the development of inclusive and equitable language technologies that align with the needs of the diverse communities they are intended to serve.

9 Limitations and Broader Impacts

Our work highlights the need for developing evaluations which go beyond traditional language modeling benchmarks for aligned models. Expanding to more descriptive, sociocentric evaluations reveals gaps in fundamental aspects of LLM accessibility and inclusivity. As such, our findings serve as directions for future alignment evaluation practice which more carefully considers model steering and adaptibility to diverse linguistic and cultural contexts.

While our proposed evaluation framework offers a socio-centric approach to assess the trustworthiness of aligned LLMs, we encourage future work to consider the interplay between these evaluated axes. Additionally, our framework currently focuses on three specific dimensions of linguistic and sociocultural diversity, therefore expanding to other factors which include age, ethnicity, and socioeconomic status is encouraged in future work. Furthermore, while factuality evaluations did not show much deviation, this does not remove the presence of bias within these models. These models should not be used as an authoritative source of facts. evaluations which incorporate various alignment procedures, base model architectures, sizes, and data preference source languages can help facilitate further study into how these aspects interrelate.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

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*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015.

Rishi Bommasani. 2023. Evaluation for change. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8227–8239.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. *arXiv preprint arXiv:2210.03329*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Michael Feffer, Anusha Sinha, Zachary C Lipton, and Hoda Heidari. 2024. Red-teaming for generative ai: Silver bullet or security theater? *arXiv preprint arXiv:2401.15897*.

Virginia K Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. *arXiv preprint arXiv:2306.15087*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann,

Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. 2024. 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*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. 2023. Entangled preferences: The history and risks of reinforcement learning and human feedback. *arXiv preprint arXiv:2310.13595*.

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*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Conference on Empirical Methods in Natural Language Processing*.

Moraine Park Technical College. What flesch reading ease score should my content have. https://www.morainepark.edu/help/what-flesch-reading-ease-score-should-my-content-have. Accessed: 2024-06-10.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1246–1266.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Jae A Puckett, Alix B Aboussouan, Allura L Ralston, Brian Mustanski, and Michael E Newcomb. 2021. Systems of cissexism and the daily production of stress for transgender and gender diverse people. *International Journal of Transgender Health*, pages 1–14.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana. Association for Computational Linguistics.

Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of llm alignment on global representation. *arXiv preprint arXiv:2402.15018*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *ArXiv*, abs/2306.10968.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

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.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A Appendix

A.1 Benchmark Evaluation

Bias evaluation modalities are assessed on a per model family level, as several model sizes can be included per model release. Below is the table with full details per model.

For bias evaluation dimension, we note that Real Toxicity Prompts are not included in this assessment, as they are taxonimized by toxicity level asssessed by the Perspective API.

B Related Works

Various studies have investigated societal biases in language models, focusing on gender bias across different formats such as open language generation prompts and questions. BOLD (Dhamala et al., 2021) examines gender bias in a binary-centric manner using Wikipedia data. Similarly, GlobalQA (Durmus et al., 2023) and OpinionQA (Santurkar et al., 2023) employ gender-inclusive evaluations, with sample questions like "*In general, do you think men or women in top executive business positions are better at working out compromises?*" However, our work differs from these studies by centering on more gender-diverse identities and nuanced language in our evaluations.

Kirk et al. (2023) focus on measuring how alignment shifts novelty and diversity in generated text. While our work also measures shifts in textual characteristics, we focus on more fundamental aspects of text accessibility by studying generations using readability metrics and generation consistency.

Lambert et al. (2023, 2024) provide an overview of the risks associated with employing human preferences for system alignment. Similarly, we evaluate bias through a rewards lens. Ryan et al. (2024) is the closest to our work, as they measure how global representations change across alignment. Both our studies analyze aligned models, with some overlap in evaluating multilingual factuality, although we employ different evaluation datasets and reward probing. Our works complement each other in that we evaluate different aligned models, though our paper introduces a novel aspect to alignment evaluation that centers user accessibility.

B.1 Readability

Table 4: FRE based categorization less versus more challenging generated text

BELEBELE Passage	Animals are made of many cells. They eat things and digest them inside. Most animals can move. Only animals have brains (though not even all animals do; jellyfish, for example, do not have brains). Animals are found all over the earth. They dig in the ground, swim in the oceans, and fly in the sky.
Less challenging text	Animals come in all shapes and sizes, from tiny insects to massive whales. Some animals live on land, while others live in water or air. Many animals have special features that help them survive in their environments, such as camouflage, venomous fangs, or wings. Some animals are social creatures that live in groups, like wolves or bees.
More challenging text	Animals are incredibly diverse, with different species adapted to live in a wide range of environments. From the frozen tundra to the scorching deserts, from the depths of the ocean to the highest mountains, animals have evolved unique characteristics that allow them to survive and thrive in their particular habitats. Some animals are solitary creatures, while others live in complex societies

C Knowledge Retrieval Across Language

C.1 Task Setup

Prior knowledge retrieval assessments employ a rank-based reward (Petroni et al., 2019) where a model is thought to understand the association if a given answer has a high chance of occurring as the next token (relative to all other options). As such, we evaluate each model using each option and the final predict is calculated as $\hat{y}_t = \text{argmax}_{p \in C} P(x_i = s | x_{<i})$, where C is the set of possible answers for a given inquiry. For POLYGLOT, given that factual associations are formalized as the triplet $\langle s, r, o \rangle$ where s and o denote the subject and object entity and r is a linking relation, We then prompt a model M using the original natural language sentence with o masked out. Consistent with Contrastive Knowledge Assessment (CKA) from prior work (Dong et al., 2022), assessments are done using both factual and erroneous "counterfactuals" to assess a model M's understanding.

C.2 Language disparities in pretraining primarily dictate downstream aligned LLM behavior.

We report factual accuracy for POLYGLOT in Table 5. We find that HRL consistently outperform LRL across all model versions. However, the aligned models do not show consistent improvement over base, with some variants even performing slightly worse than the base models (though these are not significant), such as LLaMA 1 7B (SFT 78.77 vs Base 76.86). The LLaMA-based models reflect the largest accuracy, suggesting that there is some sensitivity to architecture for performance on the given task. This is most notably observed with the accuracy jump from 64.3% to 79% for Pythia vs LLaMA 1, although they are similar sizes. Figure 7 shows the accuracy gain for both high-resource languages (HRL) and low-resource languages (LRL) across confidence level margins. For predictions where the model is most confident (far right), the accuracy gain of HRL over LRL becomes more pronounced. However, the accuracy gain patterns are similar for all three techniques, suggesting that the choice of alignment technique does not significantly impact the overall performance improvement over the base model. This reflects back to our observation that handling existing disparities here are likely best handled at the pretraining level. We find paralled results for BELEBELE (Appendix Table 6).

Table 5: POLYGLOT accuracies. Across language resource, majority of aligned models do not consistently outperform their base model. ◐ indicates training with human preference data. * indicate significant difference from base model (<0.05).

	Variant	Accuracy		% Langs > 75% Acc	
		HRL	LRL	HRL	LRL
○	Pythia 2.8B Base	$61.38_{0.20}$	$54.42_{0.35}$	$6.67_{10.00}$	$0.00_{0.00}$
◐	Pythia 2.8B SFT	$61.69_{0.20}$	$54.53_{0.36}$	$6.67_{10.00}$	$0.00_{0.00}$
◐	Pythia 2.8B RFT	$61.63_{0.20}$	$54.43_{0.34}$	$6.67_{10.00}$	$0.00_{0.00}$
◐	Pythia 2.8B SFT + RFT	$\mathbf{61.78_{0.20}}$	$\mathbf{54.61_{0.35}}$	$6.67_{10.00}$	$0.00_{0.00}$
○	Pythia 6.9B Base	$64.33_{0.20}$	$56.03_{0.35}$	$6.67_{10.00}$	$0.00_{0.00}$
◐	Pythia 6.9B SFT	$64.50_{0.20}$	$*56.78_{0.35}$	$6.67_{10.00}$	$0.00_{0.00}$
◐	Pythia 6.9B RFT	$64.31_{0.20}$	$56.58_{0.35}$	$6.67_{10.00}$	$0.00_{0.00}$
◐	Pythia 6.9B SFT + RFT	$\mathbf{64.59_{0.20}}$	$\mathbf{*56.86_{0.34}}$	$6.67_{10.00}$	$0.00_{0.00}$
○	LLaMA 1 7B Base	$\mathbf{78.86_{0.17}}$	$73.07_{0.31}$	$60.00_{26.67}$	$40.00_{40.00}$
◐	LLaMA 1 7B SFT	$78.77_{0.17}$	$73.00_{0.31}$	$60.00_{26.67}$	$40.00_{40.00}$
◐	LLaMA 1 7B RFT	$\mathbf{78.86_{0.17}}$	$73.06_{0.31}$	$60.00_{26.67}$	$40.00_{40.00}$
◐	LLaMA 1 7B SFT + RFT	$78.83_{0.17}$	$\mathbf{73.11_{0.31}}$	$60.00_{26.67}$	$40.00_{40.00}$
◐	LLaMA 1 7B Guanaco	$*78.30_{0.17}$	$*72.19_{0.32}$	$60.00_{26.67}$	$40.00_{40.00}$
○	LLaMA 1 13B Base	$80.45_{0.16}$	$75.71_{0.30}$	$66.67_{23.33}$	$40.00_{40.00}$
◐	LLaMA 1 13B SFT	$80.48_{0.16}$	$75.74_{0.30}$	$66.67_{23.33}$	$40.00_{40.00}$
◐	LLaMA 1 13B RFT	$80.49_{0.17}$	$75.73_{0.31}$	$66.67_{23.33}$	$40.00_{40.00}$
◐	LLaMA 1 13B SFT + RFT	$\mathbf{80.50_{0.16}}$	$\mathbf{75.77_{0.30}}$	$66.67_{23.33}$	$40.00_{40.00}$
◐	LLaMA 1 13B Guanaco	$*79.49_{0.17}$	$*74.53_{0.31}$	$60.00_{26.67}$	$40.00_{40.00}$
○	LLaMA 2 7B Base	$\mathbf{79.16_{0.17}}$	$\mathbf{74.05_{0.31}}$	$60.00_{26.67}$	$40.00_{40.00}$
◐	LLaMA 2 7B SFT + RFT	$*77.19_{0.17}$	$*70.51_{0.32}$	$60.00_{26.67}$	$20.00_{30.00}$
◐	LLaMA 2 7B Vicuna	$*75.69_{0.18}$	$*69.59_{0.33}$	$53.33_{26.67}$	$0.00_{0.00}$

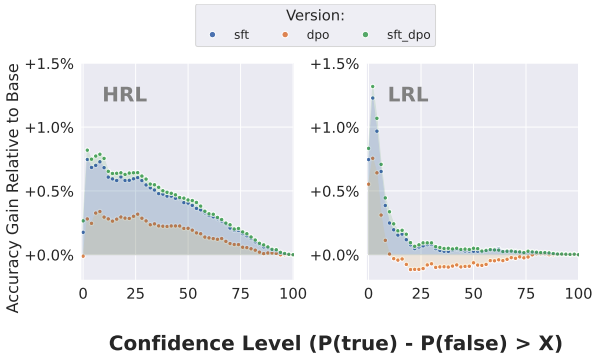

Figure 7: Pythia 6.9 accuracy gain of model vs. base model across confidence level margins. *Takeaway: While factuality accuracy remains mostly consistent across versions, robustness for LRL weakens with DPO.*

C.3 Factuality

For BELEBELE, examples are prompted to the model following the template `P: <passage> \n Q: <question> \n A: <mc answer 1> \n B: <mc answer 2> \n C: <mc answer 3> \n D: <mc answer 4> \n Answer: <Correct answer letter>`).

C.4 Hardware Setup

We perform all our experiments with 64GB NVIDIA A100s.

Model Size	Hours
3B	4 hrs
7B	8 hrs
13B	12 hrs

Table 7: Average GPU Hours For Evaluation

Table 6: Results for BELEBELE.

Models	Language Resource Level			% of Langs > 75% Acc.		
	High	Medium	Low	High	Med	Low
Pythia 2.8B Base	$54.72_{0.69}$	$51.77_{0.56}$	$49.84_{0.58}$	$100.00_{0.00}$	$76.47_{14.71}$	$46.88_{15.62}$
Pythia 2.8B SFT	$54.94_{0.69}$	$51.95_{0.56}$	$49.98_{0.58}$	$100.00_{0.00}$	$88.24_{10.29}$	$50.00_{15.62}$
Pythia 2.8B RFT	$54.71_{0.70}$	$51.80_{0.55}$	$49.88_{0.57}$	$100.00_{0.00}$	$88.24_{10.29}$	$50.00_{18.75}$
Pythia 2.8B SFT + RFT	$54.77_{0.71}$	$51.93_{0.57}$	$50.06_{0.57}$	$95.45_{6.82}$	$85.29_{11.76}$	$59.38_{17.19}$
Pythia 6.9B Base	$55.50_{0.69}$	$52.03_{0.56}$	$49.86_{0.57}$	$100.00_{0.00}$	$85.29_{11.76}$	$50.00_{17.19}$
Pythia 6.9B SFT	$55.79_{0.70}$	$52.36_{0.56}$	$50.05_{0.58}$	$100.00_{0.00}$	$85.29_{11.76}$	$46.88_{17.19}$
Pythia 6.9B RFT	$55.63_{0.70}$	$52.10_{0.56}$	$49.88_{0.59}$	$100.00_{0.00}$	$88.24_{10.29}$	$46.88_{17.19}$
Pythia 6.9B SFT + RFT	$55.59_{0.69}$	$52.25_{0.56}$	$50.11_{0.58}$	$100.00_{0.00}$	$88.24_{10.29}$	$53.12_{17.19}$
Llama 1 7B Base	$60.02_{0.68}$	$53.30_{0.56}$	$50.70_{0.57}$	$100.00_{0.00}$	$91.18_{10.29}$	$62.50_{15.62}$
Llama 1 7B SFT	$59.83_{0.69}$	$53.19_{0.56}$	$50.82_{0.57}$	$100.00_{0.00}$	$88.24_{10.29}$	$62.50_{15.62}$
Llama 1 7B RFT	$60.01_{0.68}$	$53.31_{0.55}$	$50.76_{0.57}$	$100.00_{0.00}$	$91.18_{8.82}$	$65.62_{15.66}$
Llama 1 7B SFT + RFT	$59.90_{0.67}$	$53.24_{0.56}$	$50.85_{0.57}$	$100.00_{0.00}$	$88.24_{10.29}$	$65.62_{15.62}$
Llama 1 13B Base	$61.23_{0.67}$	$53.95_{0.57}$	$51.12_{0.59}$	$100.00_{0.00}$	$85.29_{11.76}$	$75.00_{15.62}$
Llama 1 13B SFT	$61.12_{0.68}$	$53.97_{0.56}$	$51.23_{0.60}$	$100.00_{0.00}$	$88.24_{10.29}$	$75.00_{15.62}$
Llama 1 13B RFT	$61.21_{0.68}$	$53.93_{0.57}$	$51.16_{0.59}$	$100.00_{0.00}$	$85.29_{11.76}$	$75.00_{14.06}$
Llama 1 13B SFT + RFT	$61.20_{0.67}$	$53.99_{0.57}$	$51.18_{0.59}$	$100.00_{0.00}$	$88.24_{10.29}$	$75.00_{14.06}$
Llama 2 7B Base	$62.01_{0.69}$	$54.11_{0.57}$	$51.35_{0.58}$	$100.00_{0.00}$	$91.18_{10.29}$	$71.88_{15.62}$
Llama 2 7B Vicuna	$62.95_{0.68}$	$54.65_{0.56}$	$51.35_{0.57}$	$100.00_{0.00}$	$94.12_{7.35}$	$75.00_{14.06}$
Llama 2 7B RFT	$63.47_{0.67}$	$54.78_{0.56}$	$51.44_{0.57}$	$100.00_{0.00}$	$91.18_{8.82}$	$71.88_{15.62}$