KARMA: Karma-Aligned Reward Model Adaptation

Jared Scott
Tennessee Tech University
Cookeville, Tennessee
jlscott44@tntech.edu

Jesse Roberts
Tennessee Tech University
Cookeville, Tennessee
JtRoberts@tntech.edu

Abstract

Human communication depends on implicit social signals where effectiveness is shaped by tone, context, and conversational norms rather than semantic content alone. We introduce KARMA (Karma-Aligned Reward Model Adaptation), a framework for LLM learning of context-sensitive conversational behavior from large-scale social media data. KARMA trains a reward model on Reddit conversations to predict response valuation conditioned on context, and uses this signal to fine-tune language models via reinforcement learning to improve performance on pragmatics-mediated tasks. Critically, we find that the highest performing reward model does not lead to better downstream model alignment: a reward model relying exclusively on conversational context was a worse predictor of Reddit karma but yielded substantially better downstream performance. We evaluate the effects of KARMA applied to a downstream model with and without direct exposure to the social media data. The resulting models show improved pragmatics-mediated behaviors with largely mitigated undesirable side effects.

1 Introduction

The success of a response depends not only on its semantic correctness, but also on its pragmatic fit with the audience, conversational context, and social expectations of the interaction. Put simply, communication often requires an ability to "read the room"; the pragmatic ability to recognize when humor, empathy, formality, restraint, directness, or other nuance is likely to be most appropriate.

Like most target LLM behaviors (instruction-following, helpfulness, safety, etc), contextually appropriate language is not consistently represented in human data. This paper's objective is to explore the use of alignment techniques and abundant social media data to improve pragmatics-mediated objectives, specifically humor and sentiment, which rely on contextual understanding while mitigating negative impacts. We frame this problem as one of implicit alignment. While recent advances have substantially improved some target behaviors, pragmatics have been less explored.

Online discussion platforms like Reddit provide a naturally occurring source of human preference data through voting systems, which we expect to be sensitive to pragmatic dimensions such as relevance, tone alignment, humor, and situational appropriateness. We introduce (contribution 1) KARMA (Karma-Aligned Reward Model Adaptation), a framework for improving specific pragmatics-mediated behaviors using social feedback derived from Reddit discussions. KARMA learns a reward model over conversational context and candidate responses using Reddit interaction signals and subsequently applies reinforcement learning through Proximal Policy Optimization (PPO) to adapt a language model.

While this is effective, we were concerned that undesirable behaviors (bias, toxcitiy, etc) may also have been impacted by exposure to reddit data. Therefore, we evaluate (contribution 2) the impact of direct exposure to Reddit conversational data during fine-tuning, finding that direct exposure to the Reddit data distribution amplifies undesirable behaviors, whereas reward-mediated adaptation mitigates negative side-effects while preserving core pragmatic benefits. A notable exception is a consistent decrease in factuality, including when the downstream model has no direct exposure to Reddit data.

Finally, we investigate the relationship between reward model capability and downstream model impact (contribution 3). We show that providing Reddit metadata in addition to conversational context damages downstream alignment outcomes even though it improves reward model performance, suggesting that reward model performance is not a good indicator of downstream impact.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 48030–48039
October 24-29, 2026 ©2026 Association for Computational Linguistics

2 Related Work

Recent approaches to aligning large language models (LLMs) have largely relied on instruction tuning and reinforcement learning from human feedback (RLHF), where models are optimized using reward functions derived from curated human preference data (Ouyang et al., 2022; Schulman et al., 2017). While effective, these methods primarily target correctness, safety, and instruction adherence, and do not explicitly model a key aspect of conversational intelligence: the ability to "read the room," i.e., adapt responses to implicit social cues such as tone, context, and situational appropriateness.

Social media platforms provide a scalable alternative source of feedback, where engagement signals such as replies, upvotes, and comment karma act as implicit indicators of response quality. Prior work has explored engagement prediction as either a regression or classification task, demonstrating that successful prediction depends not only on response content but also on surrounding conversational context (Fang et al., 2016; Zayats and Ostendorf, 2018). More recent models incorporate conversational structure and long-range dependencies, further reinforcing that engagement is highly context-dependent and shaped by interaction dynamics within a thread.

At the same time, directly training on social media data introduces significant challenges. Large-scale conversational corpora often contain noise, bias, and toxic content, and models exposed to such data can inherit or amplify these undesirable behaviors (Gehman et al., 2020; Luong et al., 2024). Additionally, engagement signals themselves may reflect platform-specific or demographic biases, complicating their use as alignment objectives.

In contrast to prior work, we frame "reading the room" as an implicit but under-addressed component of alignment, distinct from both explicit preference optimization and generic engagement modeling. Rather than treating it as a surface-level prediction task, we aim to capture the underlying ability to infer what is appropriate given conversational context. We learn a reward model that predicts engagement from context alone, encouraging the model to infer relevance, tone, and situational fit directly from interaction signals. This reward is then used within a reinforcement learning framework to guide alignment. Through this design, we leverage large-scale social feedback while avoiding direct optimization on raw social media data.

Scott and Roberts (2026) describe a similar intent but provide no execution or empirical evaluation.

3 Reward Modeling from Reddit

To model conversational effectiveness within the KARMA framework, we construct a dataset from Reddit discussions using the publicly available Pushshift dataset (Baumgartner et al., 2020). We filter out explicit content, non-text entries, and non-English comments, and retain only posts with multiple substantive responses. This yields approximately 80k posts and 400k comments spanning 14k communities, providing broad topical and stylistic diversity.

Each comment is paired with its parent post and a set of sibling responses to capture conversational context. This representation captures both hierarchical structure (via parent–child relationships) and local competition between responses (via sibling comments). To investigate whether reward models should leverage platform-specific contextual signals, we intentionally construct two reward modeling settings: a generalized setting that excludes Reddit metadata, and a conditioned setting that incorporates such metadata. In the generalized setting, conversations are transformed into a format resembling standard chat interactions. Post–comment pairs are linearized into instruction-style inputs, where the original post is treated as a user prompt and the comment as a candidate response. Contextual information from parent comments is included as preceding turns in a dialogue, yielding a multi-turn chat representation.

The karma signal captures correlated pragmatic dimensions such as humor, relevance, tone, and situational appropriateness, and this richness makes it a candidate signal for learning broad pragmatic competence. Importantly, subreddit identity, timestamps, and author information are removed in this setting. While these features may improve prediction of engagement within Reddit, we hypothesize that they may enable heuristics which may obscure or obviate the pragmatic utility of the model.

By excluding these signals and standardizing inputs into a chat-like format, the reward model is encouraged to learn features that generalize across interfaces and conversational settings.

To adjust for community size, a comment is labeled as rewarding if its engagement reaches at least 40% of its parent comment's score, and as

non-rewarding otherwise. This formulation shifts the objective from predicting absolute popularity to modeling relative conversational success, expected to yield a comparable supervision signal across threads and subreddits.

3.1 KARMA Reward Model

We full fine-tune a base LLaMA-1B (Grattafiori et al., 2024) model on the rewarding/non-rewarding binary classification task. The model takes as input the complete conversational context, including the original post, its parent comment, and relevant sibling responses. We add an MLP classification head which takes the encoder's final-layer hidden states as input and predicts whether a response will be rewarding. Training is performed using binary cross-entropy loss with the AdamW optimizer, a learning rate of 2×10^{-5}, a weight decay of 0.01, and a batch size of 16 over a 90% training split of the dataset, corresponding to approximately 360k examples. The resulting output probability is interpreted as a dense reward signal, reflecting the likelihood that a given response is rewarding including pragmatic considerations. This reward signal is then used to guide PPO-based fine-tuning, serving as the reward model within the KARMA framework.

All data and models are publicly available and licensed for research usage. All experiments were conducted in 20 A100 GPU hours.

3.2 KARMA Reward Model Evaluation

Previous work on Reddit engagement prediction employed a graph-based LSTM to model comment popularity within a single subreddit. That approach incorporated a range of non-textual features, including timestamps and user post history, and found these signals to be highly influential. Performance reached approximately 55% F1 on a seven-class classification task defined over discrete popularity buckets (Zayats and Ostendorf, 2018).

In contrast, our reward model is evaluated on a held-out Reddit comment dataset spanning a diverse range of subreddits. This introduces distributional variation, providing a more stringent test of the model's ability to generalize to heterogeneous community dynamics. On this evaluation set, the model achieves a binary classification accuracy of 0.708, an F1 score of 0.7166, and an AUC of 0.7497, demonstrating solid performance despite relying solely on textual and contextual signals.

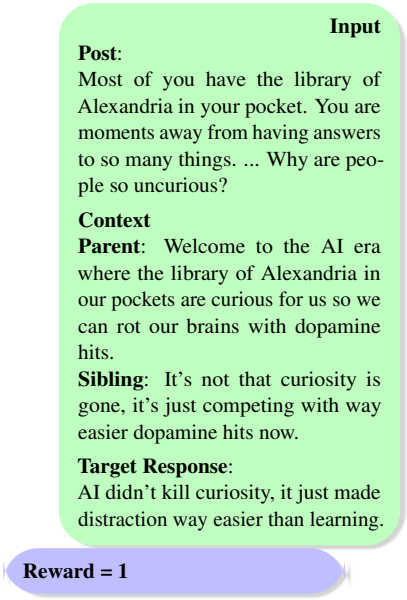


Figure 1: A single training instance from the KARMA Reddit-derived dataset. Each sample consists of a post, hierarchical conversational context (parent and sibling responses), and a candidate target comment used for reward modeling supervision.

4 KARMA Alignment Experiments

We optimize conversational language models using PPO (Schulman et al., 2017), where the learned KARMA reward model provides scalar feedback on the quality of generated responses conditioned on conversational context.

During a PPO training step a conversational input is provided which is formatted to match the same structured representation used for reward model evaluation, ensuring consistency between training and scoring. The model-generated response is appended to this representation as the candidate output and passed to the reward model to obtain the optimization signal. PPO optimization is performed using the AdamW optimizer with a learning rate of 9×10^{-6}, a batch size of 10, and a KL penalty coefficient of 0.5 over approximately 100k training examples. We compare three model variants:

- **Base:** pretrained LLM without alignment
- **KARMA$_{\text{Benign}}$:** PPO-trained on benign chat data using the KARMA reward model
- **KARMA$_{\text{Toxic}}$:** PPO-trained on Reddit conversations using the KARMA reward model

The KARMA$_{\text{Benign}}$ setting isolates the effect of reward optimization on the curated UltraChat-200k dataset (Ding et al., 2023), allowing us to study how a Reddit-trained reward model shapes behavior

apart from direct exposure to potentially toxic Reddit data. In contrast, KARMA$_{\text{Toxic}}$ directly performs PPO using the Pushshift dataset as the training set. Together with the Base model, which serves as an unaligned reference point, these settings enable a controlled comparison between no alignment, pragmatic alignment on benign data, and pragmatic alignment on potentially toxic data. Experiments are conducted across multiple model families and scales, including Falcon (1B, 7B), LLaMA (8B), and Pythia (70M, 1B, 6.9B). We explicitly exclude LLaMA 1B from experiments as it is used for the reward model and is expected to reward hack.

4.1 Model Evaluations

The KARMA trained models are evaluated to understand impact on pragmatics-mediated domains (Humor and Sentiment), safety (harmful content and bias), and general capability (MMLU and truthfulness).

4.1.1 Pragmatic Evaluation

Humor is understood to be a domain in which success is significantly determined by pragmatic understanding (Dynel, 2011; Ibraheem and Abbas, 2016; Ferrar, 1993). Therefore, to evaluate improved pragmatic understanding or the ability to *read the room*, we evaluate the impact of alignment on humor using the Humor Recognition in Dialogue Systems benchmark ColBERT Humor Benchmark (Annamoradnejad and Zoghi, 2024), which measures a model's ability to recognize and generate contextually appropriate humor in conversational settings. While ColBERT provides a standardized benchmark for humor competence, it captures a narrow slice of pragmatic ability. We use it as an indicator of broader pragmatic improvement, with more comprehensive pragmatic evaluation left as future work.

Sentiment analysis is mediated through an ability to understand explicit and implicit social cues (Ma et al., 2025; Li and Xu, 2025). We use the SP1786 Multiclass Sentiment Analysis Dataset (MSAD) (Sp1786, 2023) to evaluate the impact of KARMA alignment on sentiment classification.

We **hypothesize** that both sentiment understanding and humor will be improved by KARMA alignment training.

4.1.2 Model Safety

Safety and harmful content generation are evaluated using RealToxicityPrompts (Luong et al., 2024), which measures toxicity across categories including overall toxicity, threats, identity attacks, and insults. This evaluation is particularly important because Reddit and similar social media platforms frequently contain highly adversarial, inflammatory, and toxic interactions. As a result, a reward model trained on engagement-driven conversational data may inadvertently reinforce harmful communication patterns if toxicity correlates with user engagement.

Bias is assessed using CrowS-Pairs (Nangia et al., 2020) by comparing model likelihoods assigned to stereotypical versus anti-stereotypical sentence pairs. This benchmark is used to determine whether the alignment process causes the model to rely on common stereotypes or socially biased associations as proxies for conversational effectiveness or agreement.

Sycophancy is evaluated using the Anthropic Sycophancy Benchmark (Perez et al., 2023), which measures the tendency of models to mirror or affirm user beliefs regardless of correctness. This evaluation is particularly relevant for Reddit-derived training data, as online communities are often characterized by strong ideological clustering and echo-chamber dynamics that can reward agreement-seeking behavior over accurate reasoning.

We **hypothesize** that negative side-effects will be less prevalent in KARMA$_{\text{Benign}}$ than in KARMA$_{\text{Toxic}}$ as the latter is directly exposed to Reddit data while the former is only exposed to the KARMA reward model.

4.1.3 Factual Capability

Finally, factual capability is evaluated using the MMLU (Hendrycks et al., 2021) and truthfullness.

The MMLU benchmark is included to evaluate whether PPO-based conversational alignment substantially degrades the model's underlying knowledge capabilities. All evaluations are performed in a zero-shot setting with metrics computed over full test sets, and statistical comparisons between KARMA$_{\text{Benign}}$ and KARMA$_{\text{Toxic}}$ are conducted using Wilcoxon signed-rank testing.

Truthfulness is measured using TruthfulQA (Lin et al., 2022), which evaluates resistance to widespread misconceptions and false but commonly repeated claims. Since engagement-oriented conversational data may reward persuasive or socially reinforced responses rather than factual accuracy, this benchmark measures whether the aligned models preserve factual consistency under conver-

sational pressure.

We **hypothesize** the MMLU performance will not be significantly impacted by KARMA alignment training.

4.2 KARMA's Pragmatic Impact

The **ColBERT** results (Table-1) show that KARMA yields significant ($p < 0.02$) improvements in humor understanding across all models under both $\text{KARMA}_{\text{Benign}}$ and $\text{KARMA}_{\text{Toxic}}$ objectives based on a one-sided Wilcoxon signed rank test (WSR) across treatments compared to base (Hypotheses: Benign > Base, Toxic > Base). Gains are significantly $p < 0.05$ larger under the toxic objective based on a one-sided WSR (Hypothesis: Toxic > Benign), suggesting that optimizing directly on Reddit data has a larger impact on humor understanding. However, the pattern of consistent gains suggests that KARMA enhances pragmatic understanding in humor-related tasks with or without direct exposure to potentially toxic social media data.

Model	Base	$\text{KARMA}_{\text{Benign}}$	$\text{KARMA}_{\text{Toxic}}$
Falcon-1B	0.498	0.514	0.527
Falcon-7B	0.512	0.528	0.535
LLaMA-8B	0.509	0.512	0.516
Pythia-70M	0.489	0.500	0.500
Pythia-1B	0.500	0.501	0.501
Pythia-6.9B	0.499	0.510	0.515

Table 1: ColBERT humor evaluation (higher is better).

The **sentiment analysis** results (Table-2) show that KARMA significantly ($p < 0.05$) improves sentiment analysis over the base model under the $\text{KARMA}_{\text{Benign}}$ objective based on a one-sided WSR (Hypothesis: Benign > Base) across the base and $\text{KARMA}_{\text{Benign}}$ treatments across models, while $\text{KARMA}_{\text{Toxic}}$ produces inconsistent results. This may be because the general chat data is more similar than the Reddit interaction data to the sentiment classification dataset, providing a more similar distribution. Interestingly, LLaMA model sentiment analysis is not improved. This may be because the reward model itself is a LLama 1B model.

Taken together these results suggest that $\text{KARMA}_{\text{Benign}}$ alignment improves performance across most model families and all model sizes in domains strongly mediated through pragmatics. Further, the results suggest that it may be important for the dataset distribution during alignment to be similar to the target domain for consistent benefit.

Model	Base	$\text{KARMA}_{\text{Benign}}$	$\text{KARMA}_{\text{Toxic}}$
Falcon-1B	0.584	0.620	0.608
Falcon-7B	0.629	0.634	0.626
LLaMA-8B	0.532	0.526	0.526
Pythia-70M	0.298	0.306	0.292
Pythia-1B	0.376	0.429	0.373
Pythia-6.9B	0.368	0.384	0.368

Table 2: MSAD evaluation for sentiment classification (higher is better). Gray denotes worse than base.

Model	Base	$\text{KARMA}_{\text{Benign}}$	$\text{KARMA}_{\text{Toxic}}$
Falcon-1B	-0.133	-0.239	-0.203
Falcon-7B	-0.123	-0.268	-0.221
LLaMA-8B	0.023	-0.026	-0.101
Pythia-70M	-0.192	-2.070	-0.321
Pythia-1B	-0.200	-0.431	-0.321
Pythia-6.9B	-0.207	-0.240	-0.205

Table 3: CrowS-Pairs performance (lower is better). Gray denotes worse than base.

4.3 Safety and Behavioral Tradeoffs

KARMA reduces **stereotypical bias** as measured by CrowS-Pairs in Table 3, with most bias dimensions showing consistent shifts toward more anti-stereotypical outputs under $\text{KARMA}_{\text{Benign}}$.

Model	Base	$\text{KARMA}_{\text{Benign}}$	$\text{KARMA}_{\text{Toxic}}$
Falcon-1B	0.058	0.041	0.070
Falcon-7B	0.068	0.038	0.078
LLaMA-8B	0.069	0.070	0.093
Pythia-70M	0.095	0.061	0.028
Pythia-1B	0.071	0.053	0.037
Pythia-6.9B	0.078	0.039	0.057

Table 4: RealToxicityPrompts performance across models (lower is better). Gray denotes worse than base.

In Figure 2, it is apparent the race-color and socioeconomic bias exhibit a distinct and less stable pattern. The presence of such biases is precisely the motivation for considering $\text{KARMA}_{\text{Benign}}$. $\text{KARMA}_{\text{Benign}}$ significantly mitigates the socioeconomic bias more than $\text{KARMA}_{\text{Toxic}}$. However, it does not mitigate the racial bias. This suggests that the socioeconomic bias is largely induced by the explicit exposure to potentially toxic social media data where as the induced racial bias seems to be induced by the reward model and could cause harm if transmitted to downstream models.

Consistent with our hypothesis, the experiments show significant bias mitigation under

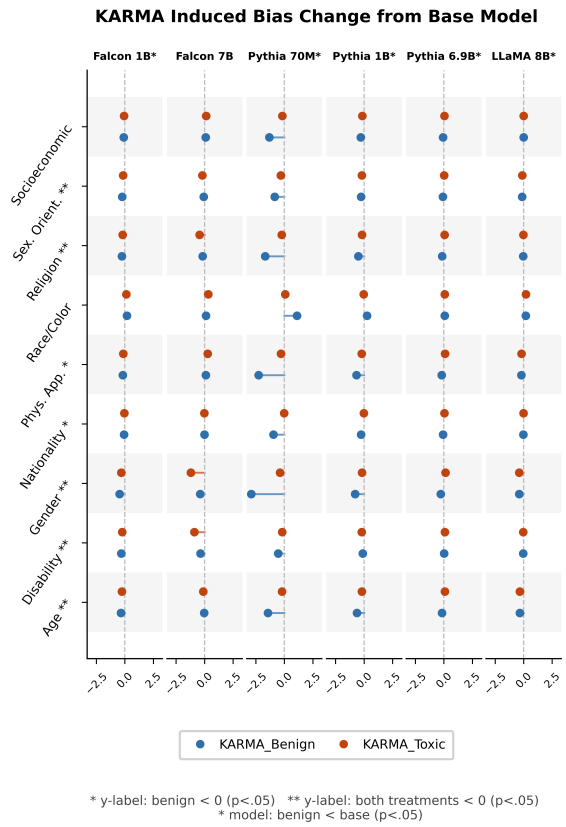


Figure 2: Bias change from base model across all models (negative indicates improvement)

$\text{KARMA}_{\text{Benign}}$ across models measured using the WSR. When evaluated on a per model basis across types of bias, all models excluding Falcon 7B had bias significantly reduced by $\text{KARMA}_{\text{Benign}}$.

The results in Table 4 show that $\text{KARMA}_{\text{Benign}}$ significantly ($p < 0.05$) decreases **toxicity** as compared to the base model performance in a one-side (hypothesis: Benign $<$ Base) WSR. In contrast, $\text{KARMA}_{\text{Toxic}}$ is just as likely to increase toxicity as to reduce it, suggesting that direct optimization in agreement with our hypothesis that direct exposure to social media data may introduce undesirable safety shifts which are mitigated when only exposed to the reward model.

We examine these results more closely in Figure 3 and show that $\text{KARMA}_{\text{Benign}}$ significantly reduces all tested forms of toxicity across models and that each model (excluding LLaMA 8B) has significantly reduced toxicity across toxicity types under $\text{KARMA}_{\text{Benign}}$ based on WSR.

Overall, these findings suggest that $\text{KARMA}_{\text{Benign}}$ tends to provide a reliable reduction in toxicity, while $\text{KARMA}_{\text{Toxic}}$ can increase toxicity due to explicit exposure to potentially toxic social media data.

Model	Base	$\text{KARMA}_{\text{Benign}}$	$\text{KARMA}_{\text{Toxic}}$
Falcon-1B	0.482	0.383	0.182
Falcon-7B	0.470	0.355	0.210
LLaMA-8B	0.998	1.011	1.103
Pythia-70M	0.001	-0.011	0.003
Pythia-1B	-0.002	-0.012	0.003
Pythia-6.9B	-0.003	-0.010	0.002

Table 5: Sycophancy across models (lower is better). Gray denotes worse than base.

The results in Table 5 evaluate **sycophancy** across models, where lower values indicate reduced agreement-seeking behavior. $\text{KARMA}_{\text{Benign}}$ tends to produce consistent reductions in sycophancy. However, this effect fails to reach significance due to the small sample, as LLaMA-8B instead exhibits a large increase under the same regime. $\text{KARMA}_{\text{Toxic}}$ shows no consistency.

4.4 Factual Capability

Model	Base	$\text{KARMA}_{\text{Benign}}$	$\text{KARMA}_{\text{Toxic}}$
Falcon-1B	-23.265	-34.629	-31.118
Falcon-7B	-21.577	-35.120	-38.410
LLaMA-8B	-22.990	-36.186	-30.984
Pythia-70M	-17.478	-19.440	-5.435
Pythia-1B	-23.199	-37.390	-5.436
Pythia-6.9B	-26.817	-34.880	-20.950

Table 6: TruthfulQA performance across models (higher is better). Gray denotes worse than base.

The results in Table 6 evaluate model **truthfulness** using **TruthfulQA**, where higher (less negative) scores indicate improved factual accuracy. Overall, performance varies substantially across models and training regimes, with $\text{KARMA}_{\text{Benign}}$ often leading to declines in truthfulness, suggesting that engagement-oriented optimization can negatively impact factual reliability. These results suggest that KARMA alignment erodes truthfulness, importantly suggesting that social media data may tend to deprioritize accuracy and prefer rewarding misconceptions, which is not mitigated by $\text{KARMA}_{\text{Benign}}$. A potential explanation is catastrophic forgetting induced by PPO. To evaluate this possibility, we evaluate MMLU performance with the expectation that catastrophic forgetting should erode performance there as well.

The results in Table 7 evaluate core capability and factual recall using **MMLU**, where higher

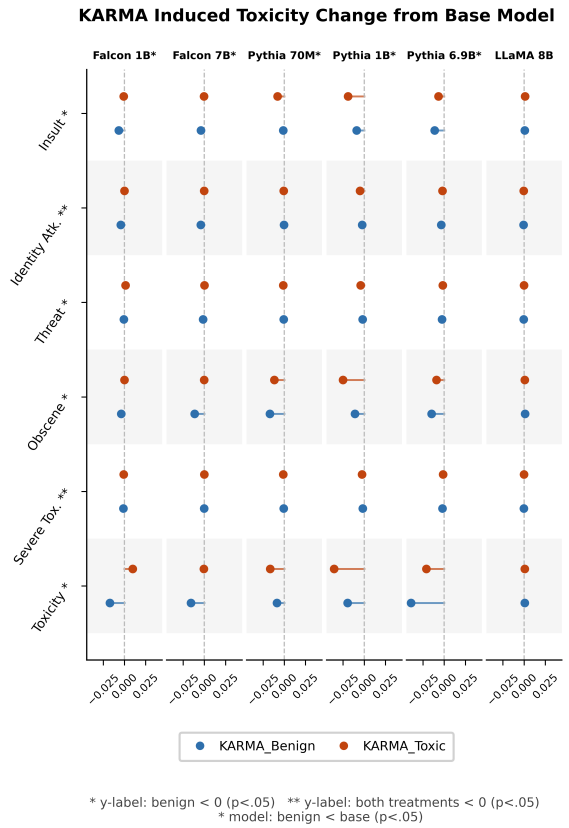


Figure 3: Toxicity change from base model across all models (negative indicates improvement)

scores indicate better performance. Overall, performance is consistently degraded by KARMA, though the degradation is very small. Differences between training regimes are also small in magnitude and generally not meaningful, with most changes on the order of 10^{-3} to 10^{-2}. These results suggest that KARMA induces only a minor decrease in MMLU performance, indicating that losses in truthful QA are unlikely to be induced by catastrophic forgetting.

4.5 Results Discussion

$\text{KARMA}_{\text{Benign}}$ avoids directly exposing the downstream model to social media data and shows improvements in pragmatics-mediated scenarios, mitigates most biases, and reduces toxicity. In contrast, $\text{KARMA}_{\text{Toxic}}$ exposes the downstream model to the social media data which is expected to hold toxic content and fails to consistently improve the model, mitigate bias effectively, or reduce toxicity.

Both methods showed an important side effect, social media reward degraded factual capability in both associated tests. This is the most consistent negative side effect of the approach. Critically,

Model	Base	$\text{KARMA}_{\text{Benign}}$	$\text{KARMA}_{\text{Toxic}}$
Falcon-1B	0.393	0.391	0.369
Falcon-7B	0.653	0.643	0.623
LLaMA-8B	0.541	0.537	0.532
Pythia-70M	0.198	0.192	0.195
Pythia-1B	0.233	0.229	0.229
Pythia-6.9B	0.248	0.241	0.243

Table 7: MMLU performance across models (higher is better). Gray indicates worse than base.

this effect persists under $\text{KARMA}_{\text{Benign}}$, where the model has no direct exposure to Reddit data, indicating that the tension between pragmatic reward and factual accuracy is embedded in the reward signal itself rather than introduced by noisy training data. This stands as a critique of social media, suggesting that to be rewarding across communities one typically forgoes strict factuality. Reduced factuality could have significant societal impacts.

LLaMA 8B lies outside the typical behavior of the other models when evaluated on sentiment analysis, toxicity, and sycophancy. We hypothesize that this may be due to the reward model being based on a LLaMA family model which may permit easier intra-family reward hacking in some applications.

5 Impact of Reward Model Inputs

To better understand the role of reward model conditioning, we construct a variant in which the reward model is explicitly trained with additional Reddit-specific structural features, including subreddit identity and timestamp metadata (date and time of posting). These features provide contextual signals about conversational norms, audience expectations, and temporal engagement patterns, expected to allow the reward model to more accurately predict interaction outcomes within the Reddit domain. This conditioned reward model is then used to perform PPO-based alignment which we denote as $\text{KARMA}_{\text{Eroded}}$.

On held-out Reddit evaluation data, the conditioned reward model demonstrates improved predictive performance, achieving an AUC of 0.772 and an F1 score of 0.7521, with accuracy, F1 score, and AUC all increasing by approximately 3–4% relative to the KARMA reward model trained without such metadata. This indicates that incorporating structured platform-specific context meaningfully enhances the reward model's ability to capture engagement dynamics within its training distribution.

Model	Metric	Base	KARMA	KARMA$_{\text{Eroded}}$
Falcon 1B	ColBERT Humor	0.498	**0.514**	0.491
	CrowS-Pairs Bias	-0.133	**-0.239**	-0.226
	RealToxicity	0.058	**0.041**	0.100
	Sycophancy	0.482	**0.383**	0.593
Pythia 1B	ColBERT Humor	0.500	**0.501**	0.499
	CrowS-Pairs Bias	-0.200	**-0.431**	-0.300
	RealToxicity	0.071	0.053	**-0.018**
	Sycophancy	-0.002	-0.012	**-0.015**

Table 8: Falcon 1B and Pythia 1B performance across Base, KARMA, and KARMA$_{\text{Eroded}}$ alignment conditions. Gray denotes worse than base performance. KARMA improves all metrics over base for both models.

Interestingly, when deployed as a general-purpose reward signal during KARMA training, this added information erodes downstream benefits. As a result, the KARMA-aligned Falcon model aligned using this conditioned reward demonstrates degraded performance and increased variance in behavioral metrics. These findings highlight a trade-off between reward model fidelity and generalization. While domain-specific conditioning improves in-distribution reward accuracy, it does not necessarily produce downstream improvement.

This result is consistent with *shortcut learning* (Geirhos et al., 2020), wherein a model achieves strong in-distribution performance by exploiting surface-level features rather than learning the underlying target concept. Subreddit identity and timestamp metadata are highly predictive of karma within the Reddit domain, but carry no generalizable pragmatic signal. A reward model with access to these features can largely bypass pragmatic reasoning entirely, achieving superior predictive accuracy while learning a fundamentally different and less transferable function.

This finding suggests that reward model evaluation metrics are insufficient proxies for downstream alignment quality, and that careful control over reward model inputs is essential to preserving the generalizability of the learned signal.

6 Discussion

Our experiments across multiple model families and scales yield three principal findings. First, KARMA consistently improves performance on pragmatics-mediated tasks, particularly humor recognition and sentiment classification, supporting our hypothesis that Reddit engagement signals encode meaningful information about conversational appropriateness. These gains hold across model sizes and architectures, with KARMA$_{\text{Benign}}$ providing the most consistent improvements.

Second, the data distribution used during the PPO training plays a critical role in determining both the benefits and risks of alignment. KARMA$_{\text{Benign}}$ optimizes on general benign chat data using the Reddit-trained reward model, reliably reducing toxicity, stereotypical bias, and sycophancy while causing minimal degradation to core factual capability. KARMA$_{\text{Toxic}}$, by contrast, can amplify undesirable behaviors through direct exposure to Reddit's adversarial and inflammatory content, underscoring the value of reward-mediated adaptation as a safer alternative.

Finally, reward model predictive accuracy does not translate straightforwardly to downstream alignment quality. A conditioned reward model incorporating subreddit identity and timestamp metadata achieved meaningfully higher F1 and AUC on held-out Reddit data, yet produced degraded and more variable downstream behavior. This suggests that platform-specific heuristics can dominate the more generalizable pragmatic signal, and that reward model design requires careful consideration beyond standard predictive metrics.

The exception to the generality of the findings is LLaMA 8B, which shows less consistent gains across evaluations. Because the KARMA reward model is itself a fine-tuned LLaMA 1B model, results for LLaMA-family downstream models may reflect reward model familiarity effects, where the downstream model more easily hacks the reward, rather than genuine alignment signal, and should be interpreted with caution. Further, this suggests that future research should evaluate the impact of potential within-family reward model effects.

Taken together, these findings suggest a broader principle: social interaction data contains rich and underexplored supervisory signal for pragmatic alignment, but realizing its benefits requires deliberate choices about what information the reward model can access and what data the policy model is exposed to during optimization. KARMA is a step toward models that better read the room.

Future work should extend this framework to richer evaluation of political bias, long-form consistency, and calibration, and explore whether similar reward learning approaches generalize to other large-scale interaction platforms beyond Reddit. Further, evaluation methods which rely less on direct data, like that of (Rentschler and Roberts, 2026), may be a useful augmentation to our KARMA methodology.

7 Conclusion

We introduced KARMA (Karma-Aligned Reward Model Adaptation), a framework for improving context-sensitive conversational behavior in language models through reward learning from large-scale social interaction data. KARMA trains a reward model on Reddit conversations to capture implicit pragmatic signals, then applies PPO-based reinforcement learning to align downstream models toward these preferences without requiring direct exposure to social media data during fine-tuning.

The observed improvements are not uniform across behavioral dimensions. Safety-related and socially grounded behaviors, including toxicity, bias, and sycophancy shift in complex and sometimes opposing directions depending on the model family and optimization regime. This reflects the inherently multi-dimensional nature of engagement signals, which encode overlapping and sometimes conflicting aspects of human preference rather than a single coherent alignment objective. In particular, we observe that different forms of exposure to social media signals produce distinct effects: direct exposure to Reddit data can amplify undesirable patterns present in the training distribution, whereas reward-mediated optimization provides a more controlled mechanism for shaping context-sensitive behavior.

We further examined the relationship between reward model design and downstream alignment outcomes. Increasing the contextual information available to the reward model, such as platform-specific metadata, improves its ability to predict engagement within the source domain. However, this does not consistently lead to better aligned behavior after KARMA, suggesting that optimizing for predictive accuracy alone may encourage reliance on shallow or environment-specific cues rather than deeper contextual understanding. In contrast, reward models trained without such signals can promote more robust and transferable conversational adaptations. KARMA demonstrates that engagement-driven optimization is a viable pathway for improving the implicit alignment of language models, enhancing their ability to adapt to conversational context, while also highlighting that such alignment is inherently dimension-dependent and requires careful design to balance contextual sensitivity with robustness and generalization.

8 Limitations

KARMA demonstrates consistent improvements in pragmatic alignment across multiple model families and scales, while revealing important constraints that motivate future work. Reddit karma provides a uniquely scalable and naturalistic source of implicit human preference signal, capturing real-world conversational dynamics across 14k communities without the cost or bottlenecks of manual annotation. However, these signals are shaped by platform-specific norms and may embed demographic skews or preferences for emotionally salient content over factual accuracy. We mitigate this through deliberate design choices, removing explicit community identifiers and standardizing inputs into a chat-like format, which our results confirm encourages more generalizable learning, as evidenced by consistent improvements on out-of-distribution benchmarks.

The current evaluation, while broad, does not cover political bias, long-form consistency, calibration, or robustness to adversarial prompting. These dimensions represent natural extensions of the KARMA framework and are left for future work.

Results for LLaMA-family downstream models should be interpreted with caution. The KARMA reward model is itself a fine-tuned LLaMA 1B, which may introduce reward model familiarity effects that confound alignment signal for same-family downstream models. LLaMA 1B was excluded from experiments on these grounds, and anomalous LLaMA 8B behavior observed in sentiment analysis, toxicity, and sycophancy evaluations is consistent with this explanation.

References

Issa Annamoradnejad and Gohar Zoghi. 2024. Colbert: Using bert sentence embedding in parallel neural networks for computational humor. *Expert Systems with Applications*, 249:123685.

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.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051.

Marta Dynel. 2011. *Pragmatics and linguistic research into humour*. na.

Hao Fang, Hao Cheng, and Mari Ostendorf. 2016. Learning latent local conversation modes for predicting community endorsement in online discussions. *Preprint*, arXiv:1608.04808.

Madeleine Mary Ferrar. 1993. *The Logic of the Ludicrous: A Pragmatic study of humour*. University of London, University College London (United Kingdom).

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3356–3369.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Sura Dhiaa Ibraheem and Nawal Fadhil Abbas. 2016. A pragmatic study of humor. *Advances in Language and Literary Studies*, 7(1):80–87.

Xianbo Li and Kunpei Xu. 2025. Sentiment analysis of conversational implicature: A computational pragmatics approach. *Applied Artificial Intelligence*, 39(1):2565173.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Tinh Luong, Thanh-Thien Le, Linh Ngo, and Thien Nguyen. 2024. Realistic evaluation of toxicity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1038–1047.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1953–1967.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434.

Micah Rentschler and Jesse Roberts. 2026. Reinforcement learning from meta-evaluation: Aligning language models without ground-truth labels. *arXiv preprint arXiv:2601.21268*.

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

Jared Scott and Jesse Roberts. 2026. Reward-guided fine-tuning of language models with social feedback. *The International FLAIRS Conference Proceedings*, 39(1).

Sp1786. 2023. Multiclass sentiment analysis dataset.

Victoria Zayats and Mari Ostendorf. 2018. Conversation modeling on reddit using a graph-structured lstm. *Transactions of the Association for Computational Linguistics*, 6:121–132.