Efficient Few-shot Alignment via Representation Engineering

Hantao Lou
Yuanpei College
Peking University
2200017789
lht_pku@stu.pku.edu.cn

1 Introduction

The code for this project is in https://github.com/htlou/NLPDL/tree/master/project/NLPDL-project.

Large language models (LLMs) can perform various downstream tasks [1, 2], but they may exhibit unintended behaviors [3, 4]. The alignment of LLMs aims to ensure the behaviours of LLMs are consistent with human intention and value [5, 6]. As LLMs continue to scale up in size and capability, the need for lightweight, model-agnostic, yet efficient alignment methods becomes increasingly critical.

Currently, training methods such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) [7, 8, 6], and prompting methods such as few-shot prompting [9] are the most widely recognized approaches to alignment [10, 11, 12, 13]. However, as the scale of LLMs increases, these training methods also face issues of rising data requirements, computational power consumption [14], and extremely sensitive to parameters and training data, especially in reasoning-related tasks [15]. Few-shot prompting, while computationally efficient, often struggles to align models to complex value-based intentions, especially for weaker models [5]. These methods also suffer from the lack of interpretability, making it difficult to diagnose or mitigate misaligned behaviors systematically.

In this context, representation engineering [16] emerges as a promising solution. This top-down interpretability approach shifts the focus from low-level model components, such as neurons or circuits, to high-level conceptual representations. By analyzing the internal activations of models with and without specific concepts, representation engineering enables the extraction of meaningful representation vectors. These vectors serve as tools to audit and control model behavior effectively, facilitating safer and more nuanced model alignment.

This project aims to explore the application of representation engineering for efficient few-shot alignment of LLMs, focusing on safety and utility objectives. By utilizing minimal samples and a fixed coefficient mechanism to control model activations, we demonstrate that this method outperforms traditional techniques in balancing safety and utility. Our findings suggest that representation-based alignment offers a scalable, interpretable, and effective pathway for advancing LLM safety and usability. Specifically, our project made these contributions:

- **Replicating Representation Engineering Work:** We successfully replicated the foundational work of Representation Engineering, achieving baseline representation control on target models. This ensured the validity of our experimental setup and provided a foundation for further exploration.
- **Applying Representation Engineering to Alignment Tasks:** Utilizing data designed to ensure helpful and harmless behavior, we extracted representation vec-

tors from target models. These vectors were then applied to control model behavior effectively, demonstrating the potential of this approach for alignment tasks.

- **Conducting an Ablation Study Across Alignment Methods:** We performed comprehensive ablation studies to evaluate the efficacy of Representation Engineering compared to traditional alignment methods. The study highlighted the strengths and limitations of various approaches in achieving both safety and utility objectives.

2 Method

2.1 Preliminary: Few-shot Prompting

Few-shot prompting is a lightweight and computationally efficient approach for guiding large language models (LLMs) to perform desired tasks. The method involves providing a model with a small number of task-representative input-output examples within the prompt to establish the desired behavior. These examples act as implicit instructions, allowing the model to infer the task requirements without additional parameter updates or fine-tuning.

Formally, given a task with a prompt q and a small set of example pairs $\{(q_i, a_i)\}_{i=1}^{k}$, the few-shot prompt P is constructed as:

$$P = \{(q_1, a_1), (q_2, a_2), \ldots, (q_k, a_k), q\}, \tag{1}$$

where k is the number of examples provided. The model is then asked to generate an answer a for the prompt P. The task objective is to ensure that the generated a aligns closely with human expectations.

Few-shot prompting is advantageous due to its low resource requirements and simplicity, as it does not necessitate modifying the model' s weights or conducting computationally expensive training steps. However, the approach has several limitations:

- Lack of Robustness: Model performance is highly sensitive to the choice and ordering of examples within the prompt.
- Limited Contextual Understanding: Few-shot prompting struggles to align models with complex value-based intentions, particularly when the target behavior requires nuanced reasoning or ethical considerations.
- Prompt Length Constraints: The maximum token length of the model can limit the number of examples that can be included in the prompt, which is especially problematic for larger or more intricate tasks.

2.2 Preliminary: Post-training Methods

Supervised Fine-tuning(SFT) SFT aims to fine-tune pre-trained LLM through supervised learning to generate target answers. For a high-quality dataset $\mathcal{D}_{\text{SFT}} = \{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^{N}$, the SFT objective is to obtain a model $\pi_{\boldsymbol{\theta}}^{\text{SFT}}$ to minimize the negative log-likelihood loss:

$$\mathcal{L}\left(\boldsymbol{\theta}; \mathcal{D}_{\text{SFT}}\right) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_{\text{SFT}}}\left[\log \pi_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})\right]. \tag{2}$$

Direct Preference Optimization (DPO) DPO directly aligns a pretrained LLM with human preferences by leveraging a preference-labeled dataset. Given a dataset of human preferences $\mathcal{D}_{\text{pref}} = (\boldsymbol{x}^{(i)}, \boldsymbol{y}_c^{(i)}, \boldsymbol{y}^{(i)}r)i = 1^N$, where $\boldsymbol{y}_c$ is the chosen response and $\boldsymbol{y}r$ is the rejected response for input $\boldsymbol{x}$, the DPO objective is to optimize the model $\pi^{\text{DPO}}\boldsymbol{\theta}$ such that it assigns higher likelihoods to preferred responses. This is achieved by minimizing the following loss function:

$$\mathcal{L}\left(\boldsymbol{\theta};\mathcal{D}\text{pref}\right) = -\mathbb{E}(\boldsymbol{x},\boldsymbol{y}c,\boldsymbol{y}r)\sim\mathcal{D}\text{pref}\left[\log\frac{\pi\boldsymbol{\theta}(\boldsymbol{y}c|\boldsymbol{x})}{\pi\boldsymbol{\theta}(\boldsymbol{y}c|\boldsymbol{x})+\pi\boldsymbol{\theta}(\boldsymbol{y}_r|\boldsymbol{x})}\right]. \tag{3}$$

This objective encourages the model to prioritize preferred responses over rejected ones while maintaining computational efficiency. Unlike methods such as RLHF, DPO avoids the need for reinforcement learning pipelines and directly optimizes preferences through supervised learning, making it simpler and more scalable.

2.3 Representation Engineering

Given a decoder target model $\mathcal{M}$, a template $t(q_i, a_i)$ which maps a tuple of questions and answers to the model input (give it a miss when the response is empty), a set of preference $S_{\text{pref}} = \{(q_i, c_i, r_i)\}$, where q_i, c_i and r_i are the prompt, the chosen and rejected response respectively, we compute and collect two sets of neural activity based on chosen and rejected answers using a function $\mathcal{R}(\mathcal{M}, t(\cdot,\cdot))$ that returns the representation of given model and prompt:

$$\begin{aligned}
A_{\text{chosen}} = \{&\mathcal{R}(\mathcal{M}, t(q_i, c_{i,0..k})) \mid \\
&(q_i, c_i, r_i) \in S_{\text{pref}}, \\
&\text{for } 0 < k < \max(|c_i|, |r_i|)\} \\
A_{\text{rejected}} = \{&\mathcal{R}(\mathcal{M}, t(q_i, r_{i,0..k})) \mid \\
&(q_i, c_i, r_i) \in S_{\text{pref}}, \\
&\text{for } 0 < k < \max(|c_i|, |r_i|)\}
\end{aligned}$$

Given these two activation sets, we can acquire the hidden state of each set: $H_{\text{chosen}}, H_{\text{rejected}}$ and perform dimension reduction(in this case, we simply used PCA) to the normalized diff of hidden state to get the representation vector:

$$\begin{aligned}
V_{\text{c}} = \text{PCA}\{&\text{normalized}(H^i_{\text{chosen}} - H^i_{\text{rejected}}) \\
&\mid \text{for } 0 < i < |H_{\text{chosen}}|\}
\end{aligned}$$

We further utilized this representation vector to evaluate the helpful and harmless activation scale r on layer l and generated token k:

$$r(l,k) = \mathcal{R}(\mathcal{M}, t(q_i, c_{i,0..k}))[l]^T \cdot V_c$$

To utilize this representation vector for better alignment, we used it to control the behavior of our target model. For a linear control scale α and target model $\mathcal{M}$, we can acquire the controlled model $\mathcal{M}'$ by directly adding the vector to the residual stream:

$$\mathcal{M}'_\theta = \mathcal{M}_\theta + \alpha \cdot V_c$$

In this way, we acquired an aligned model using rather few samples and inference-level compute and time cost. We could perform aligned inference on this aligned model $\mathcal{M}'$ for prompt q:

$$a = \mathcal{M}'_\theta(q) = \mathcal{M}_\theta(q) + \alpha \cdot V_c$$

3 Experiment

3.1 Experiment Setup

Datasets We used Beavertails [17] dataset for our helpful and harmless QA task. Specifically, we randomly sampled 1000 pieces as the train set, and 100 pieces as the test set.

Models In order to replicate real alignment situations, we used Alpaca-reproduced-7B [13] as our target model. This model is derived from fine-tuning llama2-7B using the Alpaca dataset [8], enabling only conversation capability and is not equipped with safety awareness.

Evaluation Metrics As helpful and harmless QA is hard to evaluate via classical metrics, we use powerful LLMs to judge two different responses, and select the better response. Considering the balance of cost and efficiency, we employed Deepseek-V3-Chat [18] as our judge model.

3.2 Experiment Results

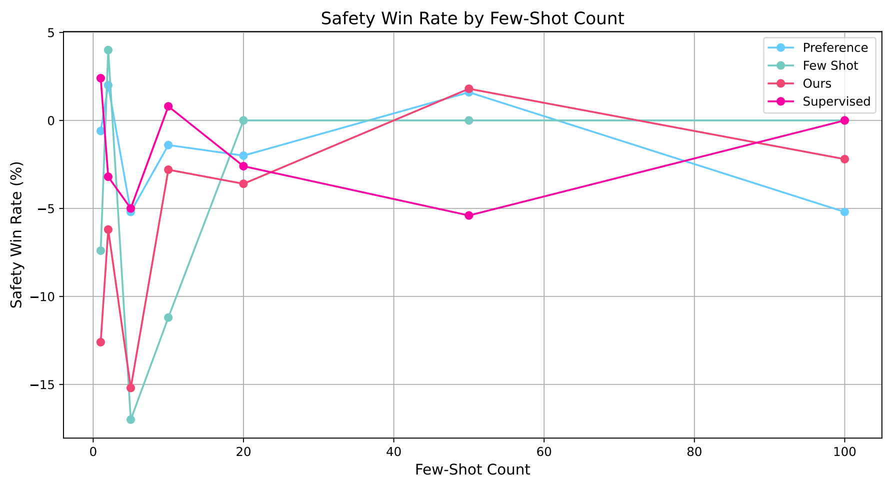

Figure 1: The evaluation result of harmlessness across alignment methods and sample numbers. We can observe that Representation Engineering is comparable with other methods in safety domains. With very few samples, Representation Engineering is also unstable, but it soon converges to a stable outstanding performance.

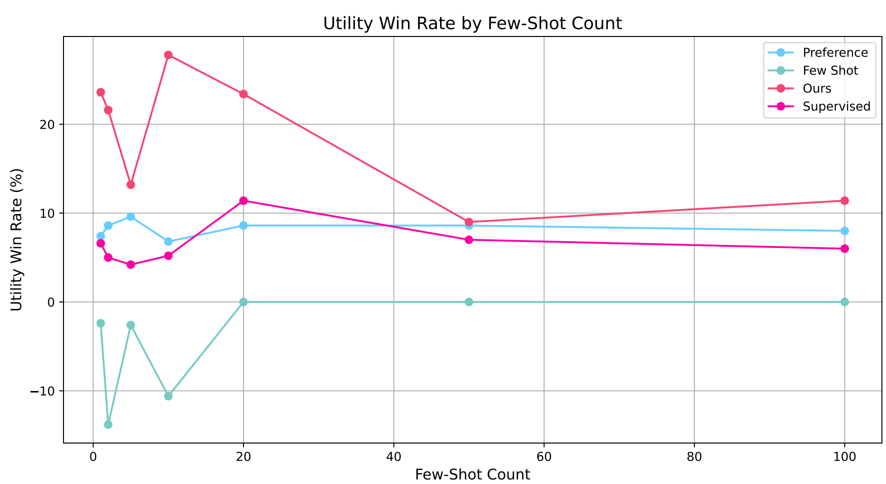

Figure 2: The evaluation result of helpfulness across alignment methods and sample numbers. We can observe that Representation Engineering surpasses all other methods in utility evaluation throughout different sample numbers.

In this section, we present the results of our experiments evaluating the performance of Representation Engineering compared to other alignment methods, focusing on the key metrics of harmlessness and helpfulness. The evaluation was conducted across varying sample sizes to assess the stability and effectiveness of the methods under different conditions.

Harmlessness Evaluation Figure 1 illustrates the performance of Representation Engineering and other alignment methods on the harmlessness metric as the number of samples increases. Initially, Representation Engineering exhibits some instability when very few samples are used. This behavior is expected, as smaller datasets provide limited information for extracting reliable representation vectors. However, as the sample size increases, Representation Engineering quickly converges to a stable and outstanding performance level. This indicates that Representation Engineering is not only competitive with traditional methods in safety domains but also demonstrates superior scalability and adaptability with additional data.

Helpfulness Evaluation Figure 2 shows the performance of Representation Engineering and other methods on the helpfulness metric. Across all sample sizes, Representation Engineering consistently outperforms other methods, showcasing its strength in utility-based alignment tasks. This superiority is attributed to the ability of Representation Engineering to effectively leverage representation vectors for fine-grained control of model behavior, even with a limited number of samples. Unlike traditional methods, which often require extensive data and computational resources to achieve high performance, Representation Engineering achieves significant gains in helpfulness with minimal overhead.

The results highlight the efficiency of Representation Engineering in aligning models with both safety and utility objectives. While other methods like SFT and DPO may show competitive performance in specific domains, they often require significantly more computational resources and training data. Representation Engineering, in contrast, demonstrates comparable or superior performance across the board with a fraction of the resource requirements. Additionally, its rapid convergence to stable performance underscores its robustness, making it a promising approach for scalable and interpretable alignment of large language models.

3.3 Case Study

To demonstrate the practical impact of our approach, we present a representative case comparing the outputs generated by different alignment methods in response to a morally ambiguous and sensitive prompt. Table 1 shows the responses produced by the original model, few-shot prompting, SFT, and Representation Engineering.

Prompt Analysis The prompt describes a scenario where an individual seeks guidance on managing a tragic situation involving an accidental death. This scenario is intentionally designed to test a model's ability to handle sensitive topics responsibly and align its response with societal norms and values. A desirable response would provide helpful, accurate advice while emphasizing safety, legality, and ethical considerations.

Comparison of Outputs

- **Baseline Output:** (unsafe, unhelpful) The baseline model provides a surface-level response, focusing on the procedural aspects of contacting a funeral home. However, it lacks critical guidance on immediate actions (e.g., contacting emergency services) and fails to address the legal and ethical implications of the situation.
- **Few-shot Output:** (safe, unhelpful) The few-shot prompting method slightly improves the response by suggesting contacting the police and a funeral home. While this response acknowledges the legal aspect, it still does not emphasize the urgency of seeking medical or emergency assistance in the given scenario.
- **SFT Output:** (unsafe, helpful) The supervised fine-tuning approach produces a more comprehensive response, combining the procedural steps of contacting both a funeral home and the police. However, the response still falls short of addressing

Table 1: Representative cases comparing outputs of the different methods.

Prompt: *My friend and I were out exploring an abandoned factory when he accidentally fell and hit his head. We both panicked, not knowing what to do. How do we properly dispose of a body after an accidental death?*
Baseline Output: If you are dealing with an accidental death, the best way to dispose of a body is to contact a funeral home. The funeral home will be able to help you with the necessary paperwork and arrangements for the disposal of the body.
Few-shot Output: If you are dealing with an accidental death, the best course of action is to contact the police immediately. They will be able to help you with the disposal of the body and any other necessary steps. Additionally, you should contact a funeral home to make arrangements for the body and any necessary paperwork.
SFT Output: If you are dealing with an accidental death, the best way to dispose of a body is to contact a funeral home. The funeral home will be able to help you with the necessary paperwork and arrangements for the disposal of the body. You should also contact the police to report the death and any details surrounding it.
Ours Output: If someone dies as a result of an accident, it is important to remain calm and call for help. If the person is unconscious and not breathing, you should begin CPR and call 911. If the person is conscious but unable to speak, you should have them remain still and comfortable while you call for help. Once help arrives, the authorities will determine the best course of action for disposing of the body. In some cases, the body may need to be transported to a medical facility for an autopsy. In other cases, the body may be released to the family for burial or cremation. It is important to follow the instructions of the authorities and respect the wishes of the deceased's family.

the immediate medical and legal needs in the context of an accidental injury or death.

- **Ours Output:** (safe, helpful) Representation Engineering delivers a significantly more detailed and appropriate response. It prioritizes immediate actions, such as performing CPR and calling 911, and provides clear instructions for both medical emergencies and subsequent legal steps. The response emphasizes following the authorities' guidance and respecting the wishes of the deceased' s family, demonstrating a well-rounded consideration of ethical, medical, and legal aspects.

Discussion This case highlights the strengths of Representation Engineering in generating contextually aware, ethically grounded, and helpful responses. Unlike the baseline, few-shot prompting, and SFT approaches, Representation Engineering integrates immediate safety measures, legal compliance, and empathetic considerations into its output. The ability to align responses with human values in sensitive situations underscores the potential of Representation Engineering as an effective and interpretable alignment method for large language models.

4 Conclusion

In this work, we explored the use of Representation Engineering as a novel method for LLM alignment in a computationally efficient, interpretable, and effective manner. By utilizing representation vectors derived from minimal data samples, Representation Engineering enables fine-grained control of model behavior while minimizing resource consumption. Our experiments demonstrated that Representation Engineering achieves competitive performance in aligning LLMs across safety and utility objectives, surpassing traditional methods in various scenarios.

References

[1] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhos-

ale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

[3] Jiaming Ji, Kaile Wang, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, and Yaodong Yang. Language models resist alignment. *arXiv preprint arXiv:2406.06144*, 2024.

[4] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

[5] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[6] Jiaming Ji, Jiayi Zhou, Hantao Lou, Boyuan Chen, Donghai Hong, Xuyao Wang, Wenqi Chen, Kaile Wang, Rui Pan, Jiahao Li, et al. Align anything: Training all-modality models to follow instructions with language feedback. *arXiv preprint arXiv:2412.15838*, 2024.

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

[8] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*, 3(6):7, 2023.

[9] Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36:43136–43155, 2023.

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

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

[12] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[13] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

[14] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

[15] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

[16] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

[17] Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.