← Back to **Author Console** (/group?id=NeurIPS.cc/2025/Conference/Authors#your-submissions)

Let's Try Again: Eliciting Multi-Turn Reasoning in Language Models via Simplistic Feedback

(/pdf?id=ocWZ3aKdos)

Licheng Liu (/profile?id=~Licheng_Liu5), Zihan Wang (/profile?id=~Zihan_Wang23), Linjie Li (/profile?id=~Linjie_Li1), Chenwei Xu (/profile?id=~Chenwei_Xu2), Yiping Lu (/profile?id=~Yiping_Lu1), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), Manling Li (/profile?id=~Manling_Li1)

02 May 2025 (modified: 18 Sept 2025) Submitted to NeurIPS 2025 Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions (/revisions?id=ocWZ3aKdos) BibTeX

Keywords: Reinforcement Learning, Large Reasoning Model

Abstract:
Multi-turn problem solving is a critical yet challenging scenario in the practical application of Large Reasoning Models (LRMs), commonly encountered in domains such as chatbots, programming assistants, and education. Recently, reasoning models like DeepSeek-R1 has shown the promise of reinforcement learning (RL) methods in enhancing model reasoning capabilities. However, we observe that models trained with existing single-turn RL paradigms often \textbf{lose their ability to solve problems across multiple turns}, struggling to revise answers based on context and exhibiting repetitive responses. This raises new challenges in preserving reasoning abilities while enabling multi-turn contextual adaptation.
In this work, we find that simply allowing models to engage in multi-turn problem solving where they receive only unary feedback (e.g., "Let's try again") after incorrect answers can help recover both single-turn and interactive multi-turn reasoning skills. We introduce \textbf{Unary Feedback as Observation (UFO)} for reinforcement learning, a method that explicitly leverages minimal yet natural user feedback during iterative problem-solving which can be easily applied to any existing single-turn RL training paradigm. Experimental results show that RL training with UFO preserves single-turn performance while improving multi-turn reasoning accuracy by 14%, effectively utilizing sparse feedback signals when available. To further reduce superficial guessing and encourage comprehensive reasoning, we explore reward structures that incentivize thoughtful, deliberate answers across interaction turns. Code and models will be publicly released.

Checklist Confirmation: I confirm that I have included a paper checklist in the paper PDF.
Supplementary Material: zip (/attachment?id=ocWZ3aKdos&name=supplementary_material)
Financial Support: ll5622@ic.ac.uk
Reviewer Nomination: Zihan Wang (/profile?id=~Zihan_Wang23)
Responsible Reviewing: We acknowledge the responsible reviewing obligations as authors.
Primary Area: Reinforcement learning (e.g., decision and control, planning, hierarchical RL, robotics)
LLM Usage: Editing (e.g., grammar, spelling, word choice)
Declaration: I confirm that the above information is accurate.
Submission Number: 6157

Filter by reply type... Filter by author... Search keywords... Sort: Newest First
Everyone Program Chairs Submission6157 Authors Submission6157...
28 / 28 replies shown

Submission6157 Area... | Submission6157... | Submission6157... | Submission6157... | Submission6157... | Submission6157... | ✖

Add: **Withdrawal**

Paper Decision

Decision by Program Chairs 17 Sept 2025, 07:41 (modified: 18 Sept 2025, 08:25) Program Chairs, Authors

Revisions (/revisions?id=bFjMGEiccr)

Decision: Reject

Comment:

The authors note that models trained under standard single-turn RL frameworks often struggle with multi-turn problem-solving. To address this, they introduce simple verbal unary feedback ("Let's Try Again") that prompts the model to retry after an incorrect response, leading to an improvement in multi-turn reasoning accuracy. The method appears simple and effective, but the empirical evaluation is somewhat limited as judged by the reviewers. The baselines also appear to be insufficient.

Official Comment by Authors

Official Comment

by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))

09 Aug 2025, 05:07

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

Comment:

We thank the AC and reviewers for the valuable feedback. We are encouraged that all the reviewers agreed that our method is simple yet effective. We believe this work makes a meaningful contribution to the multi-turn RL community and advances human–AI interaction. We also appreciate that all reviewers raised their scores after discussion. We summarize the reviewers' main concerns and how we addressed them.

---

1. **Limited Scope of Evaluation**: We trained 5 open-source models and evaluated the effectiveness of UFO on 9 datasets, including Math reasoning, STEM reasoning, QA and general tasks. We found our UFO framework showed great generalization ability and consistently improved performance across diverse domains, demonstrating its robustness and broad applicability. This finding also suggests that UFO can, to some extent, enhance the model's self-reflection ability, as evidenced by performance gains on out-of-distribution datasets.
2. **Insufficient theoretical analysis**: We added a mathematical analysis explaining why off-the-shelf large models tended to produce repeated and identical answers. Building on this, we distinguished our contributions from the self-emergent reflective behaviors in long Chain-of-Thought reasoning and from the objectives of max-entropy single-turn RL. We also explained why UFO is effective in mitigating the repetition problem, making our study more than just an empirical study. By combining both the analysis of why the existing problem occured and how our method addressed it, we laid a solid foundation for our work.
3. **Effectiveness beyond prompt complexity**: We conducted experiments on replacing simple feedback with more complex prompt, and found that UFO was both sufficient and efficient. We also showed that the repetition problem could not be simply addressed by more complex prompt using theoretical analysis.

We will incorporate all of our clarifications, additional analyses, and our latest experimental results in our final version. We thank the AC and reviewers for their invaluable time and feedback to make our work better.

Best regards,

The authors

Author AC Confidential Comment by Authors

Author AC Confidential Comment

by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))

05 Aug 2025, 23:31 Program Chairs, Senior Area Chairs, Area Chairs, Authors

Comment:

Dear Area Chair,

With two days left in the rebuttal period, only one of our reviewers have responded. We have provided detailed replies to every comment, including new results asked by the reviewers. We also believe any outstanding concerns stem from presentation rather than substance.

Could you please prompt the reviewers to indicate whether our clarifications resolve their points? If not, we are ready to supply further clarifications, experiment updates, and analysis before the deadline.

Thank you for your assistance!

Official Review of Submission6157 by Reviewer 6LhU

Official Review by Reviewer 6LhU 02 Jul 2025, 11:18 (modified: 18 Sept 2025, 10:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 6LhU

Revisions (/revisions?id=l12WsKibcA)

Summary:

The authors observe that models trained with existing single-turn RL paradigms often lose their ability to solve problems across multiple turns. They simply add verbal unary feedback (Let's Try Again) to encourage the model to try again when the model is wrong, which helps improve multi-turn reasoning accuracy by 14%.

Strengths And Weaknesses:

Strengths:

1. The paper introduces a simple and effective trick that can help enhance multi-turn reasoning capabilities.
2. The method can help recover both single-turn and interactive multi-turn reasoning skills.

Weaknesses:

1. Some data and experimental details are missing, such as the sizes of the training set and the validation set.
2. The paper only conduct experiments based on the Qwen-3B model, lacking sufficiency.
3. If "Let's Try Again" is incorporated into the single-turn training, will the model be able to correct the incorrect answers?
4. Whether the context length affect the method?
5. In Figure 6, it is not evident how the absence of feedback affects the model.

Quality: 3: good
Clarity: 3: good
Significance: 2: fair
Originality: 3: good
Questions:

Please see weaknesses.

Limitations:

yes

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Ethical Concerns: NO or VERY MINOR ethics concerns only

Paper Formatting Concerns:

No Formatting Concerns.

Code Of Conduct Acknowledgement: Yes

Responsible Reviewing Acknowledgement: Yes

Final Justification:

I give the final score to borderline accept.

Rebuttal by Authors

Rebuttal

by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))

30 Jul 2025, 04:11 (modified: 31 Jul 2025, 05:19) Program Chairs, Authors

Revisions (/revisions?id=zkjy6zA01i)

[Deleted]

Rebuttal by Authors

Rebuttal

by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))

31 Jul 2025, 05:19 (modified: 31 Jul 2025, 05:23) Program Chairs, Authors

Revisions (/revisions?id=0gvj75VRxo)

[Deleted]

Rebuttal by Authors

Rebuttal

by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))

31 Jul 2025, 05:23 (modified: 31 Jul 2025, 05:25) Program Chairs, Authors

Revisions (/revisions?id=itI3ouMXfB)

[Deleted]

Rebuttal by Authors

Rebuttal

by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))

31 Jul 2025, 05:25 (modified: 31 Jul 2025, 05:28) Program Chairs, Authors

Revisions (/revisions?id=D6xc7ANKK5)

[Deleted]

Rebuttal by Authors

Rebuttal

by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))

31 Jul 2025, 05:28 (modified: 31 Jul 2025, 05:48) Program Chairs, Authors

Revisions (/revisions?id=zCWf9wZPpg)

[Deleted]

Rebuttal by Authors

Rebuttal

by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))

31 Jul 2025, 05:48 (modified: 31 Jul 2025, 13:54)

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

Revisions (/revisions?id=wJ5HQpUsa9)

Rebuttal:

Thank you for your review!

We answer your questions in the following:

1. We appreciate your suggestion on improving reproducibility. We will include more comprehensive descriptions of the training and validation setups in the final version. In addition, we plan to open-source our code to provide full access to the data and experimental procedures.

2. We have conducted additional experiments using models of different sizes and other open-source architectures.
   We also evaluated our method on datasets beyond reasoning tasks, which demonstrates the generalization capability of our UFO framework.
   These results will be added to the final version to further strengthen the robustness and applicability of our approach.

Performance Table (Success Rate Across Models and Datasets)

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-1.5B-Instruct	0.109	0.117	0.266	0.219	0.625	0.024	0.031	0.523	0.352
RL on MMQ-Math	0.748	0.201	0.847	0.227	**0.655**	0.192	**0.095**	0.438	0.344
+5turn UFO	**0.836**	**0.268**	**0.881**	**0.273**	0.648	**0.226**	0.093	**0.609**	**0.348**

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-3B-Instruct	0.523	0.283	0.680	0.516	0.758	0.078	0.039	0.752	0.422
RL on MMQ-Math	0.797	0.320	0.930	0.501	0.776	0.195	0.129	0.668	0.483
+5turn UFO	**0.885**	**0.408**	**0.953**	0.523	0.875	0.266	0.152	**0.852**	**0.609**

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
RL on HotQA	0.724	0.318	0.891	0.484	0.813	0.383	0.166	0.715	0.493
+5turn UFO	0.727	0.292	0.850	**0.578**	**0.883**	**0.442**	**0.168**	0.766	0.489

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-7B-Instruct	0.564	0.321	0.563	0.625	0.836	0.133	0.047	0.723	**0.641**
RL on MMQ-Math	0.851	0.336	0.952	0.508	0.848	0.263	0.141	0.734	0.523
+5turn UFO	**0.930**	**0.421**	**0.968**	**0.569**	**0.848**	**0.286**	**0.164**	**0.805**	0.588

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Llama3.2-1B-Instruct	0.023	0.008	0.008	0.008	0.023	0.008	0.000	0.039	0.000
RL on MMQ-Math	0.539	0.211	0.523	0.203	0.570	0.195	0.008	0.578	**0.328**
+5turn UFO	**0.648**	**0.268**	**0.563**	**0.266**	**0.602**	**0.211**	**0.016**	**0.664**	**0.328**

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Llama3.2-3B-Instruct	0.508	0.203	0.484	0.477	0.773	0.297	0.060	0.656	0.492
RL on MMQ-Math	0.867	0.242	0.922	0.469	0.781	0.445	0.133	0.711	0.609
+5turn UFO	**0.922**	**0.320**	**0.938**	**0.508**	**0.820**	**0.398**	**0.148**	**0.828**	**0.664**

We also did experiments by replacing PPO with GRPO:

Model	MMQ	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-1.5B-Instruct	0.109	0.117	0.266	**0.219**	**0.625**	0.024	0.031	0.523	0.352
RL on MMQ-Math	0.750	0.219	0.859	0.180	0.570	0.249	**0.070**	0.523	0.289
+**5turnUFO**	**0.814**	**0.268**	**0.875**	0.203	0.617	**0.266**	**0.070**	**0.563**	**0.383**

Model	MMQ	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-3B-Instruct	0.523	0.283	0.680	**0.516**	0.758	0.078	0.039	0.752	0.422
RL on MMQ-Math	0.758	**0.352**	0.891	0.320	0.773	0.242	**0.141**	0.670	**0.461**
+**5turnUFO**	**0.883**	0.348	**0.914**	0.344	**0.781**	**0.320**	0.086	**0.766**	**0.461**

Legend:

- MMQ-Math = Math subset of MetamathQA (Math)
- TQA = TheoremQA (Math)
- GSM8k = GSM8k (Math)
- GPQA = GPQA (Science)
- STEM = MMLU-STEM (Science)
- HotQA = HotPotQA (QA)
- ConQA = ConcurrentQA (QA)
- MMLU = MMLU (General)
- MMLU-Pro = MMLU-Pro (General)

---

3. We consider two scenarios:
   Unary feedback cannot be incorporated during single-turn training with $T_{\max} = 1$.
   As described in Lines 123–132, the episode terminates at $T_{\max}$, meaning the agent never receives any feedback, and thus the learning signal from unary feedback is unavailable.

If unary feedback is used during evaluation while the model was trained with single-turn RL, Figure 4 illustrates that the performance still lags significantly behind that of UFO trained with 5-turn interactions.
This highlights the value of our UFO framework during training.

4. We appreciate the reviewer's concern regarding potential performance degradation due to increased context length. **We did not observe notable performance degradation as context length increased.** As detailed in Appendix D, our method is designed to remain effective under a reasonable context budget. In our experiments, the context length was capped at 8192 tokens, consistent with prior influential work in RL-for-reasoning such as 1-shot RLVR [1] and DeepScaleR [2].
   Additionally, we note that most existing studies in this area, including Reflexion [3], 1-shot RLVR [1], and AbsoluteZero [4], do not explicitly examine the effect of context length on reasoning performance. This appears to reflect a broader trend: current RL-for-reasoning methods typically operate under standard context limits without treating context size as a primary variable of interest.

5. We thank you for pointing this out. As noted in Lines 260–261, we conducted a controlled comparison between **5-turn training with and without unary feedback** and figure 6(a) shows that models trained with feedback achieve a **peak improvement of over 8%**. This pattern is consistent across multiple datasets, where feedback-enabled models outperform no-feedback baselines by **4–8%**. To reflect the stability of this finding more clearly, we will consider rephrasing the result in the final version (e.g., *"maximum consistent gain"*).

[1] Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Reinforcement Learning for Reasoning in Large Language Models with One Training Example. arXiv:2504.20571, 2025.

[2] Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. DeepScaleR: Surpassing O1-Preview with a 1.5 B Model by Scaling RL. agentica-org/DeepScaleR-1.5B-Preview (Hugging Face), 2025.

[3] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. NeurIPS, 2023.

[4] Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Hue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute Zero: Reinforced Self-play Reasoning with Zero Data. arXiv:2505.03335, 2025.

Official Comment by Authors

Official Comment

by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))

04 Aug 2025, 00:10

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 6LhU

Comment:
Dear Reviewer 6LhU,

Thank you again for your insightful feedback. We have carefully addressed your comments with detailed, point-by-point revisions and have included new experimental results to better support our main claims.

We would greatly appreciate your thoughts on whether these updates have resolved your concerns. Thank you again for your time and consideration.

Best regards,

The Authors

Replying to Official Comment by Authors

Official Comment by Reviewer 6LhU

Official Comment by Reviewer 6LhU 06 Aug 2025, 10:54

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 6LhU

Comment:

Thanks for your reply! I will raise my score.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer 6LhU

Mandatory Acknowledgement by Reviewer 6LhU 06 Aug 2025, 10:55

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

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Replying to Official Comment by Reviewer 6LhU

Official Comment by Authors

Official Comment

by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))

06 Aug 2025, 23:52

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 6LhU

Comment:

Thank you for your valuable feedback and time. We appreciate your comments to make our work better.

Official Review of Submission6157 by Reviewer kzS2

Official Review by Reviewer kzS2 20 Jun 2025, 04:33 (modified: 18 Sept 2025, 10:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer kzS2

Revisions (/revisions?id=vHICosBVPH)

Summary:

This paper addresses a critical challenge in applying Large Reasoning Models (LRMs) to practical, multi-turn problem-solving scenarios: models trained with standard single-turn Reinforcement Learning (RL) often lose their ability to revise answers based on context, exhibiting repetitive responses and failing to incorporate feedback. The authors introduce Unary Feedback as Observation (UFO), a simple yet effective method that leverages minimal, natural user feedback (e.g., "Let's try again") during iterative problem-solving. UFO integrates this sparse, negative-only feedback into the model's observation history, allowing it to learn revision strategies without requiring explicit positive feedback or detailed turn-by-turn supervision. The model is trained using PPO on static single-turn reasoning datasets transformed into multi-turn interaction episodes. Experimental results on the METAMATHQA (MATH partition) dataset show that RL training with UFO improves multi-turn reasoning accuracy by 14% (Succ@5) compared to single-turn RL baselines with an equivalent inference budget.

Strengths And Weaknesses:

Strengths

1. the paper is easy to follow
2. The method is simple and effective in math tasks.
3. The paper goes beyond just proposing UFO and investigates reward structures (decay, repetition penalty) to encourage more efficient and diverse reasoning, which is a valuable contribution to understanding how to guide these models.

Weaknesses:

1. Limited Scope of Evaluation for both dataset and model: Experiments are solely conducted on the MATH partition of METAMATHQA. While math reasoning is a strong testbed, the generalizability of UFO to other types of reasoning tasks (like logic puzzles or science problems (GPQA)) is not demonstrated. The experiments use Qwen-2.5-3B-Instruct. While a capable model, testing on a wider range of model architectures and sizes could strengthen the claims of generality.
2. The primary comparison is against single-turn RL. While UFO's strength is its minimal feedback, a brief discussion or comparison (even if theoretical) against other multi-turn RL approaches (such as Ragen (https://arxiv.org/pdf/2504.20073) and ArCHer (https://arxiv.org/abs/2402.19446)) that might use more structured (but still automated) feedback could provide further context.
3. As indicated in S1 (https://arxiv.org/pdf/2501.19393), the sequential scaling is much effective than the parallel scaling. From my understanding, the multi-turn RL and single-turn RL implemented in this paper can be regarded as sequential scaling and parallel scaling, respectively. Therefore, the improvement of pass@k of multi-turn RL over single-turn RL may not sufficiently support the effectiveness of the proposed method. A portion of this gain could stem from the structural advantage of the sequential format itself.

Quality: 3: good
Clarity: 3: good
Significance: 2: fair
Originality: 3: good
Questions:

1. Section 3.1 and 3.2 are too redundant, can merge to one section.
2. It is good to implement a baseline that uses the same multi-turn setup but, instead of a learned RL policy, uses a fixed heuristic after an incorrect answer (e.g., simply re-prompting the model with "That was incorrect. Let's try a different approach to solve this problem." without any specific RL training for this interaction). It is better to implement other multi-turn RL methods like Ragen (https://arxiv.org/pdf/2504.20073) and ArCHer (https://arxiv.org/abs/2402.19446).
3. The paper shows UFO improves success rates and reduces repetition. However, it's less clear how the reasoning process qualitatively changes. Does UFO encourage models to make more sophisticated revisions (e.g., identifying flawed assumptions, correcting logical steps) or does it primarily enable them to try more distinct, perhaps superficially different, solution paths until one works?
4. Can you try other backbone models like llama3.1-8b-instruct or larger models like qwen-2.5-7B-instruct?

Limitations:

yes

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.

Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
NA

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
The author address my concerns and the work is solid and novel. I wil give a positive score

Rebuttal by Authors

Rebuttal
by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))
30 Jul 2025, 06:53 (modified: 31 Jul 2025, 05:20) Program Chairs, Authors
Revisions (/revisions?id=Q06DZKLszT)

[Deleted]

Rebuttal by Authors

Rebuttal
by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))
31 Jul 2025, 05:20 (modified: 31 Jul 2025, 13:54)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors
Revisions (/revisions?id=f5xZZl8Lhb)

Rebuttal:
Thank you for your review!

Regarding your comment on "limited scope of evaluation across both datasets and models", We have conducted additional experiments using models of different sizes and other open-source architectures.
We also evaluated our method on datasets beyond reasoning tasks, which demonstrates the generalization capability of our UFO framework.
These results will be added to the final version to further strengthen the robustness and applicability of our approach.

Performance Table (Success Rate Across Models and Datasets)

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-1.5B-Instruct	0.109	0.117	0.266	0.219	0.625	0.024	0.031	0.523	0.352
RL on MMQ-Math	0.748	0.201	0.847	0.227	**0.655**	0.192	**0.095**	0.438	0.344
+5turn UFO	**0.836**	**0.268**	**0.881**	**0.273**	0.648	**0.226**	0.093	**0.609**	**0.348**

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-3B-Instruct	0.523	0.283	0.680	0.516	0.758	0.078	0.039	0.752	0.422
RL on MMQ-Math	0.797	0.320	0.930	0.501	0.776	0.195	0.129	0.668	0.483

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
+5turn UFO	**0.885**	**0.408**	**0.953**	0.523	0.875	0.266	0.152	**0.852**	**0.609**
RL on HotQA	0.724	0.318	0.891	0.484	0.813	0.383	0.166	0.715	0.493
+5turn UFO	0.727	0.292	0.850	**0.578**	**0.883**	**0.442**	**0.168**	0.766	0.489

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-7B-Instruct	0.564	0.321	0.563	0.625	0.836	0.133	0.047	0.723	**0.641**
RL on MMQ-Math	0.851	0.336	0.952	0.508	0.848	0.263	0.141	0.734	0.523
+5turn UFO	**0.930**	**0.421**	**0.968**	**0.569**	**0.848**	**0.286**	**0.164**	**0.805**	0.588

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Llama3.2-1B-Instruct	0.023	0.008	0.008	0.008	0.023	0.008	0.000	0.039	0.000
RL on MMQ-Math	0.539	0.211	0.523	0.203	0.570	0.195	0.008	0.578	**0.328**
+5turn UFO	**0.648**	**0.268**	**0.563**	**0.266**	**0.602**	**0.211**	**0.016**	**0.664**	**0.328**

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Llama3.2-3B-Instruct	0.508	0.203	0.484	0.477	0.773	0.297	0.060	0.656	0.492
RL on MMQ-Math	0.867	0.242	0.922	0.469	0.781	0.445	0.133	0.711	0.609
+5turn UFO	**0.922**	**0.320**	**0.938**	**0.508**	**0.820**	**0.398**	**0.148**	**0.828**	**0.664**

We also did experiments by replacing PPO with GRPO:

Model	MMQ	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-1.5B-Instruct	0.109	0.117	0.266	**0.219**	**0.625**	0.024	0.031	0.523	0.352
RL on MMQ-Math	0.750	0.219	0.859	0.180	0.570	0.249	**0.070**	0.523	0.289
+5turnUFO	**0.814**	**0.268**	**0.875**	0.203	0.617	**0.266**	**0.070**	**0.563**	**0.383**

Model	MMQ	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-3B-Instruct	0.523	0.283	0.680	**0.516**	0.758	0.078	0.039	0.752	0.422
RL on MMQ-Math	0.758	**0.352**	0.891	0.320	0.773	0.242	**0.141**	0.670	**0.461**
+5turnUFO	**0.883**	0.348	**0.914**	0.344	**0.781**	**0.320**	0.086	**0.766**	**0.461**

Legend:

- MMQ-Math = Math subset of MetamathQA (Math)
- TQA = TheoremQA (Math)
- GSM8k = GSM8k (Math)
- GPQA = GPQA (Science)
- STEM = MMLU-STEM (Science)
- HotQA = HotPotQA (QA)
- ConQA = ConcurrentQA (QA)
- MMLU = MMLU (General)
- MMLU-Pro = MMLU-Pro (General)

Thank you for raising the question about comparisons with **RAGEN** and **ArCHer**.
First, we would like to clarify that our method is implemented on top of the **RAGEN** framework, as noted in Line 201.

In addition, we experimented with more structured forms of feedback — for example:

> *"Restate the problem in your own words to ensure understanding. Break down the problem into smaller steps, explaining each calculation in detail. Verify each step and re-check your calculations for accuracy. Use proper mathematical notation and maintain consistency with the context of the question. Be cautious at every step."*

Surprisingly, training with such elaborate feedback **did not** result in performance improvements.

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-3B-UFO	0.885	**0.408**	**0.953**	**0.523**	**0.875**	0.266	0.152	**0.852**	**0.609**
Qwen2.5-3B-Complex	**0.906**	0.375	0.938	0.492	0.852	**0.352**	**0.172**	0.797	0.516

This finding further highlights the **efficiency and sufficiency** of our minimal unary feedback setup.

We appreciate your concern about the validity of the performance gains reported. First, as stated in Lines 234–235, we apply the same evaluation criterion in both the parallel and sequential settings, where we record success if any of the 5 responses is correct. This ensures a fair comparison. While S1 employs **majority voting** in their parallel setup, this differs from our success criterion and may not be directly comparable.

Second, in Figure 4, we report results using the same metric ($Succ@k$), clearly showing the improvement brought by our UFO method. Importantly, UFO and multi-turn training is inseparable. When training using single-turn RL, the episode terminates immediately after a reward is given (see Lines 123–132). That is, the agent receives no feedback if $T_{\max} = 1$, making multi-turn RL essential for leveraging feedback effectively.

We answer your questions in the following:

1. Thank you for your suggestion regarding clarity. We agree that the presentation would benefit from a unified section, and we will revise the final version to formally define both the problem setting and the UFO framework together.
2. We appreciate the feedback prompt suggestion. We implemented a variant of multi-turn training without RL, using a heuristic feedback prompt similar to what you described.

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-3B-UFO	**0.885**	**0.408**	**0.953**	**0.523**	**0.875**	**0.266**	**0.152**	**0.852**	**0.609**
Qwen2.5-3B-Heursitic	0.531	0.275	0.688	**0.523**	0.758	0.078	0.039	0.744	0.414

We implemented UFO with RAGEN and it proved to be effective. Further implementation details are provided in the supplementary material.

3. We have included four illustrative cases in the appendix:
   - *Pre-training Behavior*
   - *Post Single-turn RL*
   - *Successful Adaptation via Multi-turn RL with UFO*
   - *Reasoning Drift under Multi-turn RL with UFO*

   These cases were selected to represent both typical and boundary behaviors, highlighting changes before and after training and illustrating both successful and failed adaptation.
4. We agree that evaluating UFO across different backbone models and model sizes strengthens the credibility of our findings. We have accordingly conducted additional experiments, and the results are reported above.

Official Comment by Authors

Official Comment

by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))

04 Aug 2025, 00:11

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer kzS2

Comment:

Dear Reviewer kzS2,

Thank you again for your insightful feedback. We have carefully addressed your comments with detailed, point-by-point revisions and have included new experimental results to better support our main claims.

We would greatly appreciate your thoughts on whether these updates have resolved your concerns. Thank you again for your time and consideration.

Best regards,

The Authors

Replying to Rebuttal by Authors

Official Comment by Reviewer kzS2

Official Comment by Reviewer kzS2 06 Aug 2025, 02:37

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

Comment:

Thanks for addressing my concerns. I will raise my score.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer kzS2

Mandatory Acknowledgement by Reviewer kzS2 06 Aug 2025, 02:37

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

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Replying to Official Comment by Reviewer kzS2

Official Comment by Authors

Official Comment

by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))

06 Aug 2025, 09:16

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

Comment:
Thank you very much for your thoughtful feedback. We truly appreciate your recognition and are grateful for your willingness to raise the score. Your comments have been very helpful in improving our work.

Official Review of Submission6157 by Reviewer s9Qq

(https://openreview.net/forum?id=ocWZ3aKdos&noteId=FTdcRC9F3U)

Official Review by Reviewer s9Qq 11 Jun 2025, 01:13 (modified: 18 Sept 2025, 10:40)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer s9Qq

Revisions (/revisions?id=FTdcRC9F3U)

Summary:
The paper addresses the observed limitation where LLMs trained with single-turn RL often lose their ability to perform multi-turn reasoning, exhibiting repetitive answers and failing to incorporate in-context feedback. To tackle this, the authors propose Unary Feedback as Observation (UFO), a simple method that integrates minimal textual feedback (e.g., "Try Again") into multi-turn RL training on existing static datasets. Experimental results on mathematical reasoning tasks show that UFO helps models recover and improve multi-turn reasoning accuracy by 14% while preserving single-turn performance. The work also explores reward shaping strategies, including reward decay and a repetition penalty, to encourage more diverse and deliberate reasoning.

Strengths And Weaknesses:
Strengths

1. **Simplicity and Practicality:** The proposed UFO is remarkably simple and can be easily integrated into existing single-turn RL training paradigms, enabling multi-turn training on static datasets without costly fine-grained human annotations or complex environments.
2. **Empirical Effectiveness:** The paper demonstrates solid empirical gains, showing that UFO can significantly improve multi-turn reasoning accuracy while maintaining single-turn performance. The reward shaping strategies are also shown to encourage more efficient and diverse problem-solving behaviors.

Weaknesses

1. **Limited Conceptual Novelty of Multi-Turn Feedback:** The advantage shown in Figure 6(a) for explicit feedback, while positive, appears relatively small compared to the overall multi-turn gains demonstrated elsewhere (e.g., Figure 3, which shows 14% gain). This raises questions about the fundamental impact of the "simple multi-turn feedback" itself. The intermediate reward in Equation (9) which penalizes repetition, seems akin to promoting diversity or increasing entropy, making the distinction from single-turn RL with max-entropy unclear. It's not fully elucidated how this "simple multi-turn feedback" differs in essence from a model's long Chain-of-Thought (CoT) processes that might naturally involve self-correction or "wait/but" type internal dialogue.
2. **Insufficient Theoretical Analysis:** The paper is primarily an empirical study, as acknowledged in the NeurIPS checklist. It lacks deeper theoretical analysis explaining why the proposed method, particularly the simple multi-turn interaction, effectively leads to more diverse and self-reflective reasoning. A more formal understanding of the underlying mechanisms would strengthen the contributions, since it feels like single-turn RL with max-entropy.
3. **Preliminary Study Design:** The initial observation that single-turn trained models repeat answers when given a simple "try again" prompt could have been more robustly explored. Not trying more detailed user feedback in this preliminary stage limits the understanding of the specific failure mode, due to lack of multi-turn ability or just simple prompt does not elicit the reasoning?
4. **Limited Model and Data Scope:** The experimental conclusions are primarily based on a single base model (Qwen-2.5-3B-Instruct) and the PPO algorithm. It would be beneficial to evaluate the method's generalizability by testing with other base models and value-free RL algorithms (e.g., GRPO). Furthermore, as experiments are mainly on an in-distribution dataset, validating conclusions on out-of-distribution test sets is crucial for robust generalization.

Quality: 3: good

Clarity: 3: good
Significance: 2: fair
Originality: 2: fair
Questions:

1. You state that single-turn RL policies are "trained without access to interaction history". Could you elaborate on how information from previous turns might implicitly influence single-turn models, for instance, if the entire dialogue history is placed in the context but only the last turn's answer is directly rewarded?
2. Please elaborate on the theoretical or mechanistic reasons why the proposed multi-turn training, particularly with the repetition penalty (Equation 9), encourages more diverse answers. How does this go beyond simply maximizing entropy in the output distribution in traditional single-turn RL, and what specific aspects of "multi-turn" or "feedback as observation" contribute to this diversity?

Limitations:
Yes

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
No major formatting concerns were observed. The paper generally adheres to the NeurIPS format.

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
I have carefully reviewed the authors' rebuttal, which I found to be exceptionally thorough and convincing. The authors have provided substantial new theoretical analysis and empirical experimental results that address my initial major concerns. My recommendation has been updated accordingly from borderline reject to accept. However, I am not so familiar with the theoretical details of the UNO method; the analogy of "Sampling Without Replacement" seems reasonable. And for the method of using minimal yet natural user feedback, I am not sure why it exhibits a great performance gain, and I still doubt it, although I updated the score to 4.

Rebuttal by Authors

Rebuttal
by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))
30 Jul 2025, 23:00 (modified: 31 Jul 2025, 05:21) Program Chairs, Authors
Revisions (/revisions?id=cHyrYjt5Ze)

[Deleted]

Rebuttal by Authors

Rebuttal
by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))
31 Jul 2025, 05:21 (modified: 31 Jul 2025, 13:54)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors
Revisions (/revisions?id=M1Ct62SgXh)

Rebuttal:

Thank you for your review!

Regarding your comment on the **impact of 'simple multi-turn feedback'**, we mentioned that there are 8% peak gain in line 260 and 261. Feedback-enabled models consistently outperform no-feedback baselines by 4–8%; we will clarify this more clearly in the final version.

We thank the reviewer for raising the question regarding the **similarity between our repetition penalty and max-entropy RL**[1]. We answer it below.

Regarding the **difference between UFO and a long CoT**, we think the two differ in several ways:

- CoT relies on the model to spot and fix its own errors, which is often vague and unreliable[2], [3]. In contrast, we give clear, minimal feedback after mistakes to trigger revision directly.
- CoT self-correction is unpredictable and hard to generalize. Our method offers a stable, general correction process that works with rule-base prompts.
- CoT's correction is hidden in reasoning steps, often unclear and hard to spot. Ours makes revision explicit and trackable.

Thank you for raising the concern about **insufficient theoretical analysis**. We now provide a formal justification.

We model multi-turn solving as an MDP (line 120), where each state s_t encodes the question q, the model's past attempts $a_1, \ldots, a_{t-1}$, and the corresponding feedback $f_1, \ldots, f_{t-1}$. Let Π_{par} denote the class of parallel policies and sample k answers i.i.d. from the same distribution only given q, and let Π_{seq} be the class of sequential policies that condition on s_t and can adapt after each failure, as UFO does. Clearly, $\Pi_{\text{par}} \subset \Pi_{\text{seq}}$, since sequential policies can choose to ignore history. It follows that:

$\max_{\pi \in \Pi_{\text{seq}}} \mathbb{E}[\text{Succ@}k] \geq \max_{\pi \in \Pi_{\text{par}}} \mathbb{E}[\text{Succ@}k]$.

To illustrate this, we construct a sequential policy that starts like the parallel policy but adapts based on feedback. At turn t, if a failure occurs, the policy avoids previously errors, increasing the conditional success probability $p'_t \geq p_t$. Thus, the success rate $P_{\text{seq}} = 1 - \prod_t (1 - p'_t)$ is at least as high as that of the parallel policy, by Blackwell dominance[4].

Max-entropy RL, which injects per-step randomness without conditioning on failures, optimizes $\mathbb{E}[R(y|q)] + \alpha \mathcal{H}[\pi(\cdot|q)]$. In contrast, UFO optimizes over trajectories $\mathbb{E}[\sum_t r_t | s_t]$, where the state evolves with interaction history. Here, entropy is applied after conditioning on failure, allowing adaptive rather than random behavior. Our reward further promotes effective reasoning: exponential decay favors earlier success, while $Eq.\ 9$ imposes a global penalty to reduce repetition.

UFO can be seen as approximate **sampling without replacement** over latent reasoning hypotheses $\mathcal{H}$. Each failure narrows the posterior, and the repetition penalty encourages exploring new reasoning paths. In contrast, parallel sampling **samples with replacement**, often repeating failed strategies. This encourages greater diversity and reflection.

In summary, our theoretical analysis shows that:

- Sequential policies strictly dominate parallel ones in expected success.
- Feedback-based adaptation is more expressive than max-entropy.
- Diversity arises not from randomness, but from structured, feedback-aware exploration over the solution space.

Regarding the problem of **'Preliminary Study Design'**, we conducted tests using more detailed prompts:

> *"Incorrect. Try solving it with a different approach."*

> *"Let's try a different way."*

However, we found that richer prompts did not meaningfully change the behavior of single-turn trained models. This suggests the issue lies in a deeper lack of adaptability without multi-turn supervision.

To support our claim, we provide a theoretical analysis:

This behavior arises from repeatedly sampling from a fixed base distribution $q(\cdot|x)$, regardless of prior feedback. For a parallel policy sampling k responses $i.i.d.$ from q, the chance of duplicates is: $\Pr[A_i = A_j] = \sum_y q(y|x)^2$ which increases as q becomes more peaked.

In practice, SFT or RLHF-tuned LLMs often produce sharp output distributions focused on a few high-probability completions. This is due to their training goals:

- SFT minimizes cross-entropy, pushing the model to favor frequent answers.[5]
- RLHF maximizes reward, often collapsing to a few top-ranked outputs.[6]

Both objectives select for policies $\pi(\cdot|x)$ that maximize:

$\mathbb{E}_{y\sim\pi(\cdot|x)}[\log p\mathrm{ref}(y|x)]$ for SFT,

or $\mathbb{E}_{y\sim\pi(\cdot|x)}[r(y)]$ for RLHF.

This results in low-entropy distributions: $\mathcal{H}[\pi(\cdot|x)] = -\sum_y \pi(y|x)\log\pi(y|x)$, Low entropy directly reflects a peaked distribution: most mass is concentrated on a few outputs, increasing the chance of repetition. Hence, the repetition stems from the model's peaked distribution.

We conducted experiments on a larger scope:

PerformanceTable

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-1.5B-Instruct	0.109	0.117	0.266	0.219	0.625	0.024	0.031	0.523	0.352
RLonMMQ-Math	0.748	0.201	0.847	0.227	**0.655**	0.192	**0.095**	0.438	0.344
+5turnUFO	**0.836**	**0.268**	**0.881**	**0.273**	0.648	**0.226**	0.093	**0.609**	**0.348**
--									
Qwen2.5-3B-Instruct	0.523	0.283	0.680	0.516	0.758	0.078	0.039	0.752	0.422
RLonMMQ-Math	0.797	0.320	0.930	0.501	0.776	0.195	0.129	0.668	0.483
+5turnUFO	**0.885**	**0.408**	**0.953**	0.523	0.875	0.266	0.152	**0.852**	**0.609**
RLonHotQA	0.724	0.318	0.891	0.484	0.813	0.383	0.166	0.715	0.493
+5turnUFO	0.727	0.292	0.850	**0.578**	**0.883**	**0.442**	**0.168**	0.766	0.489
--									
Qwen2.5-7B-Instruct	0.564	0.321	0.563	0.625	0.836	0.133	0.047	0.723	**0.641**
RLonMMQ-Math	0.851	0.336	0.952	0.508	0.848	0.263	0.141	0.734	0.523
+5turnUFO	**0.930**	**0.421**	**0.968**	**0.569**	**0.848**	**0.286**	**0.164**	**0.805**	0.588
--									
Llama3.2-1B-Instruct	0.023	0.008	0.008	0.008	0.023	0.008	0.000	0.039	0.000
RLonMMQ-Math	0.539	0.211	0.523	0.203	0.570	0.195	0.008	0.578	**0.328**
+5turnUFO	**0.648**	**0.268**	**0.563**	**0.266**	**0.602**	**0.211**	**0.016**	**0.664**	**0.328**
--									
Llama3.2-3B-Instruct	0.508	0.203	0.484	0.477	0.773	0.297	0.060	0.656	0.492
RLonMMQ-Math	0.867	0.242	0.922	0.469	0.781	0.445	0.133	0.711	0.609
+5turnUFO	**0.922**	**0.320**	**0.938**	**0.508**	**0.820**	**0.398**	**0.148**	**0.828**	**0.664**

We also did experiments by replacing PPO with GRPO:

Model	MMQ-Math	TQA	GSM8k	GPQA	STEM	HotQA	ConQA	MMLU	MMLU-Pro
Qwen2.5-1.5B-Instruct	0.109	0.117	0.266	**0.219**	**0.625**	0.024	0.031	0.523	0.352
RLonMMQ-Math	0.750	0.219	0.859	0.180	0.570	0.249	**0.070**	0.523	0.289
+5turnUFO	**0.814**	**0.268**	**0.875**	0.203	0.617	**0.266**	**0.070**	**0.563**	**0.383**
Qwen2.5-3B-Instruct	0.523	0.283	0.680	**0.516**	0.758	0.078	0.039	0.752	0.422
RLonMMQ-Math	0.758	**0.352**	0.891	0.320	0.773	0.242	**0.141**	0.670	**0.461**
+5turnUFO	**0.883**	0.348	**0.914**	0.344	**0.781**	**0.320**	0.086	**0.766**	**0.461**

Legend:
-MMQ-Math=Math-subset of MetamathQA(Math)
-TQA=TheoremQA(Math)
-GSM8k(Math)
-GPQA(Science)
-STEM=MMLU-STEM(Science)
-HotQA=HotPotQA(QA)
-ConQA=ConcurrentQA(QA)
-MMLU(General)
-MMLU-Pro(General)

We answer your questions in the following:

1. We agree that one could concatenate previous turns into the prompt and reward only the final response. However, it remains a single-turn RL setup under our formulation. The policy receives a static context, **lacks state transitions**, and **optimizes single-step reward objective**: $\max_\pi \mathbb{E}_{y\sim\pi(\cdot|x)}[r(y)]$ regardless of how long the prompt is. As a result, the model lacks trajectory-level learning and **cannot assign credit across interaction steps**. Thus, our analysis of single-turn RL still applies.

In multi-turn RL, the model is trained over full trajectories$(s_1, a_1, r_1, \ldots, s_T, a_T, r_T)$. Even if only the final output r_T receives a scalar reward, we apply Generalized Advantage Estimation (GAE) to compute token-level advantages: $A_t^{\text{GAE}} = \sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}, \quad \text{where } \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$. This allows the model to update earlier actions based on final outcomes, enabling temporal credit assignment that single-turn RL lacks.

2. See above analysis.

---

[1]Onur Celik, Zechu Li, Denis Blessing, Ge Li, Daniel Palenicek, Jan Peters, Georgia Chalvatzaki, and Gerhard Neumann. DIME: Diffusion-Based Maximum Entropy Reinforcement Learning. ICML 2025.

[2]Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. arXiv preprint arXiv:2310.01798, 2023.

[3]Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. TACL 2024.

[4]Blackwell, D. (1951). "Comparison of experiments." Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 93–102.

[5]Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Ruoyu Sun, and Zhi-Quan Luo. Entropic Distribution Matching in Supervised Fine-tuning of LLMs: Less Overfitting and Better Diversity. arXiv preprint arXiv:2408.16673, 2024.

[6]Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu.Understanding the Effects of RLHF on LLM Generalisation and Diversity. arXiv preprint arXiv:2310.06452, 2023.

Official Comment by Authors

Official Comment

by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))

04 Aug 2025, 00:12

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer s9Qq

Comment:

Dear Reviewer s9Qq,

Thank you again for your insightful feedback. We have carefully addressed your comments with detailed, point-by-point revisions and have included new experimental results to better support our main claims.

We would greatly appreciate your thoughts on whether these updates have resolved your concerns. Thank you again for your time and consideration.

Best regards,

The Authors

Replying to Rebuttal by Authors

Official Comment by Reviewer s9Qq

(https://openreview.net/forum?id=ocWZ3aKdos&noteId=wvXmEGt6bz)

Official Comment by Reviewer s9Qq 04 Aug 2025, 22:27

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

Comment:

Thank you for the comprehensive and insightful rebuttal. The new theoretical analysis provides a clear and compelling distinction from max-entropy RL, and the extensive new experiments across multiple models and tasks successfully address my concerns about the limited scope. You have resolved all of my major concerns, and consequently, I have updated my score.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer s9Qq

Mandatory Acknowledgement by Reviewer s9Qq 04 Aug 2025, 22:34

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

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Replying to Official Comment by Reviewer s9Qq

Official Comment by Authors

Official Comment
by Authors (Yiping Lu (/profile?id=~Yiping_Lu1), Zihan Wang (/profile?id=~Zihan_Wang23), Han Liu (/profile?id=~Han_Liu4), Avirup Sil (/profile?id=~Avirup_Sil1), +4 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6157/Authors))
05 Aug 2025, 11:25
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

We sincerely appreciate your encouraging feedback and are delighted that our revisions helped clarify the key distinctions and broaden the scope. Thank you again for your time and for updating your score.