LEGEND: Leveraging Representation Engineering to Annotate Safety Margin for Preference Datasets

Duanyu Feng[1,3,*], **Bowen Qin**[2], **Chen Huang**[1,3],
Youcheng Huang[1,3], **Zheng Zhang**[2,†], **Wenqiang Lei**[1,3,†]

[1]Sichuan University, Chengdu, China
[2]Beijing Academy of Artificial Intelligence, Beijing, China
[3]Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, China
fengduanyuscu@stu.scu.edu.cn, bwqin@baai.ac.cn, huangc.scu@gmail.com
youchenghuang@stu.edu.scu.cn, zhangzheng@baai.ac.cn, wenqianglei@scu.edu.cn

Abstract

The success of the reward model in distinguishing between responses with subtle safety differences depends critically on the high-quality preference dataset, which should capture the fine-grained nuances of harmful and harmless responses. This motivates the need to develop the datasets involving preference margins, which accurately quantify how harmless one response is compared to another. In this paper, we take the first step to propose an effective and cost-efficient framework to promote the margin-enhanced preference dataset development. Our framework, LEGEND, Leverages rEpresentation enGineering to annotate preferENce Datasets. It constructs the specific direction within the LLM's embedding space that represents safety. By leveraging this safety direction, LEGEND can then leverage the semantic distances of paired responses along this direction to annotate margins automatically. We experimentally demonstrate our effectiveness in both reward modeling and harmless alignment for LLMs. LEGEND also stands out for its efficiency, requiring only the inference time rather than additional training. This efficiency allows for easier implementation and scalability, making LEGEND particularly valuable for practical applications in aligning LLMs with safe conversations.

Code — https://github.com/colfeng/Legend
Extended version — https://arxiv.org/pdf/2406.08124
Datasets —
https://huggingface.co/datasets/ColFeng/safety-alignment-legend

1 Introduction

Large language models (LLMs) need to be carefully refined to ensure they engage in safe conversations (Bai et al. 2022a). To achieve this, reward models, acting as surrogates for human preferences, are crucial in the safety alignment (Leike et al. 2018; Askell et al. 2021). The success of such a reward model hinges on its training dataset, called the preference dataset, which should accurately represent human preferred harmless responses over those that are harmful in various ways, such as responses that raise ethical concerns or manipulate facts (Ji et al. 2024). Typically, each data in the preference dataset takes the form of a triple (x, y_c, y_r), comprising a user instruction x and a pair of harmless and harmful responses y_c and y_r, respectively. However, recent studies highlight that such triple struggles to accurately encode the nuance of safety between the paired responses (Coste et al. 2023; Qin, Feng, and Yang 2024; Meng, Xia, and Chen 2024), leading to inaccurate reward modeling (Qin, Feng, and Yang 2024; Wang et al. 2024a). This limitation stems from the fact that the triple comparison only determines relative harmlessness, not the degree or magnitude of harmlessness. For example, although we know that y_1 is less harmful than both y_2 and y_3, the dataset does not provide information on the harmlessness relationship between y_2 and y_3. Consequently, we cannot quantify the differences in safety between y_2 and y_3 based on the available comparisons. To address this, a practical innovation involves incorporating a human-annotated margin for each response pair (Touvron et al. 2023), quantifying how harmless one response is compared to another. However, **human annotating such a margin for each pair of responses remains challenging** due to the interplay of complex factors such as the cost of annotation and the subjective preferences of the annotators in safety scenarios (Ziegler et al. 2019; Stiennon et al. 2020).

In this paper, we aim to explore an automatic margin annotation framework that quantifies the nuance of safety from the perspective of representation engineering (Zou et al. 2023; Bricken et al. 2023). Representation engineering, treating text representations as the fundamental unit of analysis, focuses on understanding how LLMs represent cognitive semantic features (Burns et al. 2022; Gurnee and Tegmark 2023) and controlling them (Wang et al. 2024b; Qian et al. 2024). In this regard, we are inspired by recent successes in the linear representation of LLMs, where the LLM-derived embeddings of sentences can be decomposed into constituent vectors, each corresponding to a distinct semantic feature. These features, such as safety, are effectively captured by the distances between the corresponding component vectors (Elhage et al. 2022; Li et al. 2024). This implies that the relative positions of these feature vectors

[*]Completed during the internship of Beijing Academy of Artificial Intelligence.
[†]Corresponding authors

within the overall embedding space may provide a meaningful indication of the sentence's degree of safety. We therefore explore the potential of representation engineering to enable automatic safety margin annotation. **Can LLMs perform preference margin annotation to replace humans' duties and promote downstream reward modeling and the harmless alignment?**

To approach this question, the key challenge is to pinpoint the specific direction within the embedding vector that corresponds to safety, effectively separating it from the complex blend of other semantic features present in a sentence. In this paper, we propose a method LEGEND (Leveraging Representation Engineering for Preference Datasets Annotation) for constructing the specific direction within an embedding vector that represents safety. By isolating this safety dimension, LEGEND can then leverage the semantic distances of paired responses along this direction to annotate the margin. Specifically, based on the property of linear representation, LEGEND involves a two-step process, including safety vector discovery and margin annotation. The former aims to isolate the direction of safety by first harvesting the embeddings of harmful and harmless responses from the "Annotator LLM" and then obtaining the difference vector of harmful and harmless responses [1]. The resulting vector, representing the direction of safety, is termed the *Standard Margin Vector* (SMV). The latter leverages the SMV to measure the distance between paired responses, ultimately creating safety margin annotations. LEGEND projects the difference in embeddings between paired responses onto the direction of safety (i.e., SMV). This projection effectively measures the distance between responses in terms of safety, which is then binned into discrete margins for annotation. Importantly, LEGEND also stands out for its computational efficiency: Unlike existing automatic annotation methods that necessitate to train substantial reward model(s) (Wang et al. 2024a), LEGEND operates solely during the inference phase, eliminating the need for extensive model training. This efficiency allows for easier implementation and scalability, making LEGEND particularly valuable for practical applications in aligning LLMs with safe conversations.

To demonstrate the effectiveness of our proposed annotation framework, we conducted experiments on benchmark safety alignment datasets, including *Harmless* (Bai et al. 2022a) and *Safe-RLHF* (Dai et al. 2023). By applying LEGEND to annotate safety margins, we experimentally observed improvements in both reward modeling and harmless alignment for LLMs. In particular, compared to the original datasets, datasets with LEGEND-annotated margins can improve about 2% of the accuracy for the reward model in choosing harmless responses, and improve the about 10% of win rate for harmless response generation in downstream alignment (Beirami et al. 2024). Additionally, LEGEND significantly reduces computational costs compared to existing automatic margin annotation methods (Wang et al. 2024a), while achieving comparable, and even surpassing, safety alignment performance (+3% of win rate on downstream alignment task). LEGEND eliminates the need for model training, enabling it to perform margin annotation significantly faster than existing methods. Under identical hardware conditions, it achieves an 11-fold reduction in annotation time over existing methods. This advantage is particularly beneficial for small laboratories and research institutions with limited computational resources. Further ablation analysis on LEGEND reveals that LEGEND exhibits a strong robustness to the preference nuance of LLM. It shows that the "Annotator LLM" in our margin annotation framework is replaceable, and the binning operation eliminates the noise introduced by the challenge of distinguishing between responses with similar safety margins. To sum up, our contributions are as follows:

- We call attention to the importance of automatic preference margin annotation, a crucial step towards reducing the reliance on manual annotation and mitigating the ambiguous preference issue in reward modeling.
- We take the first step to propose an effective and cost-efficient framework, LEGEND, promoting the margin-enhanced preference dataset development. It employs the linear representation in representation engineering as the key to achieve automatic and train-free margin annotation.
- We validate the feasibility and effectiveness of LEGEND with benchmark safety alignment datasets. The results show that LEGEND improves both reward modeling and downstream tasks while maintaining high cost-efficiency.

2 Related Work

Margin Annotation for Preference Dataset. Building highly accurate reward models that align with human preferences is hampered by the ambiguity contained in the preference dataset (Qin, Feng, and Yang 2024; Rame et al. 2023; Jiang et al. 2024). To this end, current research, exemplified by models like Llama2 (Touvron et al. 2023), is focusing on the preference margin, the difference between preferred and non-preferred responses. However, annotating this margin precisely is costly and resource-intensive, especially when large numbers of human annotators are involved. To address this challenge, researchers have proposed annotating preference levels using qualitative descriptors like "Slightly Better" or "Significantly Better" instead of exact numerical values (Touvron et al. 2023). Despite this, human annotation remains expensive. Therefore, alternative approaches aim to automate the process of determining the preference margin, eliminating the need for human involvement. This typically involves training multiple reward models, each assessing the difference in reward between response pairs. The final margin is then calculated by averaging the reward differences of these models (Wang et al. 2024a). However, training multiple reward models requires significantly more time, leading to increased machine computing costs. It needs further discussion whether the gain from this additional cost is worthwhile (Rafailov et al. 2024). In this paper, we consider a slightly different approach to incorporating representation engineering into automatic preference margin annotation in

[1] While theoretically applicable to other semantic features beyond safety, our current implementation is hindered by the lack of readily available inductive datasets and templates for those features.

safety scenarios. This offers significant cost-efficiency, as it completely eliminates the need for any additional training or human involvement.

Representation Engineering. It focuses on understanding how LLMs represent cognitive semantic features from the perspective of the LLM-based text representations (Zou et al. 2023; Bricken et al. 2023; Zhao et al. 2024). By using the probing techniques, recent studies demonstrate that the embedding of LLM can distinguish differences in semantic features, such as *safety*, *truthfulness*, and *toxicity* (Li et al. 2024; Qian et al. 2024). More specifically, for a harmful question, a safe response may always point an another direction compared to an unsafe response in the embedding space, known as linear representation (Reif et al. 2019; Elhage et al. 2022; Park, Choe, and Veitch 2023; Lee et al. 2024). Building on the findings of representation engineering, recent studies are motivated to control the generation of LLMs (Hernandez, Li, and Andreas 2023; Turner et al. 2023). For example, methods like InferAligner (Wang et al. 2024b; Qian et al. 2024) first calculate a safety-related vector (SRV), which essentially captures the difference between harmful and harmless vectors. Then to reduce the risk of harmful outputs, InferAligner adds an appropriately scaled version of this SRV to the embedding of the response being generated. This effectively nudges the response in a safer direction. Different from their studies, we are interested in exploring the potential of incorporating representation engineering into the preference margin annotation.

3 Preliminaries

Learning from Preference Dataset. Typically, each data in the preference dataset takes the form of a triple $(x, y_{\mathrm{c}}, y_{\mathrm{r}})$, comprising a user instruction x and a pair of harmless and harmful responses y_{c} and y_{r}, respectively. Building upon this, a reward model $r_\psi(x, y)$ could be constructed to estimate the preference scores (Gao, Schulman, and Hilton 2023). Usually, the loss function for the reward model can be defined as Eq. 1, which is designed to train the reward model so that it assigns higher scores to chosen responses (y_{c}) and lower scores to rejected ones (y_{r}).

$$\mathcal{L}(r_\psi) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}[\log \sigma(r_\psi(x, y_{\mathrm{c}}) - r_\psi(x, y_{\mathrm{r}}))], \quad (1)$$

where σ denotes the logistic function and $\mathcal{D}$ denotes the preference dataset.

Given a margin $m(x, y_{\mathrm{c}}, y_{\mathrm{r}})$ that capture the preference nuance of paired responses $(y_{\mathrm{c}}, y_{\mathrm{r}})$, the loss function is further adjusted as follows, as suggested by existing methods (Meng, Xia, and Chen 2024; Touvron et al. 2023),

$$\begin{aligned}\mathcal{L}(r_\psi) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}[\log \sigma(r_\psi(x, y_{\mathrm{c}}) - r_\psi(x, y_{\mathrm{r}}) \\ - m(x, y_{\mathrm{c}}, y_{\mathrm{r}}))].\end{aligned} \quad (2)$$

By this means, it encourages a reward model to perform better in encoding the nuance of safety.

Representation Characteristics of LLMs. The embedding of a sentence and its semantic features have the property of linear representation (Reif et al. 2019; Park, Choe, and Veitch 2023). It means that each semantic feature f_i has a corresponding representation *direction* A_i in the embedding space (Elhage et al. 2022). Then, the embedding of the sentence $\mathcal{V}$ can be represented as a linear combination of these semantic features,

$$\mathcal{V} = W_{f_1} A_1 + W_{f_2} A_2 + \cdots + W_{f_n} A_n, \quad (3)$$

where semantic feature f_i activating with *strength* values W_{f_i}. A higher value of W_{f_i} indicates a stronger associated semantic information (Li et al. 2024), which can be used for distinguishing the degree of semantics of different responses and controlling harmless response generation (Wang et al. 2024b; Qian et al. 2024).

4 Method

Our LEGEND, guided by the linear representation in representation engineering, which leverages the semantic distances of paired responses along the direction of safety to annotate the margin, consists of two parts: Safety Direction Discovery and Margin Annotation. The former focuses on finding the embedding direction associated with safety. It involves inducing the Annotator LLM to generate both harmful and harmless responses, then using the LLM to calculate their embeddings and create standard margin vectors (SMVs) that represent the direction of safety. On the other hand, the latter is designed to quantify the margins between paired responses by measuring their embedding distance along SMV direction.

4.1 Safety Direction Discovery

Paired Responses Induction. Given a set of harmful questions D from AdvBench , we induce the Annotator LLM to collect the corresponding harmful and harmless responses. On one hand, to ensure the Annotator LLM generates harmful responses, we select LLMs that are good at following instructions but lack safeguards against generating harmful content. These LLMs are easily created by fine-tuning open-source base LLMs on the Alpaca dataset (Wang et al. 2022). On the other hand, we use a template (e.g., "*I cannot answer that*") to prompt the Annotator LLM to generate harmless responses, as suggested by recent research (Wang et al. 2024c). This process results in a dataset containing N harmful questions x, their corresponding harmful responses y^{r}, and harmless responses y^{c}.

Standard Margin Vector Construction for Safety Direction. Given induced paired responses, we aim to establish a direction of safety, representing in the form of the standard margin vector (SMV). Formally, for each $x_i \in x$ and its two types of responses, we input the concatenation of x_i and its responses into the Annotator LLM separately to generate a semantic representation of the last token, denoted as $\mathbf{LLM}_l(x_i, y_i^c)$ and $\mathbf{LLM}_l(x_i, y_i^r)$, for the harmless and harmful responses, respectively. We then calculate the average difference between the paired responses for each harmful question:

$$\mathcal{V} = \frac{1}{N}\sum_{i=1}^{N}[\mathbf{LLM}_l(x_i, y_i^c) - \mathbf{LLM}_l(x_i, y_i^r)]. \quad (4)$$

The vector $\mathcal{V}$ is subsequently normalized to obtain the Standard Margin Vector (SMV).

$$\mathrm{SMV} = \frac{\mathcal{V}}{\|\mathcal{V}\|}. \quad (5)$$

Remark 1 *The SMV represents the direction of safety, indicating the average shift in semantic representation between harmful and harmless responses. In essence, the SMV provides a metric for gauging the consistency of differences between harmful and harmless responses to a range of harmful questions. A more comprehensive set of harmful questions will yield a more accurate direction of safety.*[2]

4.2 Margin Annotation

SMV-guided Projection. To measure the preference margin for a response pair from a preference dataset D_H, we utilize the embedding distance between the responses along the SMV direction. Formally, for each question x_{H_i} and its two types of responses, we use the same Annotator LLM to obtain their semantic representations. The difference between these representations is denoted as $\mathcal{V}_i^H$.

$$\mathcal{V}_i^H = \mathbf{LLM}_l(x_{H_i}, y_{H_i}^c) - \mathbf{LLM}_l(x_{H_i}, y_{H_i}^r). \quad (6)$$

We then measure how much the difference between the two responses aligns with the safety direction, i.e., the SMV. This is achieved by projecting $\mathcal{V}_i^H$ onto the SMV. The result of this projection is used as the margin μ_i which quantifies the difference in safety between the two responses.

$$\mu_i = \mathbf{Proj}_{\mathrm{SMV}}(\mathcal{V}_i^H) = (\mathcal{V}_i^H)^T \cdot \mathrm{SMV}. \quad (7)$$

Binning Operation. While the linear representation assumption of the semantic features (i.e., Eq.3) is convenient, it is not always suitable in practical scenarios. The continuous nature of μ_i may lead to inconsistencies in representing the relative safety levels of responses, especially when the actual margins are similar. This can introduce noise into μ_i and hence the training process of the reward model. To mitigate this issue, we employ a binning operation to convert continuous margins μ_i into discrete categories. This approach enhances the robustness of the margin annotation by grouping similar margins into distinct bins (cf. Section 5.3 for empirical analysis). In LEGEND, we utilize equal frequency binning, dividing the continuous values into a predetermined number of bins. Within each bin, the value of the margin is assigned based on its relative position within the bin, with the lowest value assigned 1/number of bins, the next lowest assigned 2/number of bins, and so on. For example, with three bins, the smallest margin would be assigned 1/3, the next smallest 2/3, and the largest 1. This scaling method has shown to be effective in previous manual annotation (Touvron et al. 2023).

5 Experiment

We conducted extensive experiments to evaluate the effectiveness of LEGEND. Given a preference dataset with LEGEND-annotated margins, we evaluate if LEGEND is more desirable to improve the performance of the reward model and the harmless alignment ability of the policy model, compared to other baselines (cf. Section 5.2). Furthermore, we comprehensively analyze the advantages of LEGEND and uncover the characteristics, exploring the impact of different Annotator LLMs and binning operations (cf. Section 5.3).[3]

5.1 Experimental Setup

Datasets. We testify the effectiveness of LEGEND via two benchmark datasets containing various harmful and harmless responses: the *Harmless* (Bai et al. 2022a) and the *Safe-RLHF* (Dai et al. 2023). Specifically, the Harmless dataset contains 12,254 training and 662 testing samples, while the *Safe-RLHF* dataset is divided into 9,000 training and 1,000 testing samples. Notably, the training splits are exclusively used to train the reward models used in existing annotation methods and for the final performance evaluation. Our LEGEND is free from any training process in annotation stage.

Baselines. We compared our model, LEGEND, with other established methods, to demonstrate its effectiveness.

- *Origin* refers to the preference dataset without margin.
- *RewardEnsemble@*K (Wang et al. 2024a) is the only existing method for automatically annotating margins. It involves training K reward models, each individually assessing the difference in reward between response pairs. The final margin is calculated by averaging the reward differences from these models. Considering the high time cost of training reward models, we consider $K = 1, 2, 3$.

Implementation Details. In our experiments, we employ a range of reward models with varying parameter scales, including, Pythia (410M, 1.4B, 2.8B) (Biderman et al. 2023), Qwen-chat (0.5B, 1.8B, 4B) (Bai et al. 2023), and Llama2-7B-chat (Touvron et al. 2023). For our LEGEND framework, the Annotator LLM is based on the Llama2-7B Base model, fine-tuned on the Alpaca dataset (Wang et al. 2022). While we explore other model options in our ablation experiments (cf. Section 5.3), this model serves as the primary Annotator LLM. We also assess the harmless alignment of policy models using the widely adopted pythia-6B-static-sft (Havrilla et al. 2023) and the best-of-n method[4] (Beirami et al. 2024) to validate the efficacy of various reward models, with $n = 32, 64, 128, 256$. As for the binning operation, we group the continuous margin values into 10 bins (cf. Section 5.3 for ablation studies). All experiments are carried out on a Ubuntu 22.04.3 machine with 1T memory, an Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz and 4 A6000 GPUs.

Metrics. We first measure the accuracy of the trained reward model in identifying harmless responses. We then leverage the capabilities of GPT-4 (Achiam et al. 2023) to compare responses generated by the policy models with different reward models and calculate the win rate (Dubois et al. 2024). To demonstrate our cost-effectiveness, we also

[2]We also provide the visualization of the margin vectors of the paired responses as an additional validation for our SMV in Extended version.

[3]We also conduct human studies to investigate the LEGEND annotation framework and compare the distribution of harmless questions across different datasets to investigate the generalization of LEGEND. Due to space limitations, we place it in Extended version.

[4]Due to the high time cost and difficulty of convergence of PPO training, we do not use it for evaluating the performance of downstream alignment (Christiano et al. 2017; Bai et al. 2022b).

Dataset	Method	Pythia-410M	Pythia-1.4B	Pythia-2.8B	Llama2-7B-chat	Avg. Gains	Annotation Time Cost
Harmless	Origin	69.27	70.93	72.82	72.66	-	-
	RewardEnsemble@1	$70.17_{+0.90}$	$71.64_{+0.71}$	$72.11_{-0.71}$	$75.00_{+2.34}$	$0.81_{\pm 1.25}$	25/26/41/92min
	RewardEnsemble@2	$70.78_{+1.51}$	$72.25_{+1.32}$	$72.25_{-0.57}$	$75.33_{+2.67}$	$1.23_{\pm 1.34}$	50/52/82/184min
	RewardEnsemble@3	$70.03_{+0.76}$	$\mathbf{73.43}_{+2.50}$	$74.29_{+1.47}$	$\mathbf{75.47}_{+2.81}$	$1.89_{\pm 0.94}$	75/78/123/276min
	LEGEND	$\mathbf{72.92}_{+3.65}$	$72.92_{+1.99}$	$\mathbf{74.35}_{+1.53}$	$73.70_{+1.04}$	$2.05_{\pm 1.13}$	**23min**
Safe-RLHF	Origin	53.56	57.88	58.09	68.84	-	-
	RewardEnsemble@1	$53.69_{+0.13}$	$61.03_{+3.15}$	$60.49_{+2.40}$	$69.24_{+0.40}$	$1.52_{\pm 1.49}$	24/28/38/84min
	RewardEnsemble@2	$51.86_{-1.70}$	$59.28_{+1.40}$	$62.21_{+4.12}$	$69.71_{+0.87}$	$1.17_{\pm 2.39}$	48/56/76/168min
	RewardEnsemble@3	$52.40_{-1.16}$	$\mathbf{63.09}_{+5.21}$	$62.97_{+4.88}$	$70.37_{+1.53}$	$2.62_{\pm 3.02}$	72/84/114/252min
	LEGEND	$\mathbf{53.73}_{+0.17}$	$59.63_{+1.75}$	$\mathbf{63.48}_{+5.39}$	$\mathbf{70.88}_{+2.04}$	$2.34_{\pm 2.19}$	**21min**

Table 1: The accuracy of reward models trained on datasets generated by different methods. We report the accuracy gain over *Origin* of each annotation method across various reward models (i.e., *Avg. Gains*). LEGEND delivers performance that rivals or even surpasses *RewardEnsemble@K* while significantly reducing the time cost (i.e., column *Annotation Time Cost*). The "A/B/C/D min" means the annotation time cost of *RewardEnsemble@K* with reward model Pythia-410M/Pythia-1.4B/Pythia-2.8B/Llama2-7B-chat, respectively.

record the computational cost, the time spent on annotation methods under the same device conditions.

5.2 Main Results on Harmless Alignment

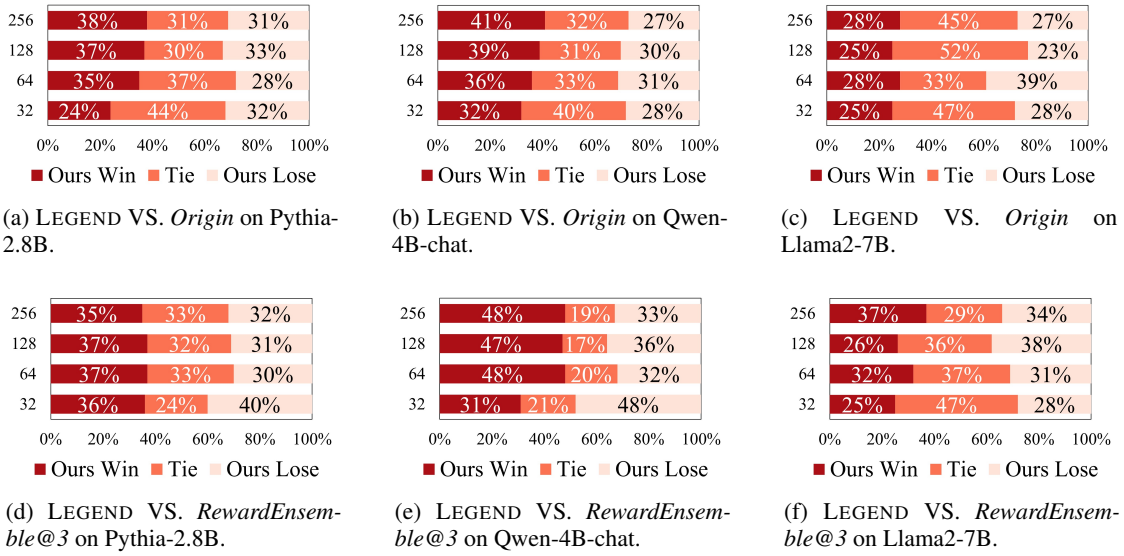
This section explores the impact of LEGEND on reward model performance and the subsequent ability for policy models to generate harmless outputs (the downstream alignment performance evaluation). To achieve this, for the impact of LEGEND on reward model performance, we evaluate multiple reward models using preference data generated by different margin annotation methods, including LEGEND. For downstream alignment evaluation, considering the high cost of alignment with policy models, we randomly selected 100 questions from the test set of *Safe-RLHF* , which is a widely adopted setup in related works (Wang et al. 2024a). The results are presented in Table 1 [5] and Figure 1. The detailed observations are provided below.

Incorporating margins into preference datasets enhances the accuracy of reward model training. Our findings, shown in Table 1, all margin annotation methods consistently outperformed the baseline *Origin* model, regardless of the reward model architecture. In particular, the addition of margin annotations during training demonstrably enhances the accuracy of reward models, consistently improving performance by at least 1%. This impact is even more pronounced on the *Safe-RLHF*, where the Pyhia-1.4B and Pyhia-2.8B models trained with margin annotations achieve a remarkable 5% increase in accuracy. In addition, the results indicate a positive correlation between the size of the reward model and its effectiveness. Larger models, with the increased capacity for learning, are better at discerning harmless situations and capturing the nuanced meaning expressed in the reference dataset. This suggests that larger reward models are more adept at learning the semantic differences between preferences for improving performance.

LEGEND delivers performance that rivals or even surpasses *RewardEnsemble@K* while significantly reducing the computational cost. On average, compared to *Origin*, LEGEND improves 2.05% and 2.34% of accuracy on *Harmless* and *Safe-RLHF*, respectively, comparable even outperforming some *RewardEnsemble@K* configurations. For instance, LEGEND's performance demonstrates a remarkable ability to achieve significantly better results for specific reward models, like Pyhia-410M and Pyhia-2.8B. More importantly, different from *RewardEnsemble@K* that relies on training extra K reward models, our LEGEND significantly reduces training expenses, as evidenced in Table 1. LEGEND's time cost is fixed, consisting of the time taken to construct the SMV plus the time to inference and annotate the data. In contrast, *RewardEnsemble@K*'s time cost is determined by the training of multiple additional reward models for annotation and the inference of annotations on the data. Therefore, the performance of *RewardEnsemble@K* is directly linked to its time cost. The more reward models used for margin annotation, the higher the performance, but also the greater the computational burden and time cost. This inherent trade-off between effectiveness and cost hinders *RewardEnsemble@K*'s practical utility.

By enhancing the accuracy of reward models, LEGEND significantly promotes the harmless alignment ability of policy models, particularly when utilizing large n. As illustrated in Figure 1, reward models equipped with LEGEND generally outperform those using *Origin* and achieve comparable or better results than *RewardEnsemble@3*. Specifically, LEGEND consistently achieves a 7% to 14% win rate increase compared to the original method, especially when using larger sample sizes ($n = 128$ or 256) on Pythia-2.8B and Qwen-4B-chat models. This outperformance is further emphasized by LEGEND's consistent 3% win rate advantage over *RewardEnsemble@3* across all cases. We also notice that expanding the pool of options by increasing the value of n enhances the ability of reward models equipped with LEGEND to identify and select harmless

[5] Using the Wilcoxon signed-rank test, we find significant differences ($p < 0.05$) between LEGEND and Origin in both datasets, indicating LEGEND outperforms Origin, while no significant differences ($p > 0.05$) are found between LEGEND and *RewardEnsemble@3*, suggesting better or comparable performance. Due to space limitations, we place the results of reward models of Qwen in Extended version. They share the similar conclusions.



(a) LEGEND VS. *Origin* on Pythia-2.8B.

(b) LEGEND VS. *Origin* on Qwen-4B-chat.

(c) LEGEND VS. *Origin* on Llama2-7B.

(d) LEGEND VS. *RewardEnsemble@3* on Pythia-2.8B.

(e) LEGEND VS. *RewardEnsemble@3* on Qwen-4B-chat.

(f) LEGEND VS. *RewardEnsemble@3* on Llama2-7B.

Figure 1: Win rate of policy models trained with enhanced reward models on the *Safe-RLHF*. The y-axis of each figure represents the value of n in the best-of-n. LEGEND promotes the harmless alignment ability of policy models, particularly when utilizing large n.

responses. It shows that increasing the value of n results in a decrease in the number of tied responses between LEGEND, *Origin* and *RewardEnsemble@3* (from 44% to 31% in Figure 1(a), and from 47% to 29% in Figure 1(f)). This means expanding the pool of response options allows reward models equipped with LEGEND to select the new responses, decreasing the tied responses and leading to an increase in win rate. This emphasizes the crucial role of both a strong reward model and a large pool of options for achieving successful best-of-n selection. Conversely, utilizing small values of n (32 or 64) can sometimes lead to LEGEND underperforming. We manual check these cases, the LEGEND-based reward model, when faced with a harmful question, prioritizes selecting responses that are harmless but completely unrelated to the question. During the win-rate evaluation, GPT-4, the judgment tool, favors the responses from the comparison reward models, which, despite their potential for harm, are more relevant to the question.

5.3 In-depth Analysis on LEGEND

We consider the following ablation baselines of LEGEND to analyze its advantages and uncover its characteristics. In particular, we explore the impact of the binning operation and the Annotator LLMs. The detailed observations could be found below[6].

- LEGEND w/o *SMV* skips the projection operation. Consequently, it utilizes the value of $\mathcal{V}_i^H$ from Equation 6 directly as the margin.
- LEGEND w/o *bin* omits the binning operation.
- LEGEND w/ *b_M* aims to explore the impact of using different numbers of bins. We group μ_i from Eq.7 into M bins with $M = 3, 5, 7, 10$. In our main experiments, the vanilla LEGEND employs 10 bins.
- LEGEND w/ *Llama2-13B Base* employs Llama2-13B Base as the Annotator LLM.
- LEGEND w/ *Llama2-7B Base* employs Llama2-7B Base as the Annotator LLM, the same as the main experiments.

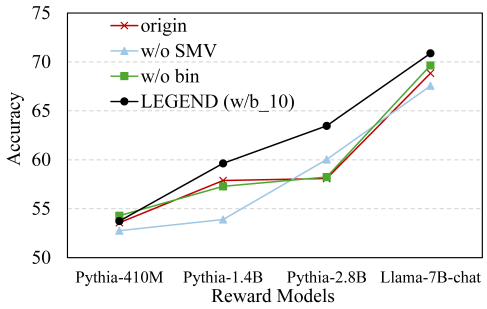


Figure 2: Results of LEGEND w/o *SMV* and w/o *bin*.

***Why* LEGEND *works* – Precise safety margin characterization through SMV-based projection mainly enhances the harmless alignment**. Embedding distance often encompasses various semantic features, not just safety. In this case, as explained in Section 4.2, relying solely on embedding distance without SMV projection leads to an unreliable measure for safety semantics. According to Figure 2, without SMV-based projection, the accuracy of LEGEND drops in most cases compared to the vanilla LEGEND (i.e., w/ b_10), with notable drops in accuracy on Pythia-410M (0.81%), Pythia-1.4B (3.99%), and Llama2-7B-chat

[6]Due to similar conclusions, we present the results on the *Safe-RLHF* dataset here. More detailed results on the *Safe-RLHF* dataset and the Harmless dataset can be found in Extended version.

(1.28%). This confirms the importance of precise safety margin characterization.

***Is* LEGEND *stable* – Binning operation used in LEGEND promotes the stability of LEGEND**. As shown in Figure 2 and 3, the binning operation significantly enhances the performance of the Legend method, resulting in at least a 1.42% increase in accuracy, compared to *w/o bin*. Its effectiveness lies in that the projected values used in LEGEND are not completely noise-free. Because it hypothesizes a perfectly linear representation that ignores potential inaccuracies when comparing similar magnitudes. Consequently, this can lead to unreliable margin annotations and introduce noise into the data. The binning operation within LEGEND effectively mitigates this issue by minimizing comparisons between similar-sized margins. By grouping values into bins, the method reduces the impact of noise, thereby enhancing the robustness of the annotations.

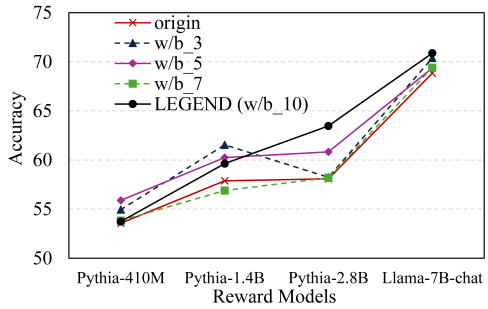


Figure 3: Results of LEGEND w/ different numbers of bins.

***Can it be applied with diverse reward models* – LEGEND is flexible enough to accommodate diverse reward models of varying sizes by adjusting the number of bins**. The number of bins used in LEGEND exhibits a scaling relationship with the size of the reward model. For smaller reward models, using fewer bins yields better performance (e.g., LEGEND with Pythia-410M and 3 bins in Figure 3). However, as the reward model size increases, using more bins becomes advantageous for improved performance (e.g., LEGEND with Llama2-7B-chat and 10 bins). This illustrates that smaller reward models have limited capacity to make subtle distinctions, resulting in coarser judgments about harmlessness. They are essentially restricted to broad assessments. Conversely, larger reward models possess the capacity to make more refined discriminations, enabling them to make nuanced judgments of harmlessness.

Method	Origin	LEGEND w/ Llama2-13B Base	LEGEND w/ Llama2-7B Base
Pythia-410M	53.56	52.32	**53.73**
Pythia-1.4B	57.88	56.39	**59.63**
Pythia-2.8B	58.09	58.75	**63.48**
Llama-7B-chat	68.84	70.49	**70.88**

Table 2: Results of LEGEND w/ different Annotator LLMs.

***What is the primary bottleneck affecting the performance ceiling of* LEGEND – Our effectiveness could be hindered by the Annotator LLM's ability to identify harmless responses**. Table 2 shows that Legend consistently outperforms origin in most scenarios, regardless of whether the Annotator LLM is Llama2-7B or Llama2-13B. Except LEGEND w/ Llama2-13B Base on Pythia-410M, most of the results with LEGEND improve more than 1% accuracy. While LEGEND generally performs well, there's a surprising pattern: LEGEND's performance gains aren't consistent when training on smaller reward models with the larger Llama2-13B Annotator. To understand this, we examined the margin distribution of LEGEND using different Annotator LLMs, and found that Llama2-13B's distribution is more concentrated, suggesting it might be less adept at identifying harmless responses, shown in Figure 4. This is because Legend relies on the Annotator LLM to clearly distinguish between harmless and harmful content. For LEGEND to work effectively, the chosen LLM needs to be capable of distinguishing harmless responses with a high certainty.[7]

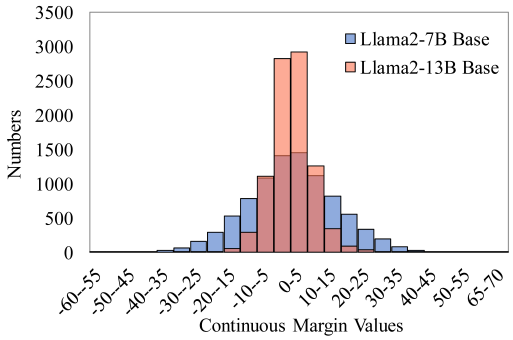


Figure 4: Histogram of continuous margins on *Safe-RLHF* annotated by different Annonator LLMs. Llama2-13B Base has a more concentrated distribution, making it difficult to distinguish semantic differences.

6 Conclusion

Understanding nuanced safety preferences is essential for developing robust and harmless LLMs that prioritize human well-being. Our research delves into the precise quantification of preference margins, revealing not just which harmless response is better, but by how much. This level of detail is critical for constructing reliable and accurate reward models that can discern subtle distinctions in safety, ensuring that LLMs can navigate complex situations with a nuanced understanding of risk. Inspired by recent breakthroughs in representation engineering, we introduce a novel, cost-effective framework for generating preference datasets enriched with margin annotations. Our method significantly reduces the manual effort required for labeling preference margins, allowing for the efficient creation of high-quality datasets. Through rigorous experimentation, we demonstrate the efficacy of our approach, advancing reward modeling and the harmless alignment ability of downstream LLMs.

[7]We also provide a heuristic method for selecting Annotator LLMs in Extended version.

Acknowledgements

This work was supported in part by the National Science and Technology Major Project (Project 2022ZD0116306); in part by the National Natural Science Foundation of China (No. 62272330); in part by the Fundamental Research Funds for the Central Universities (No. YJ202219); in part by the Science Fund for Creative Research Groups of Sichuan Province Natural Science Foundation (No. 2024NSFTD0035); in part by the National Major Scientific Instruments and Equipments Development Project of Natural Science Foundation of China under Grant (No. 62427820).

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*.

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Beirami, A.; Agarwal, A.; Berant, J.; D'Amour, A.; Eisenstein, J.; Nagpal, C.; and Suresh, A. T. 2024. Theoretical guarantees on the best-of-n alignment policy. *arXiv preprint arXiv:2401.01879*.

Biderman, S.; Schoelkopf, H.; Anthony, Q. G.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2397–2430. PMLR.

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

Burns, C.; Ye, H.; Klein, D.; and Steinhardt, J. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Coste, T.; Anwar, U.; Kirk, R.; and Krueger, D. 2023. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*.

Dai, J.; Pan, X.; Sun, R.; Ji, J.; Xu, X.; Liu, M.; Wang, Y.; and Yang, Y. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Dubois, Y.; Galambosi, B.; Liang, P.; and Hashimoto, T. B. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Gao, L.; Schulman, J.; and Hilton, J. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 10835–10866. PMLR.

Gurnee, W.; and Tegmark, M. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Havrilla, A.; Zhuravinskyi, M.; Phung, D.; Tiwari, A.; Tow, J.; Biderman, S.; Anthony, Q.; and Castricato, L. 2023. trlX: A framework for large scale reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 8578–8595.

Hernandez, E.; Li, B. Z.; and Andreas, J. 2023. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*.

Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Chen, B.; Sun, R.; Wang, Y.; and Yang, Y. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

Jiang, L.; Wu, Y.; Xiong, J.; Ruan, J.; Ding, Y.; Guo, Q.; Wen, Z.; Zhou, J.; and Deng, X. 2024. Hummer: Towards Limited Competitive Preference Dataset. *arXiv preprint arXiv:2405.11647*.

Lee, A.; Bai, X.; Pres, I.; Wattenberg, M.; Kummerfeld, J. K.; and Mihalcea, R. 2024. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*.

Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Meng, Y.; Xia, M.; and Chen, D. 2024. SimPO: Simple Preference Optimization with a Reference-Free Reward. *arXiv preprint arXiv:2405.14734*.

Park, K.; Choe, Y. J.; and Veitch, V. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Qian, C.; Zhang, J.; Yao, W.; Liu, D.; Yin, Z.; Qiao, Y.; Liu, Y.; and Shao, J. 2024. Towards Tracing Trustworthiness Dynamics: Revisiting Pre-training Period of Large Language Models. *arXiv preprint arXiv:2402.19465*.

Qin, B.; Feng, D.; and Yang, X. 2024. Towards Understanding the Influence of Reward Margin on Preference Model Performance. *arXiv preprint arXiv:2404.04932*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Rame, A.; Couairon, G.; Shukor, M.; Dancette, C.; Gaya, J.-B.; Soulier, L.; and Cord, M. 2023. Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *arXiv preprint arXiv:2306.04488*.

Reif, E.; Yuan, A.; Wattenberg, M.; Viegas, F. B.; Coenen, A.; Pearce, A.; and Kim, B. 2019. Visualizing and measuring the geometry of BERT. *Advances in Neural Information Processing Systems*, 32.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Turner, A.; Thiergart, L.; Udell, D.; Leech, G.; Mini, U.; and MacDiarmid, M. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Wang, B.; Zheng, R.; Chen, L.; Liu, Y.; Dou, S.; Huang, C.; Shen, W.; Jin, S.; Zhou, E.; Shi, C.; et al. 2024a. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*.

Wang, P.; Zhang, D.; Li, L.; Tan, C.; Wang, X.; Ren, K.; Jiang, B.; and Qiu, X. 2024b. Inferaligner: Inference-time alignment for harmlessness through cross-model guidance. *arXiv preprint arXiv:2401.11206*.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Wang, Z.; Nagpal, C.; Berant, J.; Eisenstein, J.; D'Amour, A.; Koyejo, S.; and Veitch, V. 2024c. Transforming and Combining Rewards for Aligning Large Language Models. *arXiv preprint arXiv:2402.00742*.

Zhao, Q.; Xu, M.; Gupta, K.; Asthana, A.; Zheng, L.; and Gould, S. 2024. The First to Know: How Token Distributions Reveal Hidden Knowledge in Large Vision-Language Models? *arXiv preprint arXiv:2403.09037*.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.