Exploring Intrinsic Alignments within Text Corpus

Zi Liang[1,2], Pinghui Wang[2*], Ruofei Zhang[3], Haibo Hu[1*], Shuo Zhang[2], Qingqing Ye[1], Nuo Xu[2], Yaxin Xiao[1], Chen Zhang[4], Lizhen Cui[5]

[1]The Hong Kong Polytechnic University
[2]MOE KLINNS Lab, Xi'an Jiaotong University
[3]Apple
[4]Zhejiang Createlink Technology
[5]Shandong University

zi1415926.liang@connect.polyu.hk, phwang@mail.xjtu.edu.cn, rfzhang@gmail.com, haibo.hu@polyu.edu.hk
zs412082986@stu.xjtu.edu.cn, qqing.ye@polyu.edu.hk, nxu@sei.xjtu.edu.cn
20034165r@connect.polyu.hk, zhangchen@chuanglintech.com, clz@sdu.edu.cn

Abstract

Recent years have witnessed rapid advancements in the safety alignments of large language models (LLMs). Methods such as supervised instruction fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) have thus emerged as vital components in constructing LLMs. While these methods achieve robust and fine-grained alignment to human values, their practical application is still hindered by high annotation costs and incomplete human alignments. Besides, the intrinsic human values within training corpora have not been fully exploited. To address these issues, we propose ISAAC (Intrinsically Supervised Alignments by Assessing Corpus), a primary and coarse-grained safety alignment strategy for LLMs. ISAAC only relies on a prior assumption about the text corpus, and does not require preferences in RLHF or human responses selection in SFT. Specifically, it assumes a long-tail distribution of text corpus and employs a specialized sampling strategy to automatically sample high-quality responses. Theoretically, we prove that this strategy can improve the safety of LLMs under our assumptions. Empirically, our evaluations on mainstream LLMs show that ISAAC achieves a safety score comparable to current SFT solutions. Moreover, we conduct experiments on ISAAC for some RLHF-based LLMs, where we find that ISAAC can even improve the safety of these models under specific safety domains. These findings demonstrate that ISAAC can provide preliminary alignment to LLMs, thereby reducing the construction costs of existing human-feedback-based methods.

1 Introduction

Large language model (LLM) enhanced dialogue systems, such as ChatGPT, have garnered significant attention thanks to their wide range of real-world applications (Anil et al. 2024; Achiam et al. 2023; Zhang et al. 2023) in chit-chat, information-seeking, and task-oriented business. However, they still face the problem of *unsafe response generation* (Anwar et al. 2024), which involves generating offensive, politically sensitive, unprofessional, or biased sentences, particularly under *adversarial prompts* from ill-intentioned users. For instance, chatbots like XiaoIce, Twitter bot Tay (Wolf, Miller, and Grodzinsky 2017), Blenderbot 3.0, and ChatGPT have been reported to produce offensive and racist responses after their release (Zhu, Wang, and Liu 2024). In the context of task-oriented dialogues (TOD), some research also focuses on the politeness transfer (Silva, Semedo, and Magalhães 2022) within real-world corpora.

Existing research in this field, such as supervised training and reinforcement learning (RL), detoxifies dialogue models by the feedback of human annotators. For supervised training, researchers either train safety classifiers based on annotated safety corpora (Sun et al. 2021; Dinan et al. 2019; Roller et al. 2020; Baheti et al. 2021; van Aken et al. 2018) to filter unsafe dialogues, or they annotate training samples with human-crafted responses (Ung, Xu, and Boureau 2022; Bianchi et al. 2024; Zhong et al. 2024). Based on the detoxified dataset, they then fine-tune models (Bianchi et al. 2024) via SFT. Similarly, *reinforcement learning with human feedback* (RLHF) (Ouyang, Wu, and et al. 2022; Bai et al. 2022a), involves collecting a dataset of LLM-interactive dialogues, rating each response in the dialogues, and training a reward model to guide LLMs training with RL algorithms (*e.g.* PPO (Schulman et al. 2017)) or directly training the LLMs (*e.g.* DPO (Rafailov et al. 2023)).

Though effective, these methods encounter two challenges. First, they heavily rely on human annotations for alignment. Second, these methods are often vulnerable to specific topics if the human-collected subset does not represent the entire distribution of the training corpus. Consequently, maintainers of LLMs need to recollect new subsets for emerging safety topics or scenarios due to the *adversarial evolution* phenomenon (Shachaf and Hara 2010; Dinan et al. 2019) in interactions. These challenges highlight the need for a more flexible and evolvable alignment strategy.

To address the above issues, we explore whether there are potential features (induced bias) in text corpus that naturally represent human values. In other words, we aim to *align LLMs with human values inherent in the text corpus itself, rather than injecting human values through additional annotations*. This task is challenging, as it requires discriminating the safety of a training sample without label information.

*Corresponding authors.

To this end, we propose *ISAAC*[1], a coarse-grained alignment strategy which can automatically cover the content safety of a given corpus without human feedback. ISAAC is **not** intended to replace existing human-feedback-based methods but to complement them. Inspired by the discovery that unsafe responses are typically few and semantically different in the corpus (Tufa, Markov, and Vossen 2024; Founta et al. 2018), ISAAC replaces unsafe responses with their predominantly ordinary neighbors. Specifically, it first clusters examples with similar contextual information to obtain the response distribution for given contexts. This distribution, usually long-tailed, is then used to sample multiple pseudo responses based on our proposed adaptively sharpening sampling strategy. We use these pseudo-response labels to fine-tune LLMs for preliminary alignments or to train a rephrasing model that transforms an unsafe response under a given context.

By evaluating ISAAC on several popular language models, experiments demonstrate that it produces more diverse (0.03 in DIST-2 and 0.62 in Entropy) and contextually aware (2% in perplexity) responses while maintaining safety scores comparable to existing supervised training methods. Compared to comprehensive alignment, ISAAC improves the safety score of Claude, Llama3, Qwen2 by 1.4%, 9.5%, and 9.9%, respectively.

2 Related Works

Conversational AIs & LLMs. Training general-purpose language models typically involves two steps, the pre-training stage, where the model learns general knowledge, and the fine-tuning stage, where it learns to understand user intents and to follow instructions. During fine-tuning, dialogue contexts (instructions and task inputs), as well as corresponding responses, are arranged as the text corpus for maximum likelihood estimation training. These fine-tuning datasets often combine commonly used and universal tasks to construct a multi-task dialogue service. For example, BlenderBot (Roller et al. 2020), a dialogue model trained on diverse corpora, has achieved significant improvements with increased model parameters and training data. This trend is further exemplified by LaMDA (Thoppilan et al. 2022), which not only expands its training datasets but also encompasses over 100 billion parameters. Through carefully and comprehensive human evaluation, LaMDA demonstrates that it is possible for current transformer-based LLMs to achieve real and multi-turn human-like responses. Subsequent advancements in conversational AI stem from the use of human-feedback-based reinforcement learning and instruction tuning. Models such as InstructGPT (Ouyang, Wu, and et al. 2022), Sparrow (Glaese, McAleese, and et al. 2022), and ChatGPT are fine-tuned with a well-annotated subset after the generalized pre-training. High-quality instruction-tuning corpus have thus been proven (Gunasekar et al. 2023) to significantly influence the training of LLMs. Besides, these methods also benefit from evolutionary training of RLHF. By collecting and annotating their responses to users, LLMs can evolve continuously. High-quality SFT and RLHF fully exploits the potential of LLMs, establishing them as the de facto state-of-the-art solution. Moreover, incorporating external knowledge (Xu et al. 2020) like retrieval-augmented generation (RAG) (Gao et al. 2023; Chen et al. 2024a) can further enhance LLMs' performance.

Safety of Contents & Value Alignments. As a substantial field of AI security (Zhou et al. 2024; Bai et al. 2024; Liang et al. 2024a; Li et al. 2023; Xiao et al. 2022; Liang et al. 2024b; Wang et al. 2024; Zheng et al. 2022), aligning language models with human values begins with generating safe responses on dialogue models. Dinan et al. (2019) explore supervised safe response generation early on. In their experiments, they inform annotators to produce dangerous utterances to "trigger" the models into generating unsafe responses. These dialogues were then collected to create a training dataset to improve the safety of dialogue models. Subsequently, an adversarial human-machine interaction corpus (Roller et al. 2020) were proposed, on which they remove unsafe samples and fine-tune dialogue models only with repaired safe responses, known as "baked-in" training. Other research (Liu et al. 2020) also proposed similar supervised training paradigms, such as GAN-based training framework, to enhance the fairness of dialogue models in sensitive categories like gender and racism. Baheti et al. (Baheti et al. 2021) address the issue of incorrect stance in responses, proposing a controlled-text-generation-based solution. These strategies have been inherited by supervised instruction fine-tuning (SFT), and numerous recent studies (Bianchi et al. 2024; Zhong et al. 2024) construct and fine-tune LLMs by a carefully dialogue pair selection.

In contrast to supervised training, human feedback can also be used to train reward models, which provides reinforcement learning signals to LLMs. Specifically, these methods (Ouyang, Wu, and et al. 2022; Glaese, McAleese, and et al. 2022) guide annotators to check the pre-defined safety rules or follow generalized human values, aligning LLMs to these values via reinforcement learning algorithms with binary feedback. Besides, recent studies (Rafailov et al. 2023) attempt to integrate the training procedure of reward models and LLMs to build a direct preference optimization (Rafailov et al. 2023), where the training target of LLMs is to maximum the difference in likelihoods between positive and negative user-annotated responses. Unlike to supervised fine-tuning, RL-based approaches allow LLMs to learn *preferences* among different kinds of responses even without "standard answers", enabling more flexible and powerful model alignments. Inspired by them, some works also explore the potential of self-criticism in LLMs, aiming to enable self-improvement without human annotations, by thecniques like reinforcement learning with AI feedback (RLAIF) (Bai et al. 2022b), self-correction (Welleck et al. 2023), or self-play (Chen et al. 2024b). However, the self-improvement ability of these methods has recently been questioned (Huang et al. 2023; Shumailov et al. 2024).

In summary, current studies focus on using external human annotations or exploring the potential within LLMs. However, the intrinsic distribution information of text corpora as well as its usage have not been thoroughly explored.

[1]**I**ntrinsically **S**upervised **A**lignments by **A**ssessing **C**orpus.

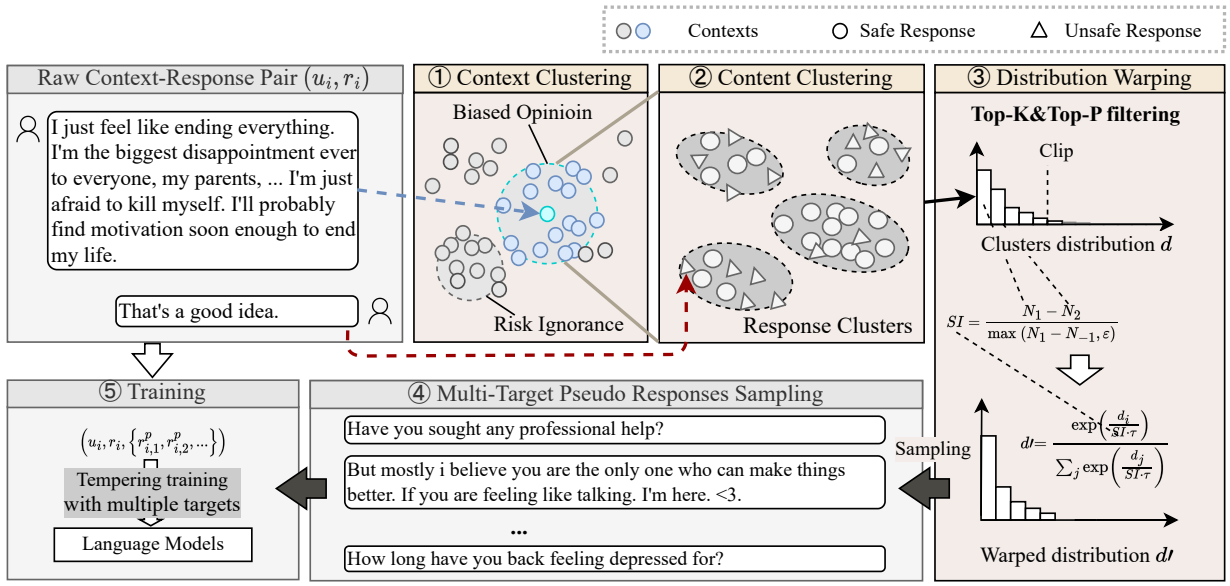

Figure 1: The sampling procedure of ISAAC, which aims to produce multiple potentially safer responses (bottom left) for each context-response pair (upper left). This is achieved through a two-step clustering (upper center) and adaptive sharpening methods (right).

3 ISAAC: Aligning with Text Corpus

In this section, we detail the method of extracting intrinsic alignment information contained in a text corpus and aligning language models accordingly. We first define the problem in Section 3.1, then discuss the deduction of intrinsic human values in Section 3.2, and finally describe the training procedure in Section 3.3.

3.1 Problem Definition

Given a text corpus $\mathcal{D}_{tr} = \{(u_i, r_i)\}_i$, where each pair (u_i, r_i) consists of the user context u_i and its corresponding response r_i, the model P_θ may learn to generate a unsafe response r_i' because r_i in $\mathcal{D}_{tr}$ is not fully aligned with human values. Therefore, value alignment usually refers to the transformation of LLMs from $p(r_i'|u_i, \theta_p)$ to $p(\hat{r}_i^s|u_i, \theta_p)$, in which $\hat{r}_i^s$, the response that well meets human values, should also be *context-related* to the inputs u_i.

Current methods either fine-tune the model parameters θ_p using human-crafted safe responses r_i^s, *i.e.*, (u_i, r_i', r_i^s) in SFT, or with a binary feedback $y_i \in \{0, 1\}$, to train reward models, i.e. (u_i, r_i, y_i) in RLHF. Our method, however, handles this problem with the raw pairs (u_i, r_i) without knowing whether r_i is safe or not. Complementarily, ISAAC relies on a basic and intuitive assumption about $\mathcal{D}_{tr}$: the number of safe responses is larger than that of unsafe responses for a given task. Section 3.2 details how to strictly define this assumption and amplify the influences of safe responses within the corpus.

3.2 Sampling Potentially Safer Responses

We first propose a primary pseudo-label sampling strategy to obtain a potentially safer response for a given dialogue context u_i.

Given a dataset $\mathcal{D}_{tr}$, we first cluster dialogues $\{(u_i, r_i)\}_i$ based on their context information to gather responses with similar topics. Specifically, for a given dialogue turn (u_i, r_i), we mine its neighbors $\mathcal{R}_c = \{(u_j, r_j)\}_j$ that are semantically similar to the *text representation* of the dialogue context u_i. We use a pre-trained NLU model (*e.g.*, RoBERTa (Liu et al. 2019)) to obtain these representations. Meta-information, such as *intent-slot combinations* in TODs, can also be included to build $\mathcal{R}_c$. Once we obtain $\mathcal{R}_c$, a second clustering step is introduced to separate it based on the semantic representations of responses. This process results in a cluster set $\mathcal{C} = \{R_1, R_2, ..., R_{N_C}\}$, where each cluster R_k represents a collection of responses that exhibit high semantic similarity. The goal of response sampling is to select a specific cluster R_k from $\mathcal{C}$, thus obtaining a potential safer response $r_i^p \in R_k$ for each given (u_i, r_i).

We conceptualize this problem as the task of distribution-based sampling. Incorporating the distribution $\mathbf{d}_c = \{|R_k|/|\mathcal{R}_c|\}_{k=1,2,...,N_C}$ of R_k, we aim to develop a sampling function $f_s(\mathbf{d}_c) \rightarrow \hat{\mathbf{d}}_c$ that accurately determines the probability of each cluster being selected. Our target is to ensure that the probability of sampling "potentially safe samples" increases after employing f_s. Unfortunately, *random sampling* may fall short in addressing this problem, as shown in Theorem 1:

Theorem 1 (Invariance of Random Sampling)**.** *Suppose the unknown probability of unsafe responses occurring in response clusters $\mathcal{C}$ is P_c', the probability of unsafe responses occurring in the sampled response $\hat{r}_i$ is $\hat{P}_c'$. Let $f_r(\mathbf{d}_c)$ denote random sampling and p_s denote the probability of sampling safe responses. Then the probability of sampling unsafe responses will not change,* i.e. *$\hat{P}_c' \equiv P_c'$ for random sampling.*

Algorithm 1: Response Sampling of ISAAC.

Input: Training dataset $\mathcal{D}_{tr}$, sampling temperature τ, filtering parameter k and p, tiny number ϵ, pseudo label number M, pre-trained language model f_p, and representation model f_r.
Output: model f'_p.

Initialization: $\mathcal{D}_p \leftarrow \emptyset, \mathcal{C} \leftarrow \emptyset$;
$\mathbf{u}$, $\mathbf{r}$=unzip($\mathcal{D}_{tr}$);
$\mathbf{E}_u$=$f_r(\mathbf{u})$;
$\mathbf{E}_r$=$f_r(\mathbf{r})$;
`// pseudo label sampling`
for $(u, r) \in \mathcal{D}_{tr}$ **do**
  $E_c \leftarrow f_r(u)$;
  `// context clustering`
  $\mathcal{R}_c \leftarrow$ TopKUtterance(E_c,$\mathbf{E}_u$,$\mathbf{u}$,$\mathbf{r}$);
  `// obtain the representation of selected responses`
  $\mathbf{E}^r_{\mathcal{R}_c} \leftarrow$ Subset($\mathbf{E}_r, \mathcal{R}_c$);
  `// content clustering`
  $\mathcal{C} \leftarrow$ DBScan($\mathbf{E}^r_{\mathcal{R}_c}$);
  $\mathbf{d}_c \leftarrow$ getClustersDistribution($\mathcal{C}$);
  `// Top-K & Top-p filtering`
  $\mathbf{d}_c \leftarrow$ clipTopKNum($\mathbf{d}_c, k$);
  $\mathbf{d}_c \leftarrow$ clipTopPProbability($\mathbf{d}_c, p$);
  `// calculate the sensitive coefficient`
  $\mathbf{s} = \frac{\mathbf{d}_c[0]-\mathbf{d}_c[1]}{max(\mathbf{d}_c[0]-\mathbf{d}_c[-1],\epsilon)}$;
  $\hat{\mathbf{d}}_c \leftarrow \emptyset$;
  `// warp the distribution`
  for $d \in \mathbf{d}_c$ **do**
    $\hat{d} \leftarrow \frac{\exp(\frac{d}{s\cdot\tau})}{\sum_j \exp(\frac{\mathbf{d}_c[j]}{s\cdot\tau})}$;
    $\hat{\mathbf{d}}_c \leftarrow \hat{\mathbf{d}}_c \cup \{\hat{d}\}$;
  end for
  $\mathbf{r}^p_c \leftarrow$ sampleFromDistribution($\hat{\mathbf{d}}_c, \mathbf{r}, M$);
  $\mathcal{D}_p \leftarrow \mathcal{D}_p \cup \{(u, r, \mathbf{r}^p_c)\}$;
end for
$f'_p \leftarrow$ train($\mathcal{D}_p, f_p$);
return f'_p;

Therefore, we introduce a definition and a theorem to propose an expected sampling strategy named *convex sampling*.

Definition 1 ($\mathcal{A} \succ \mathcal{B}$)**.** *For number series $\mathcal{A}$ and $\mathcal{B}$ with their sorted permutation series $\hat{\mathcal{A}}$ and $\hat{\mathcal{B}}$, let N_A and N_B denote their series lengths. We define $\mathcal{A} \succ \mathcal{B}$ if and only if*

1. *$N_A \geq N_B$, and*
2. *for any i-th element $b_i = \hat{\mathcal{B}}[i]$ there always exists $a_i = \hat{\mathcal{A}}[i]$ such that $b_i < a_i$.*

Theorem 2 (Convex Sampling)**.** *$\forall R_{\mathcal{C}} \in \mathcal{D}_{tr}$, let $\mathcal{C}' \subset \mathcal{C}$ denote the subset of unsafe response clusters in $\mathcal{C}$, and let $\mathcal{C} \backslash \mathcal{C}'$, its complementary set, denotes the subset of those safe clusters. Let $\mathbf{d}'_c$ and $\mathbf{d}_c \setminus \mathbf{d}'_c$ denote the cluster distributions of these two subsets. Given the* ***assumption*** *$\mathbf{d}_c \backslash \mathbf{d}'_c \succ \mathbf{d}'_c$ and the* ***condition*** *$\frac{\partial^2 f_s}{\partial \mathbf{d}_c^2} > \mathbf{0}$, we can derive the* ***conclusion*** *that $\hat{P}'_c < P'_c$, i.e. the sampled response has a higher probability of being safe than before.*

In essence, Theorem 2 suggests that a *convex sampling function* that fulfills $f''_s > \mathbf{0}$ can improve the probability of obtaining a potentially safer response if safe responses occupy the majority of the corpus, i.e. $\mathbf{d}_c \setminus \mathbf{d}'_c \succ \mathbf{d}'_c$. As discussed in Section 1, this assumption assumes the unsafe dialogues are distributed in the tail and are semantically distinct. This premise might not be fulfilled in some special situations, which we discuss in Section 4.

Based on Theorem 2, ISAAC employs two normalized functions to transform the distribution $\mathbf{d}_c$ into a sampling distribution $\hat{\mathbf{d}}_c$ to enhance the distinction between potential safe and unsafe response clusters. These functions include a temperature-based *softmax* and the *max* function.

By sampling a pseudo response label $r^p_{c,i} \sim \mathcal{R}_{\mathcal{C}}$ under $\hat{\mathbf{d}}_c$ as the expected response, the standard supervised training task can be formatted as:

$$\mathcal{L}_p = \sum_i \log p_\theta(r^p_{c,i}|u_i) = \sum_i \sum_{t=1}^{N} \log p_\theta(x_{i,t}|u_i, x_{i,<t}),$$

where $x_{i,t}$ denotes the t-th token of $r^p_{c,i}$, and N is the sequence length of training samples.

3.3 Supervised Fine-tuning for ISAAC

Based on the response sampling strategy described in Section 3.2, we now explore the training procedure of *ISAAC*.

Adaptively Sharpening. While Theorem 2 guides us in sampling safer dialogue responses, we are yet to address the differences among response clusters $\mathcal{C}$. Intuitively, if the head cluster contains most of the samples in $\mathcal{R}_{\mathcal{C}}$, we can relax the sharpening to provide more diversity for responses, as there is already sufficient sampling probability for the head cluster. In contrast, if the distribution does not show significant differences among head clusters, we may need to sharpen this distribution more intensively, to ensure the safety of responses. To this end, we propose a sensitive indicator SI:

$$SI = \frac{N_1 - N_2}{\max(N_1 - N_{-1}, \epsilon)}, \tag{1}$$

where N_1, N_2 and N_{-1} denote the size of top-2 and last clusters, and $\epsilon = 10^{-3}$ is a small number. Then, we can modify the original softmax sampling to the following formation with SI and the temperature τ:

$$f'_{exp}(\mathbf{d}_c) = \text{softmax}(\frac{\mathbf{d}_c}{SI \cdot \tau}) = \frac{\exp(\frac{N_i}{SI\cdot\tau})}{\sum_j \exp(\frac{N_j}{SI\cdot\tau})}. \tag{2}$$

Training Procedure. Inspired by curriculum learning (Soviany et al. 2022; Wang, Chen, and Zhu 2022), we divide the fine-tuning procedure into a series of sub-training stages. At each substage, we sample different responses for the same context to ensure the models learn generalized ability rather than memorizing specific sampled pseudo responses. We call this strategy as *tempering training*. Our other discovery in the fine-tuning process is that the language models can be trained and converge stably even with complex and conflicting texts. Based on this, we propose *multi-target training* for our alignment, which aims to force dialogue models to fit multiple different responses rather than a single one, encouraging the model to generate a generalized, non-specific

Method	Entropy↑	B-PPL↓	Safety ↑
Test Set	8.649	82.26	0
+random	7.051	84.03	52.30
+Detoxify	7.681	**82.85**	45.71
+PersAPI	7.616	82.88	54.09
+BBF	7.416	82.99	55.49
+BAD	7.471	82.97	51.90
+ISAAC-S eps=0.22	4.577	84.71	**78.24**
+ISAAC-S eps=0.42	**8.103**	83.25	29.74
+ISAAC-S w.o. AS	5.384	84.04	50.30
+ISAAC-M	5.124	85.00	61.88
+ISAAC-H	5.953	84.85	43.71
+GPT-4	9.75	77.37	77.64
+GPT-4+ISAAC	9.20	80.50	84.83
+Claude	**10.61**	75.80	74.25
+Claude+ISAAC	9.18	81.07	75.64
+Llama3(8B)	8.34	79.21	48.50
+Llama3(8B)+ISAAC	7.64	84.07	58.08
+Qwen2(7B)	9.98	**72.58**	78.64
+Qwen2(7B)+ISAAC	9.06	81.45	88.62
+Phi-3(3.8B)	8.56	80.90	**93.61**
+Phi-3(3.8B)+ISAAC	8.41	83.78	93.61

Table 1: Alignment comparison between ISAAC and the alignments of LLMs.

response. Formally, the original training task can be modified as:

$$\mathcal{L}_{mp} = \frac{1}{M}\sum_{l}\mathcal{L}_{p,l} = \frac{1}{M}\sum_{i}\sum_{l}^{M}\log p_\theta(r^p_{c,i,l}|u_i, r_i), \quad (3)$$

where $r^p_{c,i,l}$ denotes the l-th target pseudo response of r_i and M denotes the number of targets.

4 Experiments

4.1 Experimental Settings

Datasets. We use DiaSafety (Sun et al. 2021), a comprehensive dialogue safety dataset, as our evaluation benchmark. It consists of 11K safety-relevant instruction-response pairs across 7 unsafe categories sampled from Reddit, making it both comprehensive and challenging for evaluation. Besides, we construct a poisoned version of MultiWoz 2.1 (Eric et al. 2020) to quantify the *information missing problem* in our coarse-grained alignments. MultiWoz 2.1 is a task-oriented benchmark that contains structured dialogue data. These structured data, referred to as dialogue *actions* and *slots*, are well-suited for evaluating whether information is missing in the alignments.

Baselines. We compare ISAAC to both current supervised detoxification methods and value-aligned LLMs. For the former, our baselines include the Detoxify API [2], Perspective API (Lees et al. 2022), BBF (Dinan et al. 2019), and BAD (Roller et al. 2020). For the latter, we compare ISAAC with models such as GPT-3.5, Claude, GPT-4 (Achiam et al. 2023), Qwen-2 (Yang et al. 2024), and Llama-3 (Grattafiori, Dubey, and et al. 2024).

Evaluation Metrics. We evaluate ISAAC across three dimensions, *safety*, *text quality*, and *informativeness*.

For unseen responses generated by LLMs with different alignments, we have to evaluate their safety either by human evaluation or some approximated methods. To address the prohibitively high cost of manually annotating all generated responses, we first train a safety classifier based on the DiaSafety dataset. Specifically, we fine-tune a RoBERTa model, achieving an accuracy of 80.88%. The error of this classifier can be estimated by $|(P \cdot N_u + (1 - P_f) \cdot \frac{N-N_u}{N} - \frac{N_u}{N}|$, where P is the precision of the classifier, $P_f = \frac{\text{False Positive}}{\text{True Negative+False Positive}}$, and N and N_u denote the total and unsafe number of samples in the test set, respectively. In DiaSafety, with $N = 1095$ and $N_u = 501$, assuming $P = P_f = 0.8$, we calculate the error to be 1.7%. This suggests that while a standalone classifier evaluation is indeed useful, it may not be sufficiently accurate for detailed analysis. Therefore, we introduce human evaluation to assess the safety of responses flagged as "unsafe" by the classifier, using 5 annotators. We find this approach achieves a favorable trade-off between the evaluation accuracy and cost.

In terms of response quality, we use Entropy and DIST-N as metrics of diversity, and employ the training checkpoints of COLA (Warstadt, Singh, and Bowman 2019) and SST (Socher et al. 2013) to evaluate the *acceptance* and the *engagingness* of responses, respectively.

For informativeness, inspired by BARTScore (Yuan, Neubig, and Liu 2021), we propose two averaged logarithmic perplexity indicators to estimate the correlation between contexts and responses. Specifically, we introduce the Forward-Perplexity (F-PPL) metric, where the model is trained with contexts as inputs and original responses as outputs. Conversely, for Backward-Perplexity (B-PPL), the inputs are responses and the outputs are contexts. Besides, we adopt the *success rate*, a commonly used metric for evaluating the informativeness of task-oriented dialogue models, and BLEU-4 (Papineni et al. 2002) to estimate the quality of generated sentences, following previous research (Peng et al. 2020). We also propose two simple metrics to measure the safety of task-oriented dialogue models: unsafe response rate at the turn level (RPR) and the task level level (DPR), based on our poisoned corpus, respectively.

4.2 Implementation Details

For the rephrasing version of ISAAC, we use T5 (Raffel et al. 2020) as the backbone model, trained with a learning rate of $3e-5$ and a batch size of 4. The model is trained for 500 steps in information-missing experiments and 50000 steps in LLM experiments, following the implementation of Huggingface Transformers (Wolf, Debut, and Sanh 2020). We use DB-SCAN as the clustering algorithm used in ISAAC, setting the nearest neighbor number to 150 and epsilon to 0.22. All experiments are conducted on $8\times$ Nvidia Tesla V100 GPUs with 32GB of memory each. For LLMs and safety detection APIs, we use the default hyperparameters.

[2] https://github.com/unitaryai/detoxify

Model	Safety↑	Quality					Informativeness	
		Accep. ↑	Engage. ↑	AvgLen ↑	DIST2 ↑	Entropy ↑	F-PPL ↓	B-PPL ↓
Raw Test Set	54.25	85.41	42.35	14.64	0.63	9.14	45.78	79.51
+Detoxify	75.16	89.55	27.62	13.35	0.53	8.30	43.89	79.98
+PersAPI	76.80	90.64	26.31	13.20	0.51	8.15	43.26	80.05
+BBF	**79.63**	**90.83**	25.53	13.21	0.51	8.08	43.39	80.09
+BAD	77.99	90.58	26.78	13.12	0.52	8.17	43.14	80.04
+Detoxify+ISAAC (ours)	73.70	88.74	46.67	14.06	**0.55**	**8.79**	40.75	**79.76**
+BBF+ISAAC (ours)	77.17	89.68	48.75	**14.12**	0.52	8.68	**39.68**	79.81
+BAD+ISAAC (ours)	76.16	90.61	**49.90**	14.06	0.53	8.69	39.85	79.80
Blenderbot	54.79	89.07	65.74	2.98	0.54	7.22	18.85	80.73
+Detoxify	63.47	89.55	45.29	4.26	0.27	4.67	23.75	81.03
+PerAPI	63.93	**92.56**	41.95	**4.49**	0.24	4.39	24.54	81.09
+BBF	**64.02**	92.51	41.10	4.48	0.24	4.29	24.57	81.10
+BAD	63.65	91.78	43.57	4.32	0.26	4.51	24.00	81.06
+Detoxify+ISAAC (ours)	59.27	91.18	**68.36**	3.19	**0.43**	**6.86**	17.42	**80.82**
+BBF+ISAAC (ours)	57.90	91.51	68.12	3.26	0.41	6.79	**17.13**	80.84
+BAD+ISAAC (ours)	59.91	91.34	68.11	3.21	0.42	6.79	17.32	80.83
DialoGPT	73.51	95.67	34.74	9.49	0.26	6.89	18.41	80.48
+Detoxify	73.42	89.55	26.22	8.84	0.23	6.18	23.28	80.81
+PerseAPI	73.79	95.84	25.08	8.76	0.22	6.03	23.76	80.85
+BBF	73.42	95.90	23.46	8.66	0.21	5.85	24.61	80.91
+BAD	73.70	95.89	24.50	8.68	0.22	5.98	23.82	80.86
+Detoxify+ISAAC (ours)	75.80	95.47	40.51	9.53	**0.25**	**6.99**	**19.53**	**80.62**
+BBF+ISAAC (ours)	**76.89**	95.42	41.23	**9.58**	0.24	6.95	19.81	80.66
+BAD+ISAAC (ours)	75.98	**96.12**	**45.75**	9.53	0.25	6.96	19.57	80.64

Table 2: Experiments integrating ISAAC with safety APIs.

4.3 Safety Evaluation

We first compare the safety of ISAAC with existing LLMs, related safety methods, and APIs. To integrate ISAAC with existing closed-source LLMs, we train a rephrasing model following the implementation detailed in Section 4.2.

As shown in Table 1, we collect all unsafe samples in the DiaSafety test set and compute the safety score of ISAAC models. We categorize the unsafe response examples in the training set into three degrees, Simple (-S), Medium (-M), and Hard (-H) with fractions of 0.04, 0.1, and 0.3, respectively. These categories represent a higher fraction of unsafe examples compared to those in commonly used pretraining and supervised fine-tuning corpora.

From Table 1 we see that ISAAC enhances safety when trained on the raw corpus, with *adaptively sharpening* (AS) playing a crucial role in this alignment. Moreover, Table 1 demonstrates that safety can be balanced with response diversity by adjusting the clustering threshold ϵ. After integrating ISAAC with existing LLMs, we observe a surprising and significant improvement across all models. Specifically, with only a minimal reduction in diversity and informativeness, ISAAC leads to meaningful enhancements, particularly in Qwen2 and LLaMa-3.

Besides, we also incorporate ISAAC with some popular safety detection methods to evaluate whether ISAAC leads to statistically obvious drops in text quality compared to supervised fine-tuning. Table 2 presents the results on both the test set and various basic models. The improvements in DIST and Entropy indicate that ISAAC mitigates the issue of *trivial response generation* commonly found in template replacements used by safety APIs. Furthermore, the decrease in both F-PPL and B-PPL suggests that ISAAC enhances the correlation between responses and contexts. In addition, improvements in regular sentence metrics such as Acceptability and Engagingness, further coincides with this analysis.

4.4 Information-missing Experiments

We then explore the *information-missing problem* that arises from our proposed safety alignment strategies. To clearly compare the information-missing evaluation, we use task-oriented dialogue (TOD) as the primary evaluation task, since it includes several information-relevant metrics (*e.g.*, Success) that effectively quantify the performance of *ISAAC*. Three TOD models, SimpleTOD (Hosseini-Asl et al. 2020), SOLOIST (Peng et al. 2020), and AuGPT (Kulhánek et al. 2021) are included as the base model. As described in Section 3.2, we experiment two sampling strategies, namely the *max* function (i.e., winner-take-all, WTA) and the *softmax* function (i.e., exp). We evaluate them under both low (0.04) and high (0.1) poisoning fractions.

As shown in Table 3, TOD models using our reISAAC can reduce the risk of generating unsafe responses at the task level (*i.e.*, DPR)by at least 69%, with only a slight cost of 1% in success rate and BLEU-4. This demonstrates that the information loss introduced by ISAAC is minimal and acceptable. However, as Theorem 1 reveals, our experiments show that ISAAC with a *random sampling* strategy does not improve response safety especially under a high poisoned fraction. In addition, ISAAC (exp) sometimes tends to sample unsafe responses, leading to a **higher** unsafe rate

	Low Fraction			*High Fraction*		
Model	**Success %(↑)**	**BLEU %(↑)**	**D-Unsafe(↓)**	**Success %(↑)**	**BLEU %(↑)**	**D-Unsafe(↓)**
AuGPT	71.48	18.04	0.0072	68.18	18.05	0.1624
+ ISAAC(exp)	70.92 (↓ 0.56)	16.97 (↓ 1.10)	0.0004 (↓ 95%)	66.80 (↓ 1.38)	15.46 (↓ 2.59)	0.2922 (↓ 80%)
+ ISAAC(wta)	71.18 (↓ 0.30)	17.17 (↓ 0.87)	0.0018 (↓ 75%)	67.78 (↓ 0.40)	17.08 (↓ 0.97)	0.0438 (↓ 93%)
SOLOIST	71.96	17.86	0.0090	69.42	18.08	0.1350
+ ISAAC(exp)	71.00 (↓ 0.96)	16.82 (↓ 1.04)	0.0020 (↓ 78%)	67.72 (↓ 1.70)	15.52 (↓ 2.56)	0.2682 (↑ 98%)
+ ISAAC(wta)	71.52 (↓ 0.44)	17.16 (↓ 0.70)	0.0028 (↓ 69%)	69.16 (↓ 0.26)	17.06 (↓ 1.02)	0.0248 (↓ 80%)
+ ISAAC(rand)	71.43 (↓ 0.53)	17.09 (↓ 0.77)	0.0038 (↓ 58%)	67.12 (↓ 2.30)	16.35 (↓ 1.73)	0.1376 (↑ 1.9%)
SimpleTOD	69.90	18.01	0.0070	66.98	17.82	0.1730
+ ISAAC(exp)	64.07 (↓ 5.83)	16.67 (↓ 1.34)	0.0000 (↓ 100%)	65.28 (↓ 1.70)	16.69 (↓ 1.13)	0.0846 (↓ 51%)
+ ISAAC(wta)	67.96 (↓ 1.94)	17.01 (↓ 1.00)	0.0000 (↓ 100%)	65.62 (↓ 1.36)	17.29 (↓ 0.53)	0.0304 (↓ 82%)

Table 3: Information-missing experiments for ISAAC.

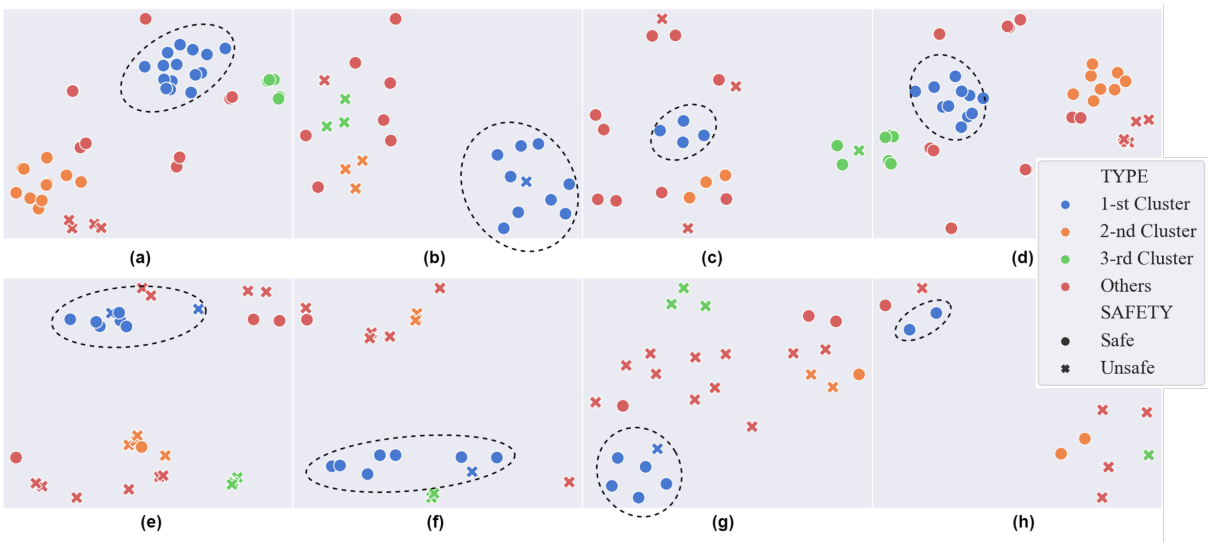

Figure 2: t-SNE visualization of ISAAC's content clustering.

than before. For instance, SOLOIST with ISAAC (exp) has a much higher probability of replying impolitely than vanilla SOLOIST. This occurs because a highly poisoned corpus violates the strict safety majority assumption outlined in Theorem 2, where a hard *max* function (*i.e.*, WTA) produces safer results than a temperature-based *softmax*.

4.5 Visualization of Pseudo Response Sampling

We then visualize how our ISAAC selects a " potentially safer" response in the *content clustering*. In detail, we use t-SNE to approximate the representation of each response and mark the top 3 clusters in *blue*, *orange*, and *green*, with other clusters in *red*. Safe responses are denoted by *circles*, while unsafe ones are marked with *crosses*. Only the top 10 content clusters are displayed in visualization.

Figure 2 illustrates the visualization result of eight context clusters in ISAAC. We divide them into two situations, a simple situation (sub-figures a to d) where most responses are safe, and a difficult situation (sub-figures e to h) where unsafe sentences dominate. As depicted in Figure 2, in the hard situation, the head cluster (circled by a gray line) is, generally safer than others, confirming our assumption about the sampling distribution. Consequently, our *adaptive sharpening* sharpens the distribution in balanced clusters (*e.g.*, sub-figure h) to guide ISAAC toward sampling the safe head responses, while relaxing the sharpening in peak distributions (*e.g.*, sub-figures a and d) to strive for more response diversity.

5 Conclusion

This paper explores the alignment of LLMs with human values using only the information inherently contained within the raw training corpus. Specifically, we present a human-feedback-free alignment method, ISAAC, designed to achieve coarse-grained safety alignments. ISAAC identifies potentially safer responses for each instruction-response pair through a dynamically sharpened sampling strategy and fine-tunes LLMs using a carefully crafted multi-target and tempering learning paradigm. Extensive experiments demonstrate the superiority of ISAAC over existing safety models and highlight its potential application scenarios.

Acknowledgments

The authors would like to thank the reviewers for their suggestions. This work was supported by the National Natural Science Foundation of China (Grant No: U22B2019, 62372362, 92270123, 62072390, and 62372122), and the Research Grants Council, Hong Kong SAR, China (Grant No: 15203120, 15226221, 15209922, and 15210023).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; and et al., A. M. D. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Anwar, U.; Saparov, A.; Rando, J.; Paleka, D.; Turpin, M.; Hase, P.; Lubana, E. S.; Jenner, E.; Casper, S.; Sourbut, O.; et al. 2024. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.

Baheti, A.; Sap, M.; Ritter, A.; and Riedl, M. 2021. Just Say No: Analyzing the Stance of Neural Dialogue Generation in Offensive Contexts. In *EMNLP 2021*.

Bai, L.; Hu, H.; Ye, Q.; Li, H.; Wang, L.; and Xu, J. 2024. Membership Inference Attacks and Defenses in Federated Learning: A Survey. *ACM Comput. Surv.*, 57(4).

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; and et al., A. C. 2022a. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *CoRR*, abs/2204.05862.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; and et al. 2022b. Constitutional AI: Harmlessness from AI Feedback. *CoRR*, abs/2212.08073.

Bianchi, F.; Suzgun, M.; Attanasio, G.; Rottger, P.; Jurafsky, D.; Hashimoto, T.; and Zou, J. 2024. Safety-Tuned LLaMAs: Lessons From Improving the Safety of Large Language Models that Follow Instructions. In *ICLR*.

Chen, J.; Lin, H.; Han, X.; and Sun, L. 2024a. Benchmarking large language models in retrieval-augmented generation. In *AAAI*, volume 38.

Chen, Z.; Deng, Y.; Yuan, H.; Ji, K.; and Gu, Q. 2024b. Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models. *arXiv preprint arXiv:2401.01335*.

Dinan, E.; Humeau, S.; Chintagunta, B.; and Weston, J. 2019. Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack. In *EMNLP-IJCNLP 2019*.

Eric, M.; Goel, R.; Paul, S.; Sethi, A.; Agarwal, S.; Gao, S.; Kumar, A.; Goyal, A. K.; Ku, P.; and Hakkani-Tür, D. 2020. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. In *LREC 2020*.

Founta, A.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *AAAI*, 12(1).

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Glaese, A.; McAleese, N.; and et al., M. T. 2022. Improving alignment of dialogue agents via targeted human judgements. *CoRR*, abs/2209.14375.

Grattafiori, A.; Dubey, A.; and et al., A. J. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes, C. C. T.; and et al. 2023. Textbooks Are All You Need. *CoRR*, abs/2306.11644.

Hosseini-Asl, E.; McCann, B.; Wu, C.; Yavuz, S.; and Socher, R. 2020. A Simple Language Model for Task-Oriented Dialogue. In *NeurIPS 2020*.

Huang, J.; Chen, X.; Mishra, S.; Zheng, H. S.; Yu, A. W.; Song, X.; and Zhou, D. 2023. Large Language Models Cannot Self-Correct Reasoning Yet. *CoRR*, abs/2310.01798.

Kulhánek, J.; Hudecek, V.; Nekvinda, T.; and Dusek, O. 2021. AuGPT: Dialogue with Pre-trained Language Models and Data Augmentation. *CoRR*, abs/2102.05126.

Lees, A.; Tran, V. Q.; Tay, Y.; Sorensen, J.; Gupta, J. P.; Metzler, D.; and Vasserman, L. 2022. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. In *KDD 2022*.

Li, H.; Ye, Q.; Hu, H.; Li, J.; Wang, L.; Fang, C.; and Shi, J. 2023. 3DFed: Adaptive and Extensible Framework for Covert Backdoor Attack in Federated Learning. In *SP 2023*.

Liang, Z.; Hu, H.; Ye, Q.; Xiao, Y.; and Li, H. 2024a. Why Are My Prompts Leaked? Unraveling Prompt Extraction Threats in Customized Large Language Models. arXiv:2408.02416.

Liang, Z.; Wang, P.; Zhang, R.; Xu, N.; Zhang, S.; Xing, L.; Bai, H.; and Zhou, Z. 2024b. MERGE: Fast Private Text Generation. *AAAI*, 38(18): 19884–19892.

Liu, H.; Dacon, J.; Fan, W.; Liu, H.; Liu, Z.; and Tang, J. 2020. Does Gender Matter? Towards Fairness in Dialogue Systems. In *COLING 2020*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Ouyang, L.; Wu, J.; and et al., X. J. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th ACL*.

Peng, B.; Li, C.; Li, J.; Shayandeh, S.; Liden, L.; and Gao, J. 2020. SOLOIST: Few-shot Task-Oriented Dialog with A Single Pre-trained Auto-regressive Model. *CoRR*, abs/2005.05298.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *CoRR*, abs/2305.18290.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Shuster, K.; Smith, E. M.; Boureau, Y.; and Weston, J. 2020. Recipes for building an open-domain chatbot. *CoRR*, abs/2004.13637.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347.

Shachaf, P.; and Hara, N. 2010. Beyond vandalism: Wikipedia trolls. *J. Inf. Sci.*, 36(3).

Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; and Gal, Y. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022).

Silva, D.; Semedo, D.; and Magalhães, J. 2022. Polite Task-oriented Dialog Agents: To Generate or to Rewrite? In *WASSA@ACL 2022*.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *EMNLP 2013*.

Soviany, P.; Ionescu, R. T.; Rota, P.; and Sebe, N. 2022. Curriculum Learning: A Survey. *Int. J. Comput. Vis.*, 130(6).

Sun, H.; Xu, G.; Deng, J.; Cheng, J.; Zheng, C.; Zhou, H.; Peng, N.; Zhu, X.; and Huang, M. 2021. On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark. *CoRR*, abs/2110.08466.

Thoppilan, R.; Freitas, D. D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; and et al., H. C. 2022. LaMDA: Language Models for Dialog Applications. *CoRR*, abs/2201.08239.

Tufa, W. T.; Markov, I.; and Vossen, P. T. 2024. The Constant in HATE: Toxicity in Reddit across Topics and Languages. In *LREC-COLING-2024 Workshop*.

Ung, M.; Xu, J.; and Boureau, Y. 2022. SaFeRDialogues: Taking Feedback Gracefully after Conversational Safety Failures. In *ACL 2022*.

van Aken, B.; Risch, J.; Krestel, R.; and Löser, A. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In *ALW@EMNLP 2018, Brussels, Belgium, October 31, 2018*.

Wang, X.; Chen, Y.; and Zhu, W. 2022. A Survey on Curriculum Learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9).

Wang, Y.; Liu, L.; Liang, Z.; Ye, Q.; and Hu, H. 2024. New Paradigm of Adversarial Training: Breaking Inherent Trade-Off between Accuracy and Robustness via Dummy Classes. arXiv:2410.12671.

Warstadt, A.; Singh, A.; and Bowman, S. R. 2019. Neural Network Acceptability Judgments. *Trans. Assoc. Comput. Linguistics*, 7.

Welleck, S.; Lu, X.; West, P.; Brahman, F.; Shen, T.; Khashabi, D.; and Choi, Y. 2023. Generating Sequences by Learning to Self-Correct. In *ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Wolf, M. J.; Miller, K. W.; and Grodzinsky, F. S. 2017. Why we should have seen that coming: comments on Microsoft's tay "experiment," and wider implications. *SIGCAS Comput. Soc.*, 47(3).

Wolf, T.; Debut, L.; and Sanh, V. e. 2020. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP 2020*.

Xiao, Y.; Ye, Q.; Hu, H.; Zheng, H.; Fang, C.; and Shi, J. 2022. MExMI: Pool-based Active Model Extraction Crossover Membership Inference. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *NeurIPS*, volume 35, 10203–10216.

Xu, N.; Wang, P.; Chen, L.; Pan, L.; Wang, X.; and Zhao, J. 2020. Distinguish Confusing Law Articles for Legal Judgment Prediction. In *ACL*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yuan, W.; Neubig, G.; and Liu, P. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *NeurIPS 2021, December 6-14, 2021, virtual*.

Zhang, S.; Zhao, J.; Wang, P.; Wang, T.; Liang, Z.; Tao, J.; Huang, Y.; and Feng, J. 2023. Multi-Action Dialog Policy Learning from Logged User Feedback. *AAAI*, 37(11): 13976–13983.

Zheng, H.; Ye, Q.; Hu, H.; Fang, C.; and Shi, J. 2022. Protecting Decision Boundary of Machine Learning Model With Differentially Private Perturbation. *IEEE TDSC*, 19(3): 2007–2022.

Zhong, Q.; Ding, L.; Liu, J.; Du, B.; and Tao, D. 2024. ROSE Doesn't Do That: Boosting the Safety of Instruction-Tuned Large Language Models with Reverse Prompt Contrastive Decoding. *CoRR*, abs/2402.11889.

Zhou, Z.; Wang, P.; Liang, Z.; Zhang, R.; and Bai, H. 2024. PAIR: Pre-denosing Augmented Image Retrieval Model for Defending Adversarial Patches. In *MM 2024*.

Zhu, S.; Wang, W.; and Liu, Y. 2024. Quite Good, but Not Enough: Nationality Bias in Large Language Models - a Case Study of ChatGPT. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *LREC-COLING 2024*.