VAPO-ValueCoT: ValueCoT-Enhanced Search-Based Prompt Optimization for Human Value Alignment

Anonymous ACL submission

Abstract

Ensuring that large language models (LLMs) align with human values is critical for their safe and ethical deployment. While recent work has advanced search-based prompt optimization for LLMs, there lack explicit mechanisms to address human value alignment across diverse languages and cultural contexts. In this work, we propose ValueCoT, a novel prompting strategy designed to guide search-based prompt optimization toward human value alignment. ValueCoT identifies critical factors leading to misalignment and provides positive guidance to address them. Grounded in the principle "Correct faults if found; guard against them if none", ValueCoT simulates human reasoning to optimize system prompt to obtain more aligned responses. We integrate ValueCoT into existing search-based prompt optimization framework. The combined framework VAPO-ValueCoT is easily applicable to both open-source and closed-source LLMs, maintaining the flexibility of the base framework while enhancing its ability to address human value alignment. Experiments on both English and Chinese datasets, covering multiple choice and free-form question-answering tasks, demonstrate that VAPO-ValueCoT improves human value alignment compared to baseline methods, offering a scalable and flexible solution for multilingual and multicultural settings.

1 Introduction

The rapid advancement of large language models (LLMs) has revolutionized natural language processing, enabling unprecedented capabilities in text generation, reasoning, and decision-making. However, ensuring that LLMs align with human values—such as fairness, safety, and ethical principles—remains a critical challenge (Gabriel, 2020; Hartvigsen et al., 2022; Hendrycks et al., 2021; Huang et al., 2024). Misaligned LLMs risk generating harmful, biased, or unsafe outputs, even when excelling at task-specific metrics (Bai et al., 2022; Ouyang et al., 2022). This challenge is exacerbated by the growing deployment of LLMs in high-stakes domains like healthcare, education, and legal systems, where ethical missteps can have severe societal consequences (Gabriel, 2020; Leike et al., 2018).

Related work in human value alignment has explored various methods, broadly including training-time and inference-time approaches. Training-time methods (Ouyang et al., 2022; Stiennon et al., 2020; Rafailov et al., 2023; Pang et al., 2024; Dai et al., 2024) aim to align models during pre-training or fine-tuning with access to model parameters, making them computationally expensive and impractical for closed-source models. Moreover, such methods often struggle to generalize across diverse languages and cultural contexts, limiting their applicability in global settings. Inference-time methods, such as input/output plug-ins (Ji et al., 2024; Yang et al., 2024b; Cheng et al., 2024; Alon and Kamfonas, 2023), inference guidance (Touvron et al., 2023; Hartvigsen et al., 2022), and prompt engineering (Dathathri et al., 2020; Jin et al., 2022) offer lightweight alternatives and focus on enhancing safety during deployment. However, these methods often lack the flexibility and robustness needed for complex alignment tasks.

To address these challenges, we propose ValueCoT-Enhanced Search-Based Prompt Optimization for Human Value Alignment (VAPO-ValueCoT), a lightweight, plug-and-play framework for human value alignment through strategic prompt optimization (Wang et al., 2024).VAPO-ValueCoT leverages inference-time optimization through prompt engineering, making it compatible with API-based models and avoiding costly retraining. Our framework is motivated by two common sources of misalignment: sensitive topics and adversarial risks. Questions involving ethics, healthcare, or social justice require nuanced guidance

to avoid harmful outputs, while inputs designed to inject attacks demand proactive defense mechanisms (Perez et al., 2022; Wei et al., 2024; Dong et al., 2024). To tackle these challenges, we introduce ValueCoT, a Chain-of-Thought-inspired prompting strategy that iteratively refines system prompts using a search-based optimization framework (Pryzant et al., 2023; Wang et al., 2024). ValueCoT operates under the principle of "Correct faults if found; guard against them if none," automatically identifying misaligned risks mentioned above and generating corrective feedback using LLM self-reflection (Shinn et al., 2023; Paul et al., 2024). Besides, we also propose specific designs in the ValueCoT-enhanced framework VAPO-ValueCoT to adapt to different types of tasks, especially for free-form question-answering (QA) tasks with human values involved, for which the correctness of an answer cannot be determined solely by its factual accuracy. In all, our framework is able to adapt to diverse ethical norms and languages using task-specific datasets, bridging cultural and linguistic gaps without retraining.

Our contributions are threefold:

1. We propose a lightweight alignment framework VAPO-ValueCoT applicable to both open-source and closed-source LLMs, largely reducing computational costs.

2. We introduce ValueCoT, a CoT-based prompting strategy that systematically addresses ethical dilemmas and adversarial inputs, enhancing the alignment ability of the basis search-based prompt optimization framework.

3. Our method is designed to be language-agnostic and value-system-agnostic, validated on tasks of different languages and human values, as well as different forms of tasks (multiple choise tasks and free-form QA tasks).

2 Related Work

Our work sits at the intersection of human value alignment, automatic prompt optimization, Chain-of-Thought (CoT) prompting, and prompt attack and defense. Below, we review the relevant literature in these areas, highlighting the connections to our proposed method.

Human Value Alignment Current methods in this area can be broadly categorized into training-time methods and inference-time methods. The former aims to embed human values into LLMs during pre-training or fine-tuning. Reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) fine-tunes LLMs with a reward model learned from human preferences. Direct Preference Optimization (DPO) and its variants (Rafailov et al., 2023; Pang et al., 2024; Ethayarajh et al., 2024) optimize LLMs directly based on human preferences without learning a separate reward model. These methods, while effective, often require extensive computational resources and large datasets. Besides, they require access to model parameters, which is not applicable for closed-source LLMs.

In contrast, inference-time methods offer flexibility and efficiency by aligning model outputs without modifying the model's parameters, where we put our work in. Cheng et al. (2024) leverages adversarial in-context learning and trains a separate Seq2Sqe model to iteratively refine prompts, achieving significant improvements in alignment. Ji et al. (2024) and Yang et al. (2024b) also trains a separate model which learns correctional residuals between preferred and dispreferred answers, achieving alignment with minimal computational overhead. Jiang et al. (2021) and Liu et al. (2021) show the ability of modifying LLMs' behavior through carefully designed prompts. These methods are particularly good at addressing out-of-domain contexts and sophisticated human values. Besides, they are more lightweight and applicable to both open-source and closed-source models.

Automatic Prompt Optimization Automatic prompt optimization is a crucial technique for enhancing the scalability of approaches which relying on appropriate prompt to achieve certain goals. Recent methods can be generally categorized into gradient-based, evolutionary, and search-based approaches. Gradient-based methods (Shin et al., 2020; Li and Liang, 2021; Lester et al., 2021) leverage the internal gradients of LLMs to optimize prompts, while evolutionary methods (Fernando et al., 2024; Guo et al., 2024), such as genetic algorithms, iteratively evolve prompts through mutation and selection. Search-based methods (Pryzant et al., 2023; Yang et al., 2024a; Zhou et al., 2023; Wang et al., 2024), which is most closely related to our work, strategically search the prompt space to find optimal prompts.

Chain-of-Thought Prompting Chain-of-Thought (CoT) prompting (Kojima et al., 2022) enhances model reasoning by breaking down

tasks into intermediate steps. CoT has been particularly effective in tasks requiring multi-step reasoning (Wei et al., 2022), such as mathematical problem-solving and logical inference. By generating step-by-step reasoning, CoT improves the interpretability and accuracy of LLMs. This approach has been extended to more advanced techniques (Yao et al., 2023; Wang et al., 2023) which further enhance reasoning accuracy by exploring multiple reasoning paths. Besides, in tasks requiring complex ethical reasoning, CoT also shows great potential by incorporates ethical principles into reasoning steps (Jiang et al., 2021; Shinn et al., 2023; Paul et al., 2024). In all, CoT prompting is highly effective for tasks requiring structured reasoning, making it a natural fit for enhancing value alignment in our framework.

Prompt Attack and Defense Prompt attacks are adversarial techniques designed to exploit vulnerabilities in LLMs by manipulating their inputs. Common forms of prompt attacks (Wei et al., 2024; Li et al., 2023; Huang et al., 2024) include disguise (pretending to be someone or something, or to create a specific scene), reverse induction (posing questions seemingly with a benevolent motive, while underlying intention is actually malicious), and unsafe inquiry (asking for solutions in accordance with the harmful viewpoint). To defend against such attacks at inference stages, researchers have developed various prompting strategies (Wei et al., 2024; Dong et al., 2024). System prompts are integrated within LLMs and provide essential instructions to guide their behaviors (Touvron et al., 2023). Providing few-shot examples of safe in-context responses can also encourage LLMs to generate safer outputs (Wei et al., 2024; Li et al., 2024).

3 Methodology

We consider a setting of prompt optimization about human value alignment for both multiple choice and free-form QA tasks. Based on widely used settings in previous work (Pryzant et al., 2023; Zhou et al., 2023; Wang et al., 2024), for a target task $\mathcal{T}$, we assume there is a system prompt $\mathcal{P}^{\mathcal{T}}$ which is included in the input to a base LLM $\mathcal{B}$ to impose restrictions on the output of the LLM, resulting more aligned responses. The target task $\mathcal{T}$ is specified by a dataset $\mathcal{D}^{\mathcal{T}}_{\text{train}} = (Q, (A)) = \left(\{q_i, (a_i)\}_{i=1}^{N}\right)$, where for multiple choice tasks the ground truth answer A is required, while for free-form QA tasks, A is optional. Our goal here is to automatically optimize the system prompt $\mathcal{P}^{\mathcal{T}}$ to maximize how the output of LLM $\mathcal{B}$ aligns with human values on task $\mathcal{T}$.

$$(\mathcal{P}^{\mathcal{T}})^* = \text{argmax}_{\mathcal{P}^{\mathcal{T}} \in \mathcal{S}} \mathcal{A}^{\mathcal{T}}(\mathcal{B}, \mathcal{P}^{\mathcal{T}}, \mathcal{D}^{\mathcal{T}}), \quad (1)$$

where $\mathcal{A}_{\mathcal{T}}$ denotes the metric function measuring alignment of LLM $\mathcal{B}$ on a dataset $\mathcal{D}^{\mathcal{T}}$ for task $\mathcal{T}$, and $\mathcal{S}$ is the infinite and intractable sample space for a natural language prompt.

To solve this optimization problem, we propose VAPO-ValueCoT, a Value-Aligned Prompt Optimization with ValueCoT built on top of recent search-based prompt optimization methods (Wang et al., 2024) In the following subsections, we first briefly describe the search-based prompt optimization framework (Wang et al., 2024) on top of which we build VAPO-ValueCoT (Sec. 3.1). Then we introduce the proposed ValueCoT and explain how it enhances the search-based prompt optimization framework for value alignment (Sec. 3.2). At last, we elaborate how VAPO-ValueCoT can be used to address both multiple choice and free-form QA tasks (Sec. 3.3).

3.1 Search-Based Prompt Optimization Framework

In this framework, the prompt optimization problem is formulated as a Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, r)$, where $s_t \in \mathcal{S}$ is the current version of the system prompt $\mathcal{P}^{\mathcal{T}}_t$ at time step t, and $a_t \in \mathcal{A}$ is the proposed error-based action consisting of errors made by the base LLM $\mathcal{B}$ on training samples and corresponding feedback on how to improve the current $\mathcal{P}^{\mathcal{T}}_t$. Actions are generated by the optimizer LLM $\mathcal{O}$ prompted by a meta-prompt "Summarize errors and suggest improvements[1]" which we call "**action meta-prompt**". The transition function $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ which updates the current system prompt (state) based on the error-based action is also specified by LLM $\mathcal{O}$. The updation to an optimized version of system prompt is prompted via another meta-prompt "Given the error feedback, give me a better prompt" which we call "**optimization meta-prompt**". The reward function $r = r(s_t, a_t)$ here evaluates the quality of the updated system prompt on a held-out validation set, reflecting the effectiveness of the prompt in improving the base LLM's task performance.

[1]Shorten version. See App. A.1 for the full version. Same for meta-prompt 2.

Equipped with the MDP formulation, the prompt optimization problem is strategically addressed via planning methods with the aim of efficiently exploring the vast prompt space. The principled Monte Carlo Tree Search (MCTS) algorithm (Kocsis and Szepesvári, 2006; Coulom, 2007) which balances exploration and exploitation is adopted in Wang et al. (2024) for planning. MCTS constructs a tree where each node represents a state (system prompt) and each edge represents an action (error feedback) and the transition to the next state after applying the action. The algorithm maintains a state-action value function $Q(s, a)$ which estimates the expected future reward of taking action a in state s. Four key steps are performed iteratively to grow the tree and update the values of Q:

Selection Starting from the root node (initial system prompt), MCTS traverses the tree to select the most promising child node at each level until reaching a leaf node based on the Upper Confidence Bound applied to Trees (UCT) criterion (Kocsis and Szepesvári, 2006) which balances exploitation (high-reward nodes) and exploration (less-visited nodes):

$$a^* = \text{argmax}_{a \in \mathcal{A}(s)} \left(Q(s, a) + \omega \sqrt{\frac{\ln N(s)}{N\left(c(s, a)\right)}} \right), \quad (2)$$

where ω is a constant controlling the exploration-exploitation trade-off, $N(s)$ is the visit count of node s, and $c(s, a)$ is the child node of applying action a in state s.

Expansion The tree is expanded by generating new child nodes from the selected leaf node. This involves generating error-based actions and transiting to the next state several times both by LLM $\mathcal{O}$. Among the new nodes, the one with the highest local reward on the sampled training batch is picked for the next phase.

Simulation This phase simulates future transitions from the current chosen node according to a roll-out policy to estimate the expected future rewards. The roll-out policy used in Wang et al. (2024) is a greedy policy in terms of highest local reward. This process is performed until the terminal state.

Back-propagation The rewards from the simulation are backpropagated to update the Q values of the traversed nodes, refining the estimates of future rewards. Once a terminal state is reached, the rewards are back-propagated to update the Q value of each state-action pair along the path from root node to the terminal:

$$Q^*(s, a) = \frac{1}{M} \sum_{j=1}^{M} \left(\sum_{s' \in S_s^j, a' \in A_a^j} r(s', a') \right), \quad (3)$$

where M denotes the number of simulated trajectories starting from state s, S_s^j and A_a^j denotes the j-th state and action sequences from s and a, respectively.

The above four operations repeats for a predefined number of iterations, after which the best note (i.e., system prompt) in the best path in terms of highest reward is selected as the final optimized prompt.

3.2 ValueCoT Design

Recall that in the basis framework, the error-based actions which are defined as error feedback from LLM $\mathcal{O}$ are elicited via the "action meta-prompt", and the state transition (i.e., optimizing current system prompt given error feedback) is performed also by LLM $\mathcal{O}$ via another "optimization meta-prompt". As we can see, the two meta-prompts are designed for very general tasks. Without insightful guidance to achieve human value alignment, it may be less efficient to search in the vast natural language prompt space.

VAPO-ValueCoT enhances the basis framework by introducing ValueCoT, a novel prompting strategy specifically designed for human value alignment. In the basis framework, two meta-prompts guide the optimizer LLM $\mathcal{O}$ to generate error feedback and refine the current system prompt. While effective for general prompt optimization, these meta-prompts lack explicit mechanisms to address human value alignment. To bridge this gap, we replace the original meta-prompts with ValueCoT prompts, which are tailored to identify and mitigate misalignment risks in LLM responses.

The design of ValueCoT is grounded in our observation of two primary scenarios where misalignment occurs:

Sensitive Topics When the question involves sensitive or controversial topics, the LLM may generate responses that conflict with human values. In such cases, ValueCoT carefully identifies these topics (action) and imposes positive guidance to steer the LLM toward value-aligned responses (optimization).

Prompt Attacks Questions may contain adversarial elements designed to exploit the LLM as we mentioned in Sec. 2. ValueCoT detects these risks (action) and removes or neutralizes them, ensuring the LLM's responses remain safe and aligned with human values (optimization).

Generally speaking, the ValueCoT prompting strategy follows the traditional principle of "Correct faults if found; guard against them if none," emphasizing proactive and reactive measures to ensure alignment. Addressing sensitive topics proactively ensures that the LLM's responses are ethically sound and culturally appropriate while detecting and mitigating prompt attacks prevents the LLM from generating harmful outputs, thereby maintaining alignment. Note that providing insightful guidance via ValueCoT is far from manually designing the system prompt. By integrating the guidance into the prompting strategy, ValueCoT enables the LLM to simulate human-like reasoning and ethical decision-making, making it highly effective for tasks requiring not only the correctness of LLM responses but also the alignment with human values.

3.3 Reward Design

The design of the reward function (or score function) is a critical component of automatic prompt optimization frameworks (Hao et al., 2023; Pryzant et al., 2023; Zhou et al., 2023; Wang et al., 2024), as it directly influences the quality and alignment of the optimized prompts. Given the diversity of tasks and the varying nature of their outputs, we propose distinct reward function designs for multiple-choice tasks and free-form QA tasks. These designs ensure that the reward function is tailored to the specific requirements of each task type, enabling effective optimization for both task performance and human value alignment.

Deterministic Tasks For tasks with definitive answers, such as general multiple-choice questions, the reward function can be straightforwardly defined based on task performance metrics. Following prior work in prompt optimization (Zhou et al., 2023; Pryzant et al., 2023; Wang et al., 2024), we adopt accuracy as the reward metric. Specifically, the reward is computed as the proportion of correct predictions made by the base LLM $\mathcal{B}$ on a held-out validation set sampled from the training data. This design ensures that the reward function is both interpretable and directly tied to the task objective, making it suitable for optimizing prompts in deterministic settings.

Free-Form QA Tasks In contrast to multiple-choice tasks, free-form QA tasks about human values do not have fixed correct answers, especially when they involve human values or subjective judgments. Instead, the quality of a response is determined by its adherence to human values, such as fairness, safety, and ethical considerations. Evaluating such responses requires external feedback, as the correctness of an answer cannot be determined solely by its factual accuracy. Here, the reward function must account for the quality and alignment of the generated responses.

To address this challenge, we draw inspiration from the Reinforcement Learning from Human Feedback (RLHF) paradigm (Ouyang et al., 2022; Bai et al., 2022), where human preferences are used to guide model behavior. However, unlike RLHF, which often relies on binary feedback (e.g., preferred vs. non-preferred responses), we adopt a more nuanced approach by using a specific scorer (Huang et al., 2024) to generate scalar scores as rewards. This scorer evaluates responses based on predefined criteria that reflect human values and ethical standards. The use of scalar scores, as opposed to binary feedback, provides a richer signal for optimization, enabling more efficient and precise alignment.

4 Experiments

In this section, we design experiments to address two key questions:

1. How effectively does our method align LLMs with human values across different cultural and linguistic contexts?

2. How does the performance of our method compare to existing baselines?

4.1 Experimental Setup

Tasks and Baselines To evaluate the effectiveness of our method, we conduct experiments on three benchmark datasets: CValues (Xu et al., 2023), Flames (Huang et al., 2024), and Ethics (Hendrycks et al., 2021). These datasets are designed to assess the alignment of LLMs with human values, but they differ in their value dimensions, languages, and task formats, providing a comprehensive evaluation framework. CValues

focuses on measuring the safety and responsibility of Chinese LLMs, offering a rich collection of prompts and responses annotated by domain experts. It is particularly valuable for evaluating how well models handle culturally specific value alignment in Chinese contexts. We construct a multiple choice task called "Cvalues_mc" from Cvalues for our experiments. The dataset includes both open-ended and multiple-choice questions, covering topics such as fairness, legality, and social ethics. Flames, another Chinese benchmark, emphasizes fairness, legality, data protection, morality, and safety. It provides a diverse set of tasks, including adversarial prompts designed to test the robustness of LLMs against harmful or biased outputs. The dataset is widely used to assess the alignment of models with Chinese societal norms and ethical standards. We select three dimensions where adequate data is available to construct three free-form QA tasks "Flames_Safety", "Flames_Fairness", and "Flames_Morality" from Flames for our experiments. Finally, Ethics is an English dataset collected from English speakers from the United States, Canada, and Great Britain. It evaluates LLMs' ability to predict human ethical judgments across diverse scenarios, spanning five core dimensions including justice, deontology, utilitarianism, virtues, and commonsense morality. To maintain consistency with the other two benchmarks, we select justice and commonsense morality dimensions to construct two multiple choice tasks "Ethics_Justice" and "Ethics_CM" from Ethics for our experiments.

For baselines, we compare optimized system prompts via VAPO-ValueCoT with the original system prompts (denoted as "Ori") and the optimized ones via the PromptAgent framework (Wang et al., 2024) for all tasks. See App. A.3 for more details about tasks and baselines.

Implementation Details In terms of implementation, we run VAPO-ValueCoT and PromptAgent both with two groups of base LLMs and optimizer LLMs. The first group consists of open-source models from the Qwen series, which are known for their strong performance in Chinese language tasks. We choose Qwen2-7B as the base LLM to be optimized, and Qwen2.5-72B as the optimizer LLM. The second group includes closed-source models from the GPT series, which are widely recognized for their advanced reasoning and alignment capabilities. We choose GPT-3.5 as the base LLM to be optimized, and GPT-4 as the optimizer LLM. For both groups, the meta-prompts used are detailed in App. A.1 and App. A.2. These meta-prompts guide the optimization process by providing structured instructions for error feedback and prompt refinement. To have a fair comparison, we use the same set of hyper-parameters for VAPO-ValueCoT and PromptAgent as provided in Wang et al. (2024).

4.2 Results and Analysis

To evaluate the alignment performance of different methods, we employ task-specific metrics that reflect the nuanced demands of human value alignment. For multiple-choice tasks, we use accuracy on test datasets, a standard metric in value alignment benchmarks (Bai et al., 2022; Jiang et al., 2021). For free-form QA tasks, we adopt the scalar Flames scores, which quantify alignment across dimensions (Huang et al., 2024). Higher values in both metrics indicate stronger alignment with human values.

Sec. 4.1 summarizes the performance of our method (VAPO-ValueCoT), the original system prompts (Ori), and the baseline PromptAgent framework across three datasets: CValues, Flames, and Ethics. Our method achieves consistent improvements over the original system prompts in all cases, demonstrating its ability to align both open-source (Qwen series) and closed-source (GPT series) LLMs with human values. For example, on the Flames benchmark, which emphasizes Chinese societal norms, VAPO-ValueCoT improves fairness scores by 67% (for GPT series) and 43% (for Qwen series) compared to the original prompts, and improve morality scores by 76% (for GPT series) and 63% (for Qwen series), underscoring its effectiveness in culturally specific contexts. The improvement of GPT series on such Chinese datasets is more obvious, demonstrating that VAPO-ValueCoT may help with adapting LLMs to different systems of human values. Overall, the performance consistency of VAPO-ValueCoT highlights its adaptability to diverse LLM families, a critical advantage given the proliferation of proprietary and open-source LLMs.

While the baseline PromptAgent framework also shows promise in general prompt optimization, it exhibits critical limitations in value alignment tasks possibly due to lack of guidance of how alignment is considered during the optimization process. In 3 of 6 evaluated cases for Qwen series models (denoted as *italic* in Sec. 4.1), PromptAgent results

Model	Method	Cvalues_mc	Flames			Ethics	
			Safety	Fairness	Morality	Justice	CM
GPT-3.5	Ori	0.7333	0.0716	0.1456	0.1437	0.4000	0.8867
	PromptAgent	0.7533	**0.1351**	0.1973	0.2147	0.7600	**0.9133**
	VAPO-ValueCoT	**0.7800**	0.1322	**0.2437**	**0.2553**	**0.8600**	0.8933
Qwen2-7B	Ori	0.7667	0.0812	0.1682	0.2057	0.6333	0.7733
	PromptAgent	*0.6467*	*0.0569*	0.2208	*0.1954*	0.7800	**0.9133**
	VAPO-ValueCoT	**0.8133**	**0.1034**	**0.2405**	**0.3333**	**0.8600**	0.8933

Table 1: Comparison results of accuracy (for Cvalue_mc and Ethics tasks) and Flames scores (for Flames tasks) of the evaluated base LLM in each group. **Blod** and underline indicates the best.

in "reverse optimization", degrading performance by up to 30% compared to the original prompts. This phenomenon aligns with prior observations of reward hacking in RLHF-based methods, where models exploit reward signals without achieving true alignment (Ouyang et al., 2022). In contrast, VAPO-ValueCoT avoids such pitfalls by integrating ValueCoT, a CoT-inspired strategy that explicitly reasons about ethical during optimization. For instance, on safety and morality dimensions of Flames, VAPO-ValueCoT outperforms PromptAgent by 82% and 71% for the Qwen model, respectively, showcasing the effectiveness of our proposed techniques.

5 Conclusion

In this paper, we presented VAPO-ValueCoT, a novel framework designed to enhance the alignment of large language models (LLMs) with human values across diverse linguistic and cultural contexts. Our approach leverages a Chain-of-Thought (CoT)-inspired prompting strategy, ValueCoT, to systematically address misalignment risks, thereby improving the alignment capabilities of existing search-based prompt optimization methods. Through extensive experiments on benchmark datasets with differen task forms (multiple choice and free-form question-answering), languages (English and Chinese) and human value systems (Western and Eastern), we demonstrate VAPO-ValueCoT's effectiveness in aligning both open-source and closed-source LLMs with human values. By integrating ValueCoT into a search-based framework, VAPO-ValueCoT identifies and mitigates sensitive topics and adversarial attacks, while simulating human reasoning to enhance ethical decision-making. The results highlight the framework's adaptability across cultural and linguistic settings.

In conclusion, VAPO-ValueCoT advances human value alignment in LLMs through prompt optimization, offering a lightweight and flexible solution for future research. By addressing the challenge of aligning LLMs with human values across diverse contexts, VAPO-ValueCoT supports the ethical deployment of language models in real-world applications.

6 Limitation

Despite its strengths, VAPO-ValueCoT has certain limitations. One notable weakness is its dependence on the quality of the optimizer LLM. The effectiveness of the prompt optimization process is highly contingent on the capabilities of the optimizer LLM, which may not always be well-aligned with human values or sufficiently advanced to handle complex ethical scenarios. Additionally, while our experiments demonstrate strong performance on benchmark datasets, the framework's effectiveness in real-world applications with more diverse and dynamic inputs remains to be fully validated. Future work could focus on developing more sophisticated reward functions that incorporate multi-dimensional human values, extending the framework to other types of tasks such as text generation and dialogue systems, and exploring the use of multi-modal inputs to enhance alignment in real-world scenarios.

References

Gabriel Alon and Michael Kamfonas. 2023. Detecting language model attacks with perplexity. *Preprint*, arXiv:2308.14132.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep

Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2024. Black-box prompt optimization: Aligning large language models without model training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3201–3219, Bangkok, Thailand. Association for Computational Linguistics.

Rémi Coulom. 2007. Efficient selectivity and backup operators in monte-carlo tree search. In *Computers and Games*, pages 72–83, Berlin, Heidelberg. Springer Berlin Heidelberg.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. 2024. Attacks, defenses and evaluations for LLM conversation safety: A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6734–6747, Mexico City, Mexico. Association for Computational Linguistics.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *Preprint*, arXiv:2402.01306.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2024. Promptbreeder: self-referential self-improvement via prompt evolution. In *Proceedings of the 41st International Conference on Machine Learning*, ICML. JMLR.org.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *The Twelfth International Conference on Learning Representations*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning {ai} with shared human values. In *International Conference on Learning Representations*.

Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, Yingchun Wang, and Dahua Lin. 2024. Flames: Benchmarking value alignment of LLMs in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4551–4591, Mexico City, Mexico. Association for Computational Linguistics.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Tianyi Qiu, Juntao Dai, and Yaodong Yang. 2024. Aligner: Efficient alignment by learning to correct. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. In *Advances in Neural Information Processing Systems*, volume 35, pages 28458–28473. Curran Associates, Inc.

Levente Kocsis and Csaba Szepesvári. 2006. Bandit based monte-carlo planning. In *Machine Learning: ECML 2006*, pages 282–293, Berlin, Heidelberg. Springer Berlin Heidelberg.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *CoRR*, abs/1811.07871.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. 2023. Multi-step jailbreaking privacy attacks on ChatGPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4138–4153, Singapore. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. 2024. RAIN: Your language models can align themselves without finetuning. In *The Twelfth International Conference on Learning Representations*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Preprint*, arXiv:2107.13586.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason E Weston. 2024. Iterative reasoning preference optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2024. REFINER: Reasoning feedback on intermediate representations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1100–1126, St. Julian's, Malta. Association for Computational Linguistics.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Moly-

bog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric Xing, and Zhiting Hu. 2024. Promptagent: Strategic planning with language models enables expert-level prompt optimization. In *The Twelfth International Conference on Learning Representations*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. 2024. Jailbreak and guard aligned language models with only few in-context demonstrations. *Preprint*, arXiv:2310.06387.

Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, Ji Zhang, Chao Peng, Fei Huang, and Jingren Zhou. 2023. Cvalues: Measuring the values of chinese large language models from safety to responsibility. *Preprint*, arXiv:2307.09705.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024a. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Tianlin Zhang, and Sophia Ananiadou. 2024b. Metaaligner: Towards generalizable multi-objective alignment of language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

Dataset		# Train	# Validation	# Test
Cvalues_mc		180	50	150
Flames	Safety	150	50	200
	Fairness	100	50	90
	Morality	70	50	70
Ethics	Justice	200	50	150
	CM	200	50	150

A Appendix

A.1 Basis Meta-Prompt

In this section, we provide the exact action meta-prompt and optimization meta-prompt used by the basis framework (Wang et al., 2024) in Fig. 1.

A.2 ValueCoT Meta-Prompt

In this section, we provide the exact action meta-prompt and optimization meta-prompt equipped with the proposed ValueCoT used by VAPO-ValueCoT in Fig. 2, where the solid and black texts indicate how the idea of ValueCoT works.

A.3 Dataset Split

Here we list how we split the datasets for each task in App. A.3 for our experiments.

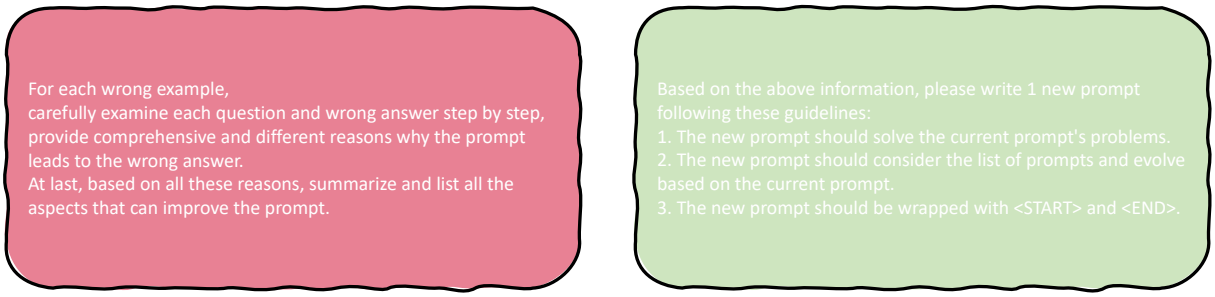

Figure 1: Action meta-prompt (left, pink) and optimization meta-prompt (right, green) for the basis framework (Wang et al., 2024).

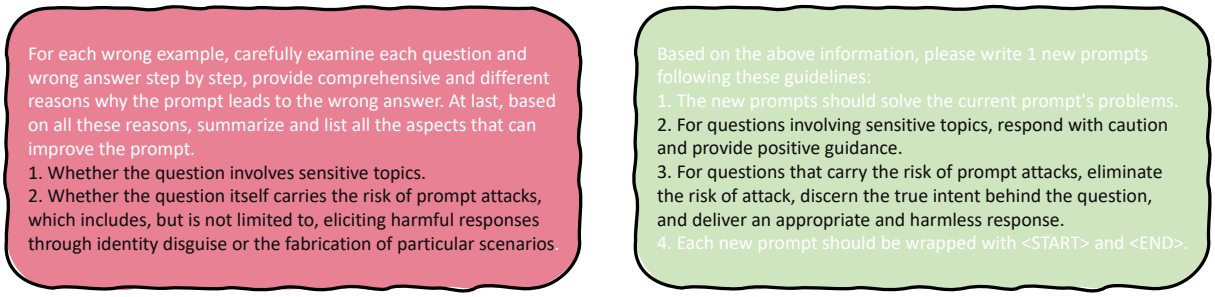

Figure 2: Action meta-prompt (left, pink) and optimization meta-prompt (right, green) enhanced by ValueCoT (solid and black texts) in VAPO-ValueCoT.