Self-Improving Foundation Models Without Human Supervision

Website: https://sites.google.com/berkeley.edu/selfimprovingfoundationmodels/home

Abstract: As foundation models (FMs) scale, they face a data bottleneck, where the growth of high-quality internet data unable to keep pace with their training needs. This is most apparent with text data already, has been a consistent problem in domains such as embodied intelligence, and is expected to soon inflict other modalities as well. *Self-improvement*, a paradigm where models generate and train on synthetic data generated from the same or other models, offers a promising solution. This paradigm differs from both supervised learning, which relies on curated human data, and reinforcement learning (RL), which depends on external rewards. Self-improvement frameworks require models to self-curate training data, often using imperfect learned verifiers, with unique challenges. This workshop will explore algorithms for self-improvement, covering topics such as synthetic data, multi-agent and multi-modal systems, weak-to-strong generalization, inference-time self-supervision, and theoretical limits.

1. Workshop Summary (expected size: 400-500 attendees; ∼80 accepted papers)

The availability of internet data, while vast, is ultimately finite or at least growing at a pace that lags behind the consumption needs of foundation models (FMs) during pre-training. Perhaps as is most evident with large language models (LLMs), even today, the projected gains from scaling up pre-training on internet data are smaller than incorporating specific test-time techniques [29]. It is projected that soon we will run out of high-quality data, worthy enough to be directly trained on via next-token prediction [34]. Similarly, real robot data in embodied or physical intelligence problems tends to be quite limited to date [24]. All is to say that as FMs scale in size and capability, we will soon hit a ***"data" bottleneck*** blocking progress. To address this, machine learning techniques that enable models to *self-improve*, *i.e.*, continually improve beyond their initial training data become essential. In theory, this can be done by training on self-generated or synthetic data that the same (or other models) produce.

The unique challenges of self-improvement as a learning paradigm. The paradigm of training on self-generated synthetic data, or what we refer to as **self-improvement**, is distinct from standard supervised and reinforcement learning (RL) in several critical ways as we discuss next. These differences underscore the need for a dedicated study of these topics. In supervised learning, models are trained on high-quality annotations from humans. Moreover, for pre-training of LLMs, high-quality data is often curated in heuristic ways that are largely independent of the learning algorithm. In contrast, self-improvement frameworks rely on the model's ability to generate its own training data (or use other models to generate this data), and thus the algorithm for data curation must now be subsumed by the learning framework. RL also involves training on model's generations, and as a result, might appear similar to the self-improvement paradigm. However, due to its generality, a generic RL algorithm (designed to cater to all downstream RL problems) might not be tailored enough for self-improvement, which poses specific constraints and conditions on improving models. For instance, in contrast to an unpredictable external environment, the only randomness in the data generation process for self-improving foundation models in many use cases corresponds to the inherent randomness in the model's own outputs. Furthermore, RL algorithms are typically meant to optimize rewards obtained from an accurate reward oracle, which is absent in the self-improvement paradigm. Here, we can only rely on querying learned verifiers or reward models which

Corresponding author(s): aviralku@andrew.cmu.edu, asetlur@cs.cmu.edu, feryal.mp@gmail.com

can fail arbitrarily. In fact, unless carefully designed, self-improvement recipes can lead to model collapse with more training, which is absent in traditional RL due to the presence of a meaningful reward signal. Thus, different from RL, the self-improvement algorithms cannot naïvely exploit the *verification-generation gap* [1, 30]. This necessitates research on self-improvement algorithms that also adapt to errors made by the learned evaluation model. We believe that such distinctions and specificity should provide far more optimistic and tailored algorithms that are more effective than a generic RL approach.

Goals of the Workshop

This workshop focuses on developing machine learning principles and algorithms for enabling self-improvement in foundation models. We aim to bring together communities working on foundation models, reinforcement learning and online learning, cognitive neuroscience, along with practitioners from various domains for fostering discussions and collaborations on several fundamental topics around this general theme of self-improvement, including but not limited to:

- Learning objectives and algorithms; what should we learn? How should we supervise training?
- Multi-agent and multi-model systems for enabling self-improvement
- Training on machine-generated synthetic data without collapse
- Autonomous online learning and reinforcement learning algorithms for FMs
- Efficiently exploiting tools and external information for self-improvement
- Theoretically characterizing conditions under which self-improvement is feasible, e.g., verification-generation gap, nature of problems where self-improvement is possible,
- Using weak supervision for improving strong models
- Gains from training with self-improvement algorithms at inference time (e.g., computational benefits, performance benefits, etc.)
- Limits of self-improvement training (e.g., when is expert data often needed?)
- **Applications:** software agents, robotic self-improvement, multi-modal systems, math, etc.

We are especially interested in downstream application of self-improvement algorithms. We explicitly encourage submissions that study applications of these algorithms on downstream problem domains. The composition of our speaker and organizer set covers different application areas of interest.

Related Previous Workshops and History.

To the best of our knowledge, our proposed workshop is the only workshop in the last five years that focuses exclusively on the problem of self-improvement with foundation models. Previous workshops at ICLR, NeurIPS and ICML on related topics specifically focus on one type of application or domain, for example workshops pertaining to agents: "open-world agents" and "LLM Agents"; RL and language: "LaReL workshop at NeurIPS 2022; and intersection of math and AI: "MATH-AI".

Perhaps the most closely related workshops to our proposal are the "System 2 reasoning" workshop at NeurIPS 2024 and "Auto RL" workshop at ICML 2024. The System 2 reasoning workshop largely focuses on studying how to imbue transformer-like models with rule-based learning, understanding, syntactic generalization, and compositionality, and while self-improvement algorithms can imbue foundation models with these capabilities, self-improvement is complementary: our goal is to study learning methods that can enable foundation models to improve without human supervision and this improvement may or may not be along these system 2 properties. Conversely, system 2 capabilities can arise from training on well-curated human data, but our goal is to study methods for improving foundation models without human data. The AutoRL workshop focuses on synergies between RL and in-context learning, with a primary focus on investigating how LLMs can help tackle big challenges in RL. Our goal is to instead study self-improvement algorithms for FMs, which is perhaps more related to the opposite direction (building RL methods for improving FMs, but with important distinctions as discussed in the introduction).

Relevance to the ICLR Community and Intellectual Excitement.

There is growing interest in self-improvement techniques, as evidenced by the number of research papers in this area at recent ML conferences (ICML, NeurIPS, and ICLR 2024-2025) and arXiv (*e.g.*, [2–23, 25–27, 27, 28, 31, 32, 32, 33, 35, 36, 36–47]). Furthermore, self-improvement is fundamentally a problem with exciting connections to different paradigms of machine learning (i.e., supervised, unsupervised, reinforcement learning, and meta learning), cognitive neuroscience and game theory, and a diverse range of application domains (natural language reasoning, robotic learning and embodiment, multi-modality). As a result, the workshop will appeal to and benefit from researchers across diverse fields in AI. Many research challenges that we outline are fundamental in nature, and will interest both theorists and empirical researchers. Beyond a rich set of research challenges, recent progress in this area (e.g., with the OpenAI release of o1) holds tremendous promise in building the best FMs, making the workshop attractive to practitioners from industry as well. This makes our workshop a timely one that discusses current progress and decides on important next steps for the research on self-improving FMs.

2. Diversity and Inclusion Statement

The organizing committee shares a strong commitment to promoting diversity in academic backgrounds, gender, and seniority. We strived to achieve a diverse representation in the invited speakers. We will actively encourage female researchers and other minorities in related fields to submit papers and participate in discussions by leveraging existing networks (WiML, Black in AI, LXAI, Queer in AI, *etc.*). Several highlights from the perspective of the proposed workshop are shown below:

- **Diversity of topics:** Our workshop covers a broad range of topics pertaining to self-improvement, including theoretical understanding, algorithm design, applications in language, multi-modal, and embodied problems. We believe that such a diverse portfolio of topics will enable multiple communities to come together and discuss topics at the intersection and for fostering new collaborations. Our speakers focus on different topics / areas as well.
- **Diversity of speakers:** Three of the seven speakers identify as females, while the rest identify as males. All speakers come from different institutions; and are at different levels of seniority in their career. Our speaker set consists of roughly an equal representation from industry and academia. Each speaker is an expert in a different area of research. One speaker is based in Asia.
- **Diversity of organizers:** Three of the six organizers identify as females. The organizers consists of 1 Assistant Professor, 2 PhD students, and 3 Research Scientists at industry labs. Two of these three research scientists hold Adjunct faculty positions at various universities. Two of the organizers are based in Europe, one is based in Canada, two are based on the US east coast, and one is based on the US west coast. Organizers are affiliated with different institutions: CMU, UC Berkeley, Google, Meta, McGill, and UCL. Our speakers come from organizations distinct from organizers. Five of the six organizers have extensive prior workshop organization experience at ICLR, NeurIPS, ICML.

3. Invited Speakers

The following speakers have confirmed their interest in participating at the workshop via an invited talk. We will also invite all invited speakers for a panel discussion.

Speaker	Affiliation	Tentative Talk Topic / Area
Noah Goodman	Stanford University	Self-improvement algorithms
Yejin Choi	Univ. of Washington & NVIDIA	Limits of self-improvement
Phillip Isola	MIT	Synthetic data in computer vision
Shunyu Yao	OpenAI / Princeton	Self-improvement for agents
Minjoon Seo	KAIST AI (Korea)	Self-improvement algorithms
Ida Momennejad	Microsoft Research	Multi-agent learning / cognitive perspectives
Ulyana Piterbarg	New York University	Synthetic data for code

Table 1: List of speakers and panelists, with their affiliations, and tentative topic of their talks.

Yejin Choi *(she/her; NVIDIA & University of Washington)*: Yejin Choi is the Wissner-Slivka Professor at the University of Washington and a senior director at NVIDIA. Her research investigates a wide variety problems across NLP and AI including fundamental limits and capabilities of large language models, neuro-symbolic fusion, commonsense knowledge, reasoning, and self-supervised learning. She is a MacArthur Fellow, named among Time100 Most Influential People in AI in 2023, and a co-recipient of 2 Test-of-Time Awards (ACL 2021 and CVPR 2021) and 8 Best and Outstanding Paper Awards at ICCV, ICML, NeurIPS, ACL, NAACL, EMNLP and AAAI.

Noah Goodman *(he/him; Stanford University)*: Noah Goodman is an Associate Professor of Psychology and CS at Stanford. His research interests lie in computational models of cognition, probabilistic programming languages, natural language semantics and pragmatics, concepts and intuitive theories. Noah is a recipient of the Sloan Fellowship, and multiple best paper awards at venues like AAAI, IEEE Transactions on Affective Computing, EDM, and Cognitive Science Society. His recent contributions (*e.g.*, STAR, Quiet-star, and Stream of Search) to inductive/deductive reasoning, search, RL on LLMs, and self-supervision stand out as some of the most promising avenues in self-improvement.

Ida Momennejad *(she/her; Microsoft Research)*: Ida Momennejad is a Principal Researcher at Microsoft Research. In this role, she focuses on building and evaluating generative AI, drawing from her research in cognitive neuroscience, reinforcement learning, and NeuroAI. She studies how humans and AI build models of the world for memory, exploration, and planning, creating behavior-inspired algorithms for learning and reasoning, such as AI for Xbox gaming. Her approach integrates reinforcement learning, neural networks, large language models, and machine learning with behavioral experiments, fMRI, and electrophysiology. Ida trained in cognitive computational neuroscience through computer science and philosophy. Her PhD was in psychology at the Bernstein Center for Computational Neuroscience.

Phillip Isola *(he/him; Massachusetts Institute of Technology)*: Phillip Isola an Associate Professor in EECS at MIT. He studies computer vision, machine learning, robotics, and AI. He completed his Ph.D. in Brain & Cognitive Sciences at MIT, and has since spent time at UC Berkeley, OpenAI, and Google Research. His work has particularly impacted generative AI and self-supervised representation learning. His research has been recognized by a Google Faculty Research Award, a PAMI Young Researcher Award, a Samsung AI Researcher of the Year Award, a Packard Fellowship, and a Sloan Fellowship. His work spans multi-modality, RL, self-improvement builds on techniques to obtain synthetic data from off-the-shelf foundation, which would be particularly interesting for the workshop attendees.

Ulyana Piterbarg *(she/her; New York University)*: Ulyana is a Ph.D. candidate in the CILVR lab at NYU Courant, co-advised by Rob Fergus and Lerrel Pinto. Her research is supported by a DeepMind Ph.D.

Scholarship and an NSF Graduate Research Fellowship. Ulyana is interested in large generative models that can solve hard tasks in settings like code synthesis, reasoning, decision-making, and open-ended interaction. Before this, she finished obtained her BS in Math at Massachusetts Institute of Technology. Her recent work has been focusing on developing ways to use synthetic data for coding problems, which is relevant to the theme of this workshop.

Shunyu Yao *(he/him; OpenAI & Princeton University)*: Shunyu is currently a researcher at OpenAI and recently completed his PhD from Princeton University where he was advised by Prof. Karthik Narsimhan. His primary focus of research is on LLM agents. Amongst his multiple works on autonomous LLM agents deployed in the wild, his works on SWE-agent and SWE-Bench present strong evaluation protocols for self-improving LLMs meant to solve software engineering problems drawn from real GitHub issues.

Minjoon Seo *(he/him; KAIST AI)*: Minjoon Seo is an Assistant Professor at KAIST AI, where he is the Director of Language & Knowledge Lab. He obtained his PhD in Computer Science at the University of Washington where he was advised by Hannaneh Hajishirzi and Ali Farhadi. He was supported by Facebook Fellowship and AI2 Key Scientific Challenges Award. His interests lie in understanding multimodal intelligence, and in particular studying the learning dynamics of multimodal language models. His works on self-exploration and reasoning in large language models, and some more recent ones on memorization in LLM pretraining are most relevant for the theme of our workshop.

4. Logistics / Planned Timeline

Session 1 (Before lunch)

Time Slot	Planned Event
8:30 - 8:40 am	Opening Remarks
8:40 - 9:15 am	Invited Talk
9:15 - 9:45 am	Contributed Talks (×3)
9:45 - 10:45 am	Poster Session
10:45 - 11:20 am	Invited Talk
11:20 - 11:55 am	Invited Talk
11:55 - 12:30 pm	Invited Talk

Session 2 (After lunch)

Time Slot	Planned Event
2:00 - 2:35 pm	Invited Talk
2:35 - 3:05 pm	Contributed Talks (×3)
3:05 - 4:05 pm	Poster Session
4:05 - 4:40 pm	Invited Talk
4:40 - 5:15 pm	Invited Talk
5:15 - 6:15 pm	Panel Discussion
6:15 - 6:30 pm	Closing Remarks

Figure 1: Tentative schedule for “Self-Improving Foundation Models without Human Supervision”.

Contributed Submissions for Main Track. We plan to accept papers (of around 8-9 pages), excluding references and appendices. We shall not accept work published in prior conferences including ICLR 2025, and will highly encourage submission of working papers. The submission process will be handled via OpenReview, and will go through a double-blind review process. The proceedings of the workshop will be non-archival and will be made available on the workshop website, and will be publicized accordingly on X (formerly twitter), and other social media platforms. We will require authors of accepted papers to provide a recorded video and a slide deck explaining their work. We shall especially encourage submissions concerning applying or describing potential problems and challenges in self-improvement. **We will abide by the mandatory accept / reject deadline of 5th March, 2025 that ICLR prescribes.**

Tiny Papers Track. In accordance with the guidelines from the ICLR workshop organization program, we will also hold a separate track on tiny papers (upto two pages). The goal of this track would be to encourage participants from under-represented minorities to submit a paper. We will follow similar qualification criteria as ICLR 2024: for instance, to be eligible, the first and/or last author on the paper should qualify URM criteria and must self-identify for the same (no reasons or rationale required).

Contributed Talks and Best Paper Award. We plan to divide the accepted papers into two groups with different presentation types – contributed talks (10 min), and posters (two poster sessions) – based on novelty, technical merit, and alignment to the workshop's goals. One paper will be conferred with the workshop's best paper award and will be presented as a contributed talk. We also plan to solicit sponsorship from industry partners (e.g. Meta and Google DeepMind). The funds will be used for awards (best paper and best presentation awards) and ICLR registration grants.

Poster Sessions, Discussion and Workshop Slack. All accepted papers will be presented as posters during **two** poster sessions. We will highly encourage authors to present in both sessions. We do not set the length of a pre-recorded talk for their flexibility. We shall recommend 5 minutes for a concise introduction, or up to 20 minutes for a full discussion, but not exceeding 30 minutes.

Further, we plan to organize a Slack for the workshop. To encourage diverse discussion, we will also create Slack channels with broad topics within self-improvement (such as different application domains and theoretical questions) and seed the discussions with starter thoughts and questions.

Invited Speakers and Panel Discussion. Invited speakers will present their views and perspectives on self-improvement. The speakers will participate in a lively panel discussion about the potential, challenges, and limitations of this field. We will invite audience questions by collecting questions ahead of time and will also enable attendees to ask questions on the spot. The panel would be moderated by two organizers.

Accessibility and Discussion. In order to increase accessibility, we will hold the workshop in a **hybrid format** available for both presenters and attendees of the workshop. In addition, we will release talk summaries, videos, and presentation slides on our workshop website, along with all accepted papers and corresponding posters. We also plan to publish a workshop summary following the event, highlighting the main emerging trends and subjects discussed in our workshop in the form of a blog post. Our workshop slack and discussion groups will remain active throughout to foster discussion among not just participants at ICLR, but also external participants who may be interested in the field.

Advertisement. We will utilize social media (X, LinkedIn, chat forums for the community, mailing lists at different universities and industry labs) to advertise the workshop if accepted. This would not only increase awareness of the work in the community, and hence result in an increase the number of submissions that we receive, but will also increase participation in the workshop.

5. Organizers

The organizing committee consists of both junior and senior members working broadly along the area of self-improvement. Most members having extensive experience in organizing previous workshops at top ML conferences (NeurIPS and ICLR) and other conferences (e.g., in robotics such as RSS).

Amrith Setlur (PhD student at CMU): Amrith Setlur is a 4^{th} year PhD student in the ML department at Carnegie Mellon University, advised by Prof. Virginia Smith. He is supported by the JP Morgan AI PhD Fellowship and his research focuses on robustness to spurious correlations, privacy/memorization, improving generalization through self-improvement techniques, synthetic data, and reinforcement learning for optimizing large language models. He was also a co-organizer for NeurIPS 2023 workshop on R0-FoMo: Workshop on robustness of zero-shot/few-shot learning in foundation models.

Aviral Kumar (Assistant Professor at CMU): Aviral Kumar is an Assistant Professor of Computer Science and Machine Learning at Carnegie Mellon University. He received his Ph.D. from UC Berkeley in 2023. His research focuses on reinforcement learning (RL) broadly, with his most notable work building algorithms for offline reinforcement learning. His recent work focuses on studying the intersection of RL with foundation models, with several works overlapping with self-improvement. He is a receipient of the C.V. & Daulat Ramamoorthy Distinguished Research Award, Facebook and Apple Ph.D. fellowships, and has been named as a semi-finalist for MIT Technology Review 35 Innovators under 35 Award in 2023. He regularly serves as an Area Chair for ICLR and NeurIPS since 2023 and has previously been the primary co-organizer for the NeurIPS workshop on offline RL (link) at NeurIPS from 2020 to 2022 (3 years), which he also helped start in 2020. He has also delivered a tutorial on offline RL at NeurIPS 2020.

Katie Kang (PhD student at UC Berkeley): Katie Kang is a 5^{th} year PhD student at UC Berkeley, advised by Prof. Sergey Levine and Prof. Claire Tomlin. She is supported by a NSF GRFP fellowship. Her research focuses on understanding the generalization behavior of modern ML models, including large language and other foundation models, therefore overlapping with the theme of training on synthetic data.

Feryal Behbahani (Staff Research scientist at Google DeepMind): Feryal Behbahani is a research scientist at Google DeepMind working on reinforcement learning (RL), meta RL, self-improvement, and more broadly learning from sub-optimal data generated by FMs. Previously, she was a research scientist leading the learning team at Latent Logic (now part of Waymo). Before that, she received her PhD in Computational Neuroscience and ML from Imperial College London, where she worked on sensorimotor perception, transfer learning and deep RL. She has experience organizing workshops like the Biological and Artificial RL workshop at NeurIPS 2019, and learning from demonstrations workshop at RSS 2018.

Rishabh Agarwal (Research scientist at Google DeepMind, Adjunct Professor at McGill):Rishabh Agarwal is a research scientist at Google Deepmind and an Adjunct Professor at McGill University working on deep reinforcement learning, often with the goal of making RL methods suitable for real-world problems. His most recent work focuses on mathematical reasoning capabilities of language models and revolves around the theme of using synthetic data to improve capabilities of large models (e.g., improving verifiers, training on suboptimal data, generative verifiers). Previously, he received his PhD in computer science at Mila. His work has received an outstanding paper award at NeurIPS. He serves as an Area Chair for ICLR and COLM. He has organized several prior workshops, including the three editions of the offline RL workshop with Aviral and an ICLR 2023 workshop on reincarnating reinforcement learning (RL).

Roberta Railenau (Research scientist at Meta, Honorary Lecturer at UCL): Roberta Railenau is a research scientist at Meta working on AI agents and tool use as a part of the Llama research team in London, UK. Previously, she received her PhD in computer science at NYU, where she worked on deep reinforcement learning. Her recent work studies methods to train verifiers and better models for code and math using external feedback and self-generated data, which aligns with the theme of this workshop. She has experience organizing workshops like the Generative Models for Decision Making at ICLR 2024, and Agent Learning in Open-Endedness (ALOE) Workshop at ICLR 2022 and NeurIPS 2023.

Potential Reviewers. We plan to invite a number of reviewers for the workshop, by reaching out to graduate students at various schools around the world, and researchers from industry. The set of reviewers will also benefit from organizers' connections and access to mailing lists. We also plan to invite authors of submissions to serve as reviewers to handle an unexpectedly high volume of submissions, following the success of this scheme from some previous workshops a subset of the organizers have organized.

References

[1] Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

[2] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

[3] Eugene Choi, Arash Ahmadian, Matthieu Geist, Oilvier Pietquin, and Mohammad Gheshlaghi Azar. Self-improving robust preference optimization. *arXiv preprint arXiv:2406.01660*, 2024.

[4] Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024.

[5] Zeyu Gan and Yong Liu. Towards a theoretical understanding of synthetic data in llm post-training: A reverse-bottleneck perspective. *arXiv preprint arXiv:2410.01720*, 2024.

[6] Kanishk Gandhi, Denise Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, and Noah D Goodman. Stream of search (sos): Learning to search in language. *arXiv preprint arXiv:2404.03683*, 2024.

[7] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large Language Models can Self-Correct with Tool-Interactive Critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

[8] Noriaki Hirose, Dhruv Shah, Kyle Stachowicz, Ajay Sridhar, and Sergey Levine. Selfi: Autonomous self-improvement with reinforcement learning for social navigation. *arXiv preprint arXiv:2403.00991*, 2024.

[9] Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457*, 2024.

[10] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023.

[11] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

[12] Lei Huang, Xiaocheng Feng, Weitao Ma, Liang Zhao, Yuchun Fan, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. Advancing large language model attribution through self-improving. *arXiv preprint arXiv:2410.13298*, 2024.

[13] Hyeonbin Hwang, Doyoung Kim, Seungone Kim, Seonghyeon Ye, and Minjoon Seo. Self-explore to avoid the pit: Improving the reasoning capabilities of language models with fine-grained rewards. *arXiv preprint arXiv:2404.10346*, 2024.

[14] Xue Jiang, Yihong Dong, Lecheng Wang, Fang Zheng, Qiwei Shang, Ge Li, Zhi Jin, and Wenpin Jiao. Self-planning code generation with large language models. *ACM Transactions on Software Engineering and Methodology*, 2023.

[15] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

[16] Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. Prover-verifier games improve legibility of llm outputs. *arXiv preprint arXiv:2407.13692*, 2024.

[17] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.

[18] Jiacheng Liu, Ramakanth Pasunuru, Hannaneh Hajishirzi, Yejin Choi, and Asli Celikyilmaz. Crystal: Introspective reasoners reinforced with self-feedback. *arXiv preprint arXiv:2310.04921*, 2023.

[19] Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

[20] Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*, 2024.

[21] Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent q: Advanced reasoning and learning for autonomous ai agents. *arXiv preprint arXiv:2408.07199*, 2024.

[22] Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Eleanor Jiang, Chengfei Lv, and Huajun Chen. Autoact: Automatic agent learning from scratch via self-planning. *arXiv preprint arXiv:2401.05268*, 2024.

[23] Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. Recursive introspection: Teaching language model agents how to self-improve. *arXiv preprint arXiv:2407.18219*, 2024.

[24] Nicholas Roy, Ingmar Posner, Tim Barfoot, Philippe Beaudoin, Yoshua Bengio, Jeannette Bohg, Oliver Brock, Isabelle Depatie, Dieter Fox, Dan Koditschek, et al. From machine learning to robotics: Challenges and opportunities for embodied intelligence. *arXiv preprint arXiv:2110.15245*, 2021.

[25] Welleck Sean, Ximing Lu, West Peter, Brahman Faeze, Shen Tianxiao, Khashabi Daniel, and Choi Yejin. Generating sequences by learning to self-correct. *arXiv preprint arXiv: 2211.00053*, 2022.

[26] Amrith Setlur, Saurabh Garg, Xinyang Geng, Naman Garg, Virginia Smith, and Aviral Kumar. Rl on incorrect synthetic data scales the efficiency of llm math reasoning by eight-fold. *arXiv preprint arXiv:2406.14532*, 2024.

[27] Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146*, 2024.

[28] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Synthetic prompting: Generating chain-of-thought demonstrations for large language models. In *International Conference on Machine Learning*, pages 30706–30775. PMLR, 2023.

[29] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

[30] Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the self-verification limitations of large language models on reasoning and planning tasks. *arXiv preprint arXiv:2402.08115*, 2024.

[31] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*, 2023.

[32] Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Haitao Mi, and Dong Yu. Toward self-improvement of llms via imagination, searching, and criticizing. *arXiv preprint arXiv:2404.12253*, 2024.

[33] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15887–15898, 2024.

[34] Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*, 2022.

[35] Ziyu Wan, Xidong Feng, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. Alphazero-like tree-search can guide large language model decoding and training. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=C4OpREezgj.

[36] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, 2024.

[37] Xiyao Wang, Linfeng Song, Ye Tian, Dian Yu, Baolin Peng, Haitao Mi, Furong Huang, and Dong Yu. Towards self-improvement of llms via mcts: Leveraging stepwise knowledge with curriculum preference learning. *arXiv preprint arXiv:2410.06508*, 2024.

[38] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[39] Ting Wu, Xuefeng Li, and Pengfei Liu. Progress or regress? self-improvement reversal in post-training. *arXiv preprint arXiv:2407.05013*, 2024.

[40] Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Xie. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36, 2024.

[41] Yuqing Yang, Yan Ma, and Pengfei Liu. Weak-to-strong reasoning. *arXiv preprint arXiv:2407.13647*, 2024.

[42] Xunjian Yin, Xinyi Wang, Liangming Pan, Xiaojun Wan, and William Yang Wang. Godel agent: A self-referential agent framework for recursive self-improvement. *arXiv preprint arXiv:2410.04444*, 2024.

[43] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

[44] Eric Zelikman, Eliana Lorch, Lester Mackey, and Adam Tauman Kalai. Self-taught optimizer (stop): Recursively self-improving code generation. *arXiv preprint arXiv:2310.02304*, 2023.

[45] Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D Goodman. Quiet-star: Language models can teach themselves to think before speaking. *arXiv preprint arXiv:2403.09629*, 2024.

[46] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024.

[47] Zhiyuan Zhou, Pranav Atreya, Abraham Lee, Homer Walke, Oier Mees, and Sergey Levine. Autonomous improvement of instruction following skills via foundation models. *arXiv preprint arXiv:407.20635*, 2024.