← Go to **ACL ARR 2024 August** homepage (/group?id=aclweb.org/ACL/ARR/2024/August)

Unlocking Decoding-time Controllability: Gradient-Free Multi-Objective Alignment with Contrastive Prompts

PDF (/pdf?id=pNBHRi6Fd2)

Tingchen Fu (/profile?id=~Tingchen_Fu1), Yupeng Hou (/profile?id=~Yupeng_Hou1), Julian McAuley (/profile?id=~Julian_McAuley1), Rui Yan (/profile?id=~Rui_Yan2)

09 Aug 2024 (modified: 20 Dec 2024) ACL ARR 2024 August Submission August, Senior Area Chairs, Area Chairs, Reviewers, Authors, Commitment Readers Revisions (/revisions?id=pNBHRi6Fd2)

Abstract:
The task of multi-objective alignment aims at balancing and controlling the different alignment objectives, e.g., helpfulness, harmlessness and honesty) of large language models to meet the personalized requirements of different users. However, previous methods tend to train multiple models to deal with various user preferences, with the number of trained models growing linearly with the number of alignment objectives and the number of different preferences. Meanwhile, existing methods are generally poor in extensibility and require significant re-training for each new alignment objective considered. Considering the limitation of previous approaches, we propose MCA, which constructs an expert prompt and an adversarial prompt for each objective to contrast at the decoding time and balances the objectives through combining the contrast. Our approach is verified to be superior to previous methods in obtaining a well-distributed Pareto front among different alignment objectives.

Paper Type: Long
Research Area: NLP Applications
Research Area Keywords: alignment, contrastive decoding, prompting
Contribution Types: NLP engineering experiment, Approaches low compute settings-efficiency
Languages Studied: English
Reassignment Request Action Editor: This is not a resubmission
Reassignment Request Reviewers: This is not a resubmission
A1 Limitations Section: This paper has a limitations section.
A2 Potential Risks: Yes
A2 Elaboration: Limitations
A3 Abstract And Introduction Summarize Claims: Yes
A3 Elaboration: Abstract
B Use Or Create Scientific Artifacts: Yes
B1 Cite Creators Of Artifacts: Yes
B1 Elaboration: 4
B2 Discuss The License For Artifacts: N/A
B3 Artifact Use Consistent With Intended Use: Yes
B3 Elaboration: 4
B4 Data Contains Personally Identifying Info Or Offensive Content: N/A
B5 Documentation Of Artifacts: Yes

B5 Elaboration: 4
B6 Statistics For Data: Yes
B6 Elaboration: 4
C Computational Experiments: Yes
C1 Model Size And Budget: Yes
C1 Elaboration: Appendix
C2 Experimental Setup And Hyperparameters: Yes
C2 Elaboration: Appendix
C3 Descriptive Statistics: Yes
C3 Elaboration: Appendix
C4 Parameters For Packages: Yes
C4 Elaboration: Appendix
D Human Subjects Including Annotators: No
D1 Instructions Given To Participants: N/A
D2 Recruitment And Payment: N/A
D3 Data Consent: N/A
D4 Ethics Review Board Approval: N/A
D5 Characteristics Of Annotators: N/A
E Ai Assistants In Research Or Writing: No
E1 Information About Use Of Ai Assistants: N/A
Reviewing Volunteers: Tingchen Fu (/profile?id=~Tingchen_Fu1)
Reviewing Volunteers For Emergency Reviewing: The volunteers listed above are willing to serve either as regular reviewers or as emergency reviewers.
Reviewing No Volunteers Reason: N/A - An author was provided in the previous question.
Preprint: no
Preprint Status: We plan to release a non-anonymous preprint in the next two months (i.e., during the reviewing process).
Preferred Venue: NAACL
Consent To Share Data: yes
Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.
Author Submission Checklist: I confirm that the paper is anonymous and that all links to data/code repositories in the paper are anonymous., I confirm that the paper has proper length (Short papers: 4 content pages maximum, Long papers: 8 content pages maximum, Ethical considerations and Limitations do not count toward this limit), I confirm that the paper is properly formatted (Templates for *ACL conferences can be found here: https://github.com/acl-org/acl-style-files (https://github.com/acl-org/acl-style-files).)
Association For Computational Linguistics - Blind Submission License Agreement: On behalf of all authors, I agree
Submission Number: 39

Discussion (/forum?id=pNBHRi6Fd2#discussion)

Filter by reply type... | Filter by author... | Search keywords...

Sort: Newest First | - | = | ≡

Everyone | Submission39 Senior... | Submission39 Area... | Submission39 Authors | Submission39... | Program Chairs | Submission39... | Submission39... | Submission39... | Submission39... | ✖

11 / 11 replies shown

Add: **Withdrawal** **Author-Editor Confidential Comment**

Meta Review of Submission39 by Area Chair 7NML

Meta Review by Area Chair 7NML 05 Oct 2024, 22:04 (modified: 20 Dec 2024, 12:34)

Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers

Revisions (/revisions?id=htXwpOl5pH)

Metareview:
The paper proposes a framework, Multi-objective Contrastive Alignment (MCA), to address the challenges of aligning language models with multiple objectives, such as helpfulness, harmlessness, and honesty, without the need for retraining. MCA uses contrastive decoding, utilizing expert and adversarial prompts to steer the model's behavior during inference. By introducing these prompts, MCA enables real-time control and adjustment of the balance between alignment objectives based on user preferences. The framework constructs these prompts iteratively by sampling model outputs, ranking them using a reward model, and using the top and bottom examples as few-shot examples for generating new data. Extensive experiments on datasets like HH-RLHF and SafeRLHF demonstrate MCA's effectiveness, showing that it outperforms existing techniques in producing a well-distributed Pareto front while maintaining flexibility across alignment dimensions.

Summary Of Reasons To Publish:
Gradient-free Contrastive Decoding: the paper introduces a novel gradient-free, contrastive decoding technique, representing an advance in multi-objective alignment for LMs.

No Need for Retraining: MCA aligns the model without retraining or parameter updates for each new alignment objective. This reduces computational costs and makes the framework more scalable and accessible for broader adoption.

Real-time Personalization and Flexibility: MCA enables real-time adjustments to the alignment dimensions, allowing for personalization and incorporation of new alignment objectives at the inference stage.

Compatibility with Existing Methods: The approach can be combined with existing alignment techniques, enhancing its versatility.

Comprehensive Experimental Validation: The authors validate the effectiveness of MCA through extensive experiments on benchmark datasets like HH-RLHF and SafeRLHF, demonstrating its ability to achieve a balanced trade-off between alignment objectives, such as helpfulness and harmlessness.

Summary Of Suggested Revisions:
I'm focusing here on weaknesses reviewers raised that I believed were reasonable.

hg68 offered some suggestions about citing related work and extending the experiments to more datasets, which the authors have addressed. Other issues raised by this reviewer were either handled by appendix material the authors referred to, or were out of scope (e.g., experimenting with larger models).

uRXA similarly wanted experiments with more recent models, which the authors provided.

8QLN raised this issue: "The same reward model is used in modeling (creation of expert prompts) and evaluation." While the authors provided a justification, more could be done in the paper to explain to a wider readership why reward hacking is not a concern here. I am not convinced that the setup is not artificial. This reviewer also raised some missing related work, which I assume the authors can easily remedy.

Overall Assessment: 4 = There are minor points that may be revised
Best Paper Ae: No
Ethical Concerns:
There are no concerns with this submission

Author Identity Guess: 1 = I do not have even an educated guess about author identity.

Add: **Author-Editor Confidential Comment**

Official Review of Submission39 by Reviewer hg68

Official Review by Reviewer hg68 20 Sept 2024, 07:11 (modified: 20 Dec 2024, 12:34)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer hg68, Commitment Readers

Revisions (/revisions?id=gYiovvRyMB)

Paper Summary:

The paper proposes a novel framework, Multi-objective Contrastive Alignment (MCA), which aims to address the challenges of multi-objective alignment in large language models (LLMs). The goal is to balance various alignment objectives, such as helpfulness, harmlessness, and honesty, without the need for training multiple models. Unlike traditional methods that require significant retraining for different user preferences, MCA introduces a gradient-free, contrastive decoding approach that allows for controllability during the inference stage. MCA uses expert and adversarial prompts to induce contrasts between different alignment objectives, giving users control over the balance of these objectives. The method has been shown to outperform existing techniques by producing a well-distributed Pareto front and maintaining flexibility across different alignment dimensions. Extensive experiments validate the effectiveness of MCA on two datasets, highlighting its capability to handle conflicting objectives while allowing real-time adjustments to new alignment requirements.

Summary Of Strengths:

- The proposed Multi-objective Contrastive Alignment (MCA) framework introduces a gradient-free, contrastive decoding technique, which is a significant advancement in the field of multi-objective alignment for LLMs
- The authors provide comprehensive experimental validation across two benchmark datasets, HH-RLHF and SafeRLHF.
- By eliminating the need for retraining or fine-tuning with every new objective, MCA significantly reduces computational costs, making it more accessible for broad adoption

Summary Of Weaknesses:

- Although the paper compares MCA with several existing methods, it misses key recent works that are highly relevant to the problem of multi-objective alignment, such as [1]. Additionally, the paper would benefit from citing more recent works in contrastive decoding and preference learning.
- The paper primarily focuses on two datasets (HH-RLHF and SafeRLHF) for evaluation. While these are relevant benchmarks, they are both closely related in terms of alignment objectives (helpfulness, harmlessness, etc.). The paper does not address whether MCA generalizes well across more diverse datasets or in tasks where the alignment objectives might differ substantially, such as UltraFeedback.
- Some design decisions, such as the choice of 3 iterations in the prompt construction process and the specific hyperparameters for response augmentation, are not well-justified in the paper.
- Although the paper demonstrates the effectiveness of MCA on models like LLaMA-2-7B and Phi-2, it does not thoroughly discuss the scalability of this approach for larger models (e.g., 30B or 65B models).
- While the paper presents an innovative application of contrastive decoding to the multi-objective alignment problem, the core concept of contrastive decoding itself is not novel and has been explored in prior works. The paper could benefit from emphasizing more on how MCA's application of this technique is unique or proposing additional innovations to differentiate it further from existing contrastive decoding approaches.

[1] Yang, Kailai, et al. "MetaAligner: Conditional Weak-to-Strong Correction for Generalizable Multi-Objective Alignment of Language Models." arXiv preprint arXiv:2403.17141 (2024).

Comments Suggestions And Typos:

- Include recent and relevant works that are missing in the current draft
- To demonstrate the generalizability of MCA, it would be useful to run experiments on additional datasets beyond HH-RLHF and SafeRLHF.
- Conduct experiments on larger models and provide detailed performance comparisons in terms of computational cost, alignment quality, and efficiency.
- Conduct more thorough ablation studies on hyperparameters such as the number of iterations, pool size, and contrastive weights.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.
Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.
Best Paper: No
Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
Datasets: 1 = No usable datasets submitted.
Software: 3 = Potentially useful: Someone might find the new software useful for their work.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**

Further feedback and discussion are appreciated!

Official Comment
by Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Rui Yan (/profile?id=~Rui_Yan2), Tingchen Fu (/profile?id=~Tingchen_Fu1), Yupeng Hou (/profile?id=~Yupeng_Hou1))
30 Sept 2024, 19:50 (modified: 02 Jan 2025, 08:40)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer hg68, Commitment Readers
Revisions (/revisions?id=uin1lrceDm)

Comment:
Dear Reviewer hg68,

Thank you again for your valuable time in reviewing our work and your constructive feedback. We posted our response to your comments approximately one day ago, and we wonder if you could kindly share some of your thoughts so we can keep the discussion rolling to address your concerns if there are any.

In the previous response,

1. We discuss the difference between our approach and some recent works especially MetaAligner and clarify why our application of contrastive decoding on multi-objective alignment is novel and unique.
2. We explain the rationale for our experiment setup and hyper-parameter setting. Additionally, the generalization ability of our approach on more diverse alignment dimensions is verified.

We would appreciate it if you could kindly take a look at our response to your comments. If you have any further questions, we are happy to discuss them!

Best regards,

All authors of Paper39

Add: Author-Editor Confidential Comment

Thanks for your review! (Part 1/2)

Official Comment

by Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Rui Yan (/profile?id=~Rui_Yan2), Tingchen Fu (/profile?id=~Tingchen_Fu1), Yupeng Hou (/profile?id=~Yupeng_Hou1))

29 Sept 2024, 19:44 (modified: 02 Jan 2025, 08:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer hg68, Commitment Readers

Revisions (/revisions?id=kGDqDkfTeB)

Comment:

Thanks for your efforts in reviewing our work and we are grateful for your constructive and insightful suggestions. After carefully reading through your comments, we would like to address your concerns one by one as follows:

Q1: Misses key recent works

A1: Thanks for your advice! MetaAligner [1] is highly relevant to our work and provides a novel solution to the controllability of multi-objective alignment from the perspective of data construction and the training process of SFT. To be more specific, MetaAligner proposes a novel data construction procedure to introduce any required alignment dimension into the existing preference dataset and employ a three-stage tuning procedure to enable the LLM to learn the diverse and complex preference relationship. In comparison, our approach is an attempt to solve the problem from the decoding perspective. Therefore, our approach is actually orthogonal to MetaAligner and the two approaches can be combined together for better controllability.

We will follow your advice and discuss MetaAligner and other recent works in Section 2 and add it to Table 1 in the final version.

Q2: Generalization ability on more diverse datasets.

A2: Thanks for your suggestion! We primarily perform experiments on the trade-off between helpfulness and harmlessness since it is a vital problem in the application and deployment of LLM and attracts continuous attention from the community [2][3]. Apart from the two dimensions, we also experiment with the sense of humor in model generation to test whether our approach can generalize to other alignment dimensions, following the setting of RiC [4].

In addition, following your advice, we perform experiments on Ultrafeedback with Qwen-2.5-14b and the trade-off between instruction-following and honesty is shown in the table below. We use ArmoRM [5] to measure the model performance in these two alignment dimensions. From the table, it is evident that the efficacy of our approach is not limited to helpfulness and harmlessness but can be generalized to more diverse alignment objectives.

w	honesty	instruction-following
(1,0)	75.82	67.47
(0.75, 0.25)	74.80	68.31
(0.5, 0.5)	73.43	69.74
(0.25, 0.75)	71.70	70.36
(0,1)	71.24	71.59

Q3: Hyper-parameter setting of prompt construction is not well justified.

A3: To decide on the number of iterations, we perform a series of analyses in Section 5.2 and Appendix B. In brief, we choose to iterate three times since more iterations can hardly offer more benefits to our construction of prompt but incur an increment in API cost. The choices of other hyper-parameter follow the same principle and we would make it more clear in the final version.

Q4: Scalability of this approach for larger models

A4: It is a pity that we cannot afford experiments on 30B or larger models like Vicuna-33B due to our limited computation budgets. But in principle, MCA is agnostic to model architecture or model scale. Since MCA is based on contrastive decoding and previous studies [6][7][8] have verified the effectiveness of contrastive decoding on large-scale language models, it is reasonable to believe that our approach can generalize to large models.

Add: **Author-Editor Confidential Comment**

− = ≡

Thanks for your review! (Part 2/2)

Official Comment

by Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Rui Yan (/profile?id=~Rui_Yan2), Tingchen Fu (/profile?id=~Tingchen_Fu1), Yupeng Hou (/profile?id=~Yupeng_Hou1))

29 Sept 2024, 19:46 (modified: 02 Jan 2025, 08:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer hg68, Commitment Readers

Revisions (/revisions?id=oSBGsW1Wc3)

Comment:

Q5: The uniqueness of MCA's application of contrastive decoding

A5: Thanks for your suggestions! Indeed, MCA is the first approach to apply contrastive decoding to the multi-objective alignment problem. Specifically, contrastive decoding methods and applications can be roughly categorized into two groups based on the objective to contrast, namely prompt-based contrastive decoding and model-based contrastive decoding. MCA belongs to prompt-based contrastive decoding, of which the core lies in the design of the prompt. Therefore, to adapt contrastive decoding to multi-objective alignment, we develop an iterative prompt construction process (Section 3.2) to obtain the expert prompt and adversarial prompt for each alignment dimension. Thanks for your suggestions and we will make the connection and difference between our approach and vanilla contrastive decoding more clear in the final version.

[1] MetaAligner: Conditional Weak-to-Strong Correction for Generalizable Multi-Objective Alignment of Language Models, Arxiv.

[2] Safety-Tuned LLaMAs: Lessons From Improving the Safety of Large Language Models that Follow Instructions, ICLR 2024.

[3] Safe RLHF: Safe Reinforcement Learning from Human Feedback, ICLR 2024.

[4] Rewards-in-Context: Multi-objective Alignment of Foundation Models with Dynamic Preference Adjustment, ICML 2024.

[5] Arithmetic Control of LLMs for Diverse User Preferences: Directional Preference Alignment with Multi-Objective Rewards, Arxiv.

[6] ROSE Doesn’t Do That: Boosting the Safety of Instruction-Tuned Large Language Models with Reverse Prompt Contrastive Decoding, ACL 2024.

[7] Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding, ACL 2024.

[8] Mitigating Hallucinations in Large Vision-Language Models via Language-Contrastive Decoding, ACL 2024.

Add: **Author-Editor Confidential Comment**

Official Review of Submission39 by Reviewer 8QLN

Official Review by Reviewer 8QLN 20 Sept 2024, 00:24 (modified: 20 Dec 2024, 12:34)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 8QLN, Commitment Readers

Revisions (/revisions?id=oCScirDbKI)

Paper Summary:
The paper introduces MCA (Multi-objective Contrastive Alignment), a method for aligning large language models with multiple objectives. MCA constructs expert and adversarial prompts for each alignment objective through an iterative process: (1) sampling examples from the model (2) Ranking according to reward model (3) Add top and bottom responses as few shot examples for sampling new data. After iteration, the top and bottom ranked examples are used to generate expert prompts. At inference time, it uses contrastive decoding with these prompts to control the model's output. The approach combines contrasts from different objectives using a weighted formula based on user preferences. Experiments were conducted on datasets like HH-RLHF and SafeRLHF, using language model backbones such as Llama-2-7b and Phi-2. The method does not require model retraining and allows for real-time adjustment of model behavior.

Summary Of Strengths:

- MCA aligns the model without updating the parameters; this allows for personalization of the solution and real time adjustments.
- The approach allows for incorporation of new alignment objectives without retraining the model.
- The authors show that the approach can be combined with existing alignment methods.

Summary Of Weaknesses:

- The main limitation is the soundness of the evaluation. The same reward model is used in modeling (creation of expert prompts) and evaluation. This problem can hide issues of reward hacking. This is a common issue of all the model being tested promoting solutions that can overfit the reward model better through different means. A round of human evaluation would be preferable.
- The process of "distilling" the reward model in a pair of expert prompts is limiting. Reward models are trained on a large set of data and they can be nuanced contextual and capture different domains (particularly by scaling training data) such as math, code, QA, RAG, etc reducing that information to two prompts will be limiting.

Comments Suggestions And Typos:

- The authors should clearly highlight that the proposed approach could potentially be used to remove (or limits) safety alignment from existing models by inverting the inference-time optimization dimensions. Although not surprising and a property shared by most decoding time methods, it is important to clearly highlight and possibly quantify such limitation.
- Similar multi-objective decoding time work was proposed this year which is worth mentioning specifically "DeAL: Decoding-time Alignment for Large Language Models" which has similar properties.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 2.5
Overall Assessment: 3.5
Best Paper: No
Ethical Concerns:

- The authors should clearly highlight that the proposed approach could potentially be used to remove (or limits) safety alignment from existing models by inverting the inference-time optimization dimensions. Although not surprising and a property shared by most decoding time methods, it is important to clearly highlight and possibly quantify such limitation.

Needs Ethics Review: No
Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
Datasets: 1 = No usable datasets submitted.
Software: 1 = No usable software released.
Knowledge Of Or Educated Guess At Author Identity: Yes
Knowledge Of Paper: Before the review process
Knowledge Of Paper Source: Preprint on arxiv
Impact Of Knowledge Of Paper: Not at all
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**

Further feedback and discussion are appreciated!

Official Comment
by Authors (👁 Julian McAuley (/profile?id=~Julian_McAuley1), Rui Yan (/profile?id=~Rui_Yan2), Tingchen Fu (/profile?id=~Tingchen_Fu1), Yupeng Hou (/profile?id=~Yupeng_Hou1))
30 Sept 2024, 20:04 (modified: 02 Jan 2025, 08:40)
👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 8QLN, Commitment Readers
Revisions (/revisions?id=pRP9LtzqOs)

Comment:
Dear Reviewer 8QLN,

Thank you again for your valuable time in reviewing our work and your constructive feedback. We posted our response to your comments approximately one day ago, and we wonder if you could kindly share some of your thoughts so we can keep the discussion rolling to address your concerns if there are any.

In the previous response,

1. We explain our evaluation method in the experiments and clarify why using the same reward model does not incur reward hacking in our problem setting.
2. We discuss the limitations of our approach and the potential malicious use of our approach. The discussion will be put into the Limitations Section in the final version.
3. We discuss the relationship between our approach and a recent work DeAL, analyzing the difference between two approaches.

We would appreciate it if you could kindly take a look at our response to your comments. If you have any further questions, we are happy to discuss them!

Best regards,

All authors of Paper39

Add: **Author-Editor Confidential Comment**

Thanks for your review!

Official Comment
by Authors (👁 Julian McAuley (/profile?id=~Julian_McAuley1), Rui Yan (/profile?id=~Rui_Yan2), Tingchen Fu (/profile?id=~Tingchen_Fu1), Yupeng Hou (/profile?id=~Yupeng_Hou1))

29 Sept 2024, 19:49 (modified: 02 Jan 2025, 08:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 8QLN, Commitment Readers

Revisions (/revisions?id=0HrDW2ZWuz)

Comment:
Thanks for your time and expertise in reviewing our work and we are encouraged by your recognition of the strengths of our work. After carefully reading through your comments, we would like to address your concerns one by one as follows:

Q1: The risk of reward hacking.

A1: Good question! We acknowledge that there is a potential risk of reward hacking if the reward model fails to be a good proxy of human preference and it is an important research question to understand and control the trade-off between different alignment objectives under reward distribution shift or noisy reward models.

However, in our study, it is a different case. In our problem formulation (Section 3.1) we assume access to the golden/user-defined reward model on different alignment dimensions since our approach emphasizes real-time alignment at inference. Meanwhile, our setting follows previous works [1][2] in this problem, which also assume access to golden reward models. Thanks for your suggestions and we will make it more clear in the final version.

Additionally, to mitigate the potential reward hacking in case the reward model is imperfect in a real scenario, we keep our approach to be gradient-free without updating the parameter towards the reward function and limit the maximum number of iterations at the iterative prompt construction process. In this way, our approach can alleviate the potential over-optimization and reward hacking.

Q2: The process of "distilling" the reward model in a pair of expert prompts is limiting.

A2: Good question! Indeed modern reward models are trained on large scales of data across various domains and it is almost impossible to obtain a prompt whose effect is exactly equivalent to a reward model and we acknowledge that our expert prompt or adversarial prompt can by no means be a perfect substitution for a reward model. Instead, getting inspiration from previous works in prompt optimization [3][4], we aim at approximately representing the human preference learned by a reward model with a textual and human-readable prompt such that we can control the effect of reward models in a more flexible way at decoding time.

Q3: Potential misuse to remove safety alignment.

A3: Thanks for your suggestions! We totally agree that malicious prompts can be used to induce undesired behaviors or safety issues from the language model. We would highlight this point in the Limitations Section and the Ethical Considerations Section, emphasizing that our approach is designed for controlling the trade-off of multi-objective alignment and should not be used for any malicious purposes.

Q4: Missing related work.

A4: Thanks for your suggestions! DeAL [5] is similar to our approach in the sense that it also provides a decoding-time solution to the alignment problem of LLMs and the difference between DeAL and ours lies in how to use the reward model. In detail, our approach "distills" or transforms the reward model into an expert prompt and an adversarial prompt, which will inevitably suffer from information loss. In contrast, DeAL directly incorporates reward models into the probability distribution of the next token prediction and the ability of the reward model is maintained. We will definitely include the discussion of DeAL in Section 2 and add it into Table 1 in the final version.

[1] Rewards-in-Context: Multi-objective Alignment of Foundation Models with Dynamic Preference Adjustment, ICML 2024.

[2] Beyond One-Preference-Fits-All Alignment: Multi-Objective Direct Preference Optimization, ACL 2024.

[3] Connecting Large Language Models with Evolutionary Algorithms Yields Powerful Prompt Optimizers, ICLR 2024.

[4] Large language models as optimizers, ICLR 2024.

[5] DeAL: Decoding-time Alignment for Large Language Models, Arxiv.

Add: **Author-Editor Confidential Comment**

Official Review of Submission39 by Reviewer uRXA

Official Review by Reviewer uRXA 17 Sept 2024, 08:06 (modified: 20 Dec 2024, 12:34)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer uRXA, Commitment Readers

Revisions (/revisions?id=ivdNkiokcs)

Paper Summary:
The paper introduces a Multi-objective Contrastive Alignment (MCA) method that aims to control different alignment objectives (e.g., helpfulness, harmlessness) during decoding time, without requiring additional model training. The core idea is to generate contrastive expert and adversarial prompts for each alignment objective, using these to steer the model's output in the desired direction.

Summary Of Strengths:
1.No Need for Retraining: One key advantage of MCA is that it does not require retraining a language model for each alignment objective. This makes it much more scalable compared to other approaches, which often require training multiple models or merging models to accommodate user preferences.

2.Versatility: The paper demonstrates MCA's ability to incorporate new alignment objectives, offering flexibility in adjusting the alignment dimensions at the decoding stage.

3.Empirical Validation: The authors provide a series of experimental results, showing that MCA performs well in achieving a balanced trade-off between alignment objectives like helpfulness and harmlessness.

Summary Of Weaknesses:
1.The timeliness of models and benchmarks: The paper uses older models such as Llama-2-7b and older benchmarks for evaluation. Using more recent models like Llama-3/3.1 or Qwen2, along with updated benchmarks like AlpacaEval, could provide more relevant results.

2.Prompt Dependency: MCA's performance may be overly dependent on the quality of the prompts used. Poorly crafted prompts could lead to suboptimal alignment or generate less desirable outputs.

Comments Suggestions And Typos:
N/A

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.
Soundness: 3.5
Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.
Best Paper: No
Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
Datasets: 1 = No usable datasets submitted.
Software: 1 = No usable software released.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**

Further feedback and discussion are appreciated!

Official Comment

by Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Rui Yan (/profile?id=~Rui_Yan2), Tingchen Fu (/profile?id=~Tingchen_Fu1), Yupeng Hou (/profile?id=~Yupeng_Hou1))

30 Sept 2024, 20:09 (modified: 02 Jan 2025, 08:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer uRXA, Commitment Readers

Revisions (/revisions?id=O928laS8ht)

Comment:

Dear Reviewer uRXA,

Thank you again for your valuable time in reviewing our work and your constructive feedback. We posted our response to your comments approximately one day ago, and we wonder if you could kindly share some of your thoughts so we can keep the discussion rolling to address your concerns if there are any.

In the previous response,

1. We explain the rationale for our experimental setup and why Alpaca is probably not a good choice for multi-objective alignment.
2. Additionally, the generalization ability of our approach on more diverse alignment dimensions is verified.
3. We discuss the potential risk of prompt dependency and our efforts in iterative prompt construction (Section 3.2) to deal with the problem.

We would appreciate it if you could kindly take a look at our response to your comments. If you have any further questions, we are happy to discuss them!

Best regards,

All authors of Paper39

Add: **Author-Editor Confidential Comment**

Thanks for your review!

Official Comment

by Authors (Julian McAuley (/profile?id=~Julian_McAuley1), Rui Yan (/profile?id=~Rui_Yan2), Tingchen Fu (/profile?id=~Tingchen_Fu1), Yupeng Hou (/profile?id=~Yupeng_Hou1))

29 Sept 2024, 19:51 (modified: 02 Jan 2025, 08:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer uRXA, Commitment Readers

Revisions (/revisions?id=m4TRl8OS7v)

Comment:

Thanks for your time and expertise in reviewing our work and we are grateful for your recognition that our approach is training-free, versatile, and empirically effective. After carefully reading through your comments, we would like to address your concerns one by one as follows:

Q1: Experiment on recent models and recent benchmarks.

A1: Thanks for your suggestions! Our experiment setup including the choice of backbone models and benchmarks mostly follows RiC [1]. When considering relevant benchmarks, we find that AlpacaEval might not be an ideal choice since it does not provide fine-grained scores on each alignment dimension, which are indispensable in our multi-objective alignment experiment. Instead, we employ the SafeRLHF dataset that is published within a year [2].

Following your advice, we perform experiments with the recently released Qwen-2.5-14b on the Ultrafeedback dataset and the trade-off between honesty and instruction-following is shown in the Table below. We use ArmoRM [3] to measure the model performance on these two alignment dimensions. We can observe from the table that our approach achieves effective control over honesty and instruction-following, verifying the generalization of our approach.

w	honesty	instruction-following
(1,0)	75.82	67.47
(0.75, 0.25)	74.80	68.31
(0.5, 0.5)	73.43	69.74
(0.25, 0.75)	71.70	70.36
(0,1)	71.24	71.59

Q2: Prompt dependency.

A2: Good question! We agree that prompt sensitivity is a shared property among LLM and their performance can be affected by the quality of the prompt [4]. Our ablation study on the prompt with the variant phi-2+keyword in Section 5.1 also verifies this point. Therefore, we are motivated to propose the iterative prompt construction (Section 3.1) to obtain LLM-preferred prompt, as black-box prompt optimization with a proprietary LLM is shown to be an effective strategy in LLM alignment [5]. And in our case, we find that the proposed black-box prompt optimization method can provide high-quality prompts, as evidenced by the reward statistics in Section 5.2. But considering possibility of suboptimal or undesired output, we will definitely follow your advice and put this point into the Limitations Section.

[1] Rewards-in-Context: Multi-objective Alignment of Foundation Models with Dynamic Preference Adjustment, ICML 2024.

[2] BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset, NeurIPS 2023 Datasets and Benchmarks Track.

[3] Arithmetic Control of LLMs for Diverse User Preferences: Directional Preference Alignment with Multi-Objective Rewards, Arxiv.

[4] Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting, ICLR 2024.

[5] Black-Box Prompt Optimization: Aligning Large Language Models without Model Training, ACL 2024.

[6] Large language models as optimizers, ICLR 2024.

Add: **Author-Editor Confidential Comment**

Terms of Use (/legal/terms)

Privacy Policy (/legal/privacy)