Strong Empowered and Aligned Weak Mastered Annotation for Weak-to-Strong Generalization

PDF (/pdf?id=cYOpBvPPP3)

Yongqi Li (/profile?id=~Yongqi_Li3), Xin Miao (/profile?id=~Xin_Miao4), Mayi Xu (/profile?id=~Mayi_Xu1), Tieyun Qian (/profile?id=~Tieyun_Qian1)

Published: 14 Dec 2024, Last Modified: 08 Feb 2025 AAAI-AIA 2025 Poster AI Alignment Track, Senior Program Committee, Program Committee, Authors Revisions (/revisions?id=cYOpBvPPP3) BibTeX CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Primary Keyword: AI Alignment (AIA) -> AIA: Superintelligence

TL;DR: A novel "Strong Empowered and Aligned Weak Mastered" framework for the weak-to-strong generalization problem.

Abstract:

The super-alignment problem of how humans can effectively supervise super-human AI has garnered increasing attention. Recent research has focused on investigating the weak-to-strong generalization (W2SG) scenario as an analogy for super-alignment. This scenario examines how a pre-trained strong model, supervised by an aligned weak model, can outperform its weak supervisor. Despite good progress, current W2SG methods face two main issues: 1) The annotation quality is limited by the knowledge scope of the weak model; 2) It is risky to position the strong model as the final corrector.

To tackle these issues, we propose a ``Strong Empowered and Aligned Weak Mastered'' (SEAM) framework for weak annotations in W2SG. This framework can leverage the vast intrinsic knowledge of the pre-trained strong model to empower the annotation and position the aligned weak model as the annotation master. Specifically, the pre-trained strong model first generates principle fast-and-frugal trees for samples to be annotated, encapsulating rich sample-related knowledge. Then, the aligned weak model picks informative nodes based on the tree's information distribution for final annotations. Experiments on six datasets for preference tasks in W2SG scenarios validate the effectiveness of our proposed method.

Supplementary Material: zip (/attachment?id=cYOpBvPPP3&name=supplementary_material)
iThenticate Agreement: Yes, I agree to iThenticate's EULA agreement version: v1beta
Reproducibility Checklist: I certify all co-authors of this work have read and completed the Reproducibility Checklist.
Submission Number: 173

Filter by reply type... | Filter by author... | Search keywords... | Sort: Newest First

Everyone | Program Chairs | Submission173... | Submission173 Authors | Submission173... | Submission173... ✖

8 / 8 replies shown

Add: **Withdrawal**

Paper Decision

Decision by Program Chairs 14 Dec 2024, 12:43 (modified: 14 Dec 2024, 12:47)
Program Chairs, Program Committee, Authors Revisions (/revisions?id=LQfDHiwdug)

Decision: Accept (Poster)

Rebuttal by Authors

Rebuttal

by Authors (Yongqi Li (/profile?id=~Yongqi_Li3), Xin Miao (/profile?id=~Xin_Miao4), Mayi Xu (/profile?id=~Mayi_Xu1), Tieyun Qian (/profile?id=~Tieyun_Qian1))

29 Nov 2024, 18:15 (modified: 30 Nov 2024, 13:33)

Program Chairs, Program Committee, Program Committee Submitted, Authors

Revisions (/revisions?id=9f9IYHy6Pb)

Rebuttal:

We sincerely thank you for your valuable comments.

Full version

Due to the 2500-character limit, please **download the full version of our response** at https://anonymous.4open.science/r/SEAM-6E70/Response.pdf (https://anonymous.4open.science/r/SEAM-6E70/Response.pdf) .

Reviewer GYYw

Q1

	AF	HS	AM	CH	AHH	SR	Avg.
SEAM w/ CoTv1 (with principles & principle definitions)	61.6	68.4	44.5	54.5	54.0	52.4	55.9
SEAM w/ CoTv2 (with principles)	61.2	69.5	44.5	56.9	53.1	51.5	56.1
SEAM w/ CoTv3 (naïve)	65.2	70.6	42.8	55.6	55.1	53.8	57.2
SEAM w/ FF-Tree	67.2	81.9	57.8	68.2	61.5	60.8	66.3

We conduct above experiments per your advice and report annotation quality of 3 CoT methods. The results clearly show FF-Tree can induce knowledge from the strong LLM.

Q2

We replace the unaligned strong LLM (Qwen-2-7B) in AuxConf and WSC Filter with an aligned one (Qwen-2-7B-Instruct). This change led to a 1.6% average decrease in risky corrections, suggesting that the misalignment in strong LLM indeed results in risky corrections to some extent.

Reviewer tTJ5

Q1

In super-alignment, humans (weak models) need to master the annotations, making the resulting quality issues unavoidable.

Q2

We have designed strategies like "Best-First Search" and "Backtrace Limits" to make the framework more efficient.

Q3

The scope of preference task is broad, covering various types, e.g., QA, summarization.

Reviewer LP5v

Q1

Different focus: Scalable oversight (SO) focuses on supervision **quality**, while W2SG/Super-alignment cares more about **safety issues**.

	AF	HS	AM	CH	AHH	SR	Avg.
Debate SO [1-2]	55.5	51.2	46.7	48.5	51.7	49.3	50.5
Ours	67.2	81.9	57.8	68.2	61.5	60.8	66.3

Poor performance of SO: The results indicate that current SO methods performs poorly.

Q2

First, OpenAI's W2S work converted all datasets into binary classification tasks, i.e., the same as our preference task (choosing the better response). Besides, the scope of preference task is broad, covering various types, e.g., QA, summarization.

> Q3

We evaluate FF-Trees using GPT-4o, rating from 1 to 5 based on: A1 (necessary principle inclusion), A2 (redundant principle exclusion), and A3 (adequacy of principle-aware thought).

	AF	HS	AM	CH	AHH	SR	Avg.
Avg (A1, A2, A3)	4.1	4.1	3.9	3.8	4.0	3.8	3.9

We can see that the FF-Tree quality achieves a good score.

[1] Debate Helps Supervise Unreliable Experts, 2023
[2] Debating with More Persuasive LLMs Leads to More Truthful Answers, 2024

(Update) Novel Improvement over CoT, important flaws in presentation, experimental setup

Official Review by Program Committee GYYw 26 Nov 2024, 01:21 (modified: 04 Dec 2024, 03:51)
Program Chairs, Senior Program Committee, Program Committee, Authors Revisions (/revisions?id=IeVUY3hkpw)

Review:
Brief Summary of Paper

- to improve W2SG, strong model generated fast and frugal decision tree (FFT) informed by pre-specified principles, which the weak supervisor used to inform labels
- this method outperforms previously proposed W2SG methods, largely based on strong and weak model confidence

Brief summary of criticisms:

- should compare FFT to baseline CoT techniques (with and without pre-specified principles)
- claim that strong models don't understand human values is unsubstantiated (and likely wrong/confused)

Overall, the approach of "use strong model reasoning to inform weak model supervision", is compelling, and while not particularly novel in the context of (language model) alignment in general (see [1-3]), has not been rigorously tested in the context of weak to strong generalization.

However, because there are strong a-priori reasons to suspect process-based oversight methods would improve over naive weak model fine-tuning, its hard to tell whether fast and frugal decision trees have particular benefits over more standard CoT prompt engineering [4-6]. That there is no engagement with prior work in CoT prompt engineering is troubling. Furthermore, while the authors perform an ablation of FFT where the strong model "provides sample related knowledge without referencing principles", no details on the elicitation strategy are provided. TLDR; I would feel much more confident in relative benefit of FFT if it was compared to baseline prompts like "does this output satisfy principles [p_1, ..., p_n], let's think step by step".

Separately, the study is premised on two supposed problems with weak-to-strong generalization

1. annotation quality of weak models is limited by knowledge scope
2. risky to position strong model as final corrector, because the strong model may "lack" particular values

In some sense this decomposition seems reasonable, if not tautological - weak models make mistakes because they are...well...weak, and strong models might do things we don't want (hence the need for supervision). But later, the paper makes a more substantive claim about problems in methods that leverage the confidence of the strong model, claiming these methods fail because "the unaligned strong LLMs do not comprehend human values related to harmlessness". However, the authors provide no evidence for this claim beyond the poor performance of the auxiliary confidence and weak-strong consistency filters, and neglect the more intuitive "null hypothesis" that poor performance is caused by mimicking errors in the weak supervision signal. In general, I would suggest the authors move away from strong claims about the cause of strong learner errors, or appeal to the incentives the strong model has to mimic weak supervisor errors rather than output according to its latent knowledge.

Given the complexity of proposed method, lack of comparison to more simple baselines, and flawed framing/presentation, I tentatively recommend rejection, though the general application of strong model reasoning to enhance weak oversight seems promising and novel in the context of weak to strong generalization.

Some small presentation/grammar/typo notes: "Finally, we call the W2SG phenomenon occurs if"

- "Second, if the aligned weak model is familiar with American culture and produces the correct annotation initially."
- the initial immigration example is hard to follow. In particular, its unclear why we would expect the strong model to know immigration is a sensitive topic, but not know that AI's should not express opinions on sensitive topics (though maybe this is downstream of my previous comment)

Update After Revision

Give the chain-of-though baselines and the modified language, I am upgrading my review to a 6, though I still have lingering questions about the experimental setup. In particular, while the weak model ultimately determined the labels, the setup implicitly assumes the strong model is faithfully trying to execute the prompt and not trying to systematically deceive the user. In other scalable oversight works (see Khan et. al. 2024 (https://arxiv.org/abs/2402.06782), Kenton et. al. 2024 (https://arxiv.org/abs/2311.08702)), the baseline "consultancy" approach seeds the strong model with the correct or incorrect perspective randomly, on the assumption that future strong models may be trying to deceive supervisors. The worry with the present work is that its always seeded with the correct perspective (i.e. follow principles x,y,z), thus breaking an important analogy with future systems. I think this is fine, as long as the authors acknolwedge this problem - future work could explore using FFTs in consultancy and debate.

[1] Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2023). Let's Verify Step by Step. ArXiv, abs/2305.20050.

[2] Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., & Higgins, I. (2022). Solving math word problems with process- and outcome-based feedback. ArXiv, abs/2211.14275.

[3] Stuhlmuller, A., Byun, J. (2022) Supervise Process, not Outcomes https://ought.org/updates/2022-04-06-process (https://ought.org/updates/2022-04-06-process)

[4] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Xia, F., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. ArXiv, abs/2201.11903.

[5] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. ArXiv, abs/2305.10601.

[6] Besta, M., Blach, N., Kubíček, A., Gerstenberger, R., Gianinazzi, L., Gajda, J., Lehmann, T., Podstawski, M., Niewiadomski, H., Nyczyk, P., & Hoefler, T. (2023). Graph of Thoughts: Solving Elaborate Problems with Large Language Models. AAAI Conference on Artificial Intelligence.

Rating: 6: Marginally above acceptance threshold

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Review Rebuttal of Submission173 by Authors

Review Rebuttal

by Authors (Yongqi Li (/profile?id=~Yongqi_Li3), Xin Miao (/profile?id=~Xin_Miao4), Mayi Xu (/profile?id=~Mayi_Xu1), Tieyun Qian (/profile?id=~Tieyun_Qian1))

08 Dec 2024, 10:06 Program Chairs, Senior Program Committee, Program Committee, Authors

Rebuttal:

Thank you so much for taking the time to read our responses and raise the score. We are pleased to know that we have addressed your concerns. We will include the comparison results with the newly added CoT baselines and improve the presentation. We will also continue to explore using strong models to generate more faithful FF-Trees in future work. Such a discussion is helpful to improve our work. Thanks again.

Official Review

Official Review by Program Committee tTJ5 25 Nov 2024, 11:00 (modified: 26 Nov 2024, 09:08)

Program Chairs, Senior Program Committee, Program Committee, Authors Revisions (/revisions?id=sXJh5ENOZ2)

Review:

Summary

The paper presents a novel framework, SEAM, to address the weak-to-strong generalization (W2SG) problem, making a contribution to the field of AI alignment and super-alignment. The authors propose leveraging the knowledge of a pre-trained strong model to empower weak annotations and position an aligned weak model as the annotation master which provides inspiration for future research.

Strengths

1. Innovative Framework: The introduction of the SEAM framework represents a novel approach to addressing the super-alignment problem in AI. By emphasizing the relationship between strong and weak models in the context of weak-to-strong generalization, the framework adds valuable insights.
2. Empirical Validation: The experiments conducted across six different datasets provide a solid foundation for evaluating the proposed method, contributing to the paper's credibility by demonstrating its effectiveness in practical scenarios.

Weaknesses

1. Dependence on annotation quality: Although the knowledge tree generated by the strong model can enhance the effectiveness of the annotations, there is still a reliance on the weak model, which may limit the accuracy of the final results, especially when the knowledge scope of the weak model is insufficient.
2. Complexity issues: The implementation of the SEAM framework may introduce a certain level of complexity, particularly in the processes of generating decision trees and selecting nodes for weak models. Effectively integrating these two aspects may pose challenges for technical implementation.
3. Generality and adaptability: Current research appears to focus on preferred tasks, lacking comprehensive evaluation of other types of tasks or datasets, which may limit the applicability and wider adoption of the methods.

Suggestions and Questions

1. Enhancing the assessment of annotation quality: To further improve the effectiveness of the SEAM framework, it is recommended to include a more comprehensive analysis of weak model performance and annotation quality in the research, especially regarding the performance of weak models across different knowledge domains.
2. Exploring simplified solutions: Research can focus on methods to simplify the framework to reduce complexity, making it more accessible and applicable. This may include developing more straightforward decision tree generation algorithms, among other approaches.
3. Expanding the experimental scope: In future research, it is recommended to validate the applicability of the SEAM framework across different task types (such as image classification) and datasets. This would provide stronger support for the framework's generalizability.

Rating: 6: Marginally above acceptance threshold

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Review Rebuttal of Submission173 by Authors

Review Rebuttal

by Authors (Yongqi Li (/profile?id=~Yongqi_Li3), Xin Miao (/profile?id=~Xin_Miao4), Mayi Xu (/profile?id=~Mayi_Xu1), Tieyun Qian (/profile?id=~Tieyun_Qian1))

08 Dec 2024, 10:13 Program Chairs, Senior Program Committee, Program Committee, Authors

Rebuttal:

Thank you for your efforts during the review phase and for your interest in and recognition of our work. We have responded to the three issues you raised and hope these responses can alleviate your concerns. Once again, we appreciate your hard work during the review process, as it is very important for improving our work.

This work attempts to improve weak supervision with the knowledge from a strong model, which is extensively studied in the field of scalable oversight. Unfortunately, this study never cites, discusses, or compares any works in the experiments in the field of scalable oversight.

Official Review by Program Committee LP5v 20 Nov 2024, 14:36 (modified: 26 Nov 2024, 09:08)

Program Chairs, Senior Program Committee, Program Committee, Authors Revisions (/revisions?id=GJNCdDGcXg)

Review:

This work introduces a new pipeline in W2SG, which uses knowledge from a strong model to improve annotation quality of a weak model.

Pros:

1. This work presents a searching-while-thinking algorithm to generate principle FF-trees that can effectively induce required knowledge from the pre-trained strong model.
2. Experiments on preference tasks show the proposed method outperforms baselines.

Cons:

1. Improving the ability of humans to supervise more capable models and improving weak supervision with the help of a strong model is extensively studied in the field of scalable oversight. Unfortunately, this study never cites, discusses, or compares any works in the experiments in the field of scalable oversight [1,2,3,4]. [1] Scalable agent alignment via reward modeling: a research direction, 2018 [2] Measuring progress on scalable oversight for large language models, 2022 [3] Debate Helps Supervise Unreliable Experts, 2023 [4] Debating with More Persuasive LLMs Leads to More Truthful Answers, 2023
2. Experiments in this work are only conducted on the preference tasks. The OpenAI authors of W2S Generalization have open-sourced NLP benchmarks (https://github.com/openai/weak-to-strong (https://github.com/openai/weak-to-strong)), including QA, dialogue, and HH tasks. Extensive experiments on diverse tasks are needed to prove the strength of the proposed method.
3. Fast-and-Frugal Tree is originally applied in a specific domain, i.e., medical decision making (Figure 3). This work uses it in an open domain for extracting knowledge from a strong pre-trained model. The quality of generated fast-and-frugal trees should be evaluated and case studies of generated fast-and-frugal trees for real testing cases should be provided.

Rating: 5: Marginally below acceptance threshold

Confidence: 5: The reviewer is absolutely certain that the evaluation is correct and very familiar with the relevant literature

Review Rebuttal of Submission173 by Authors

Review Rebuttal

by Authors (Yongqi Li (/profile?id=~Yongqi_Li3), Xin Miao (/profile?id=~Xin_Miao4), Mayi Xu (/profile?id=~Mayi_Xu1), Tieyun Qian (/profile?id=~Tieyun_Qian1))

08 Dec 2024, 10:11 Program Chairs, Senior Program Committee, Program Committee, Authors

Rebuttal:

Thank you for your review comments again. We have posted responses to alleviate your concerns about "comparison with scalable oversight", "task selection", and "the quality evaluation of generated FF-Trees". We sincerely hope that you can read our response and provide further feedback on whether your concerns have been addressed. Such feedback is very important to us. Thanks again.