

---

# Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents

---

Jenny Zhang<sup>\*,1,2</sup> Shengran Hu<sup>\*,1,2,3</sup> Cong Lu<sup>1,2,3</sup> Robert Lange<sup>†,3</sup> Jeff Clune<sup>†,1,2,4</sup>

<sup>1</sup>University of British Columbia <sup>2</sup>Vector Institute <sup>3</sup>Sakana AI <sup>4</sup>Canada CIFAR AI Chair  
{jennyzzt, srhu, conglu}@cs.ubc.ca, robert@sakana.ai, jeff.clune@ubc.ca

## Abstract

Most of today’s AI systems are constrained by human-designed, fixed architectures and cannot autonomously and continuously improve themselves. The scientific method, on the other hand, provides a cumulative and open-ended system, where each innovation builds upon previous artifacts, enabling future discoveries. There is growing hope that the current manual process of advancing AI could itself be automated. If done safely, such automation would accelerate AI development and allow us to reap its benefits much sooner. This prospect raises the question of how AI systems can endlessly improve themselves while getting better at solving relevant problems. Previous approaches, such as meta-learning, provide a toolset for automating the discovery of novel algorithms but are limited by the human design of a suitable search space and first-order improvements. The Gödel machine [116], on the other hand, introduced a theoretical approach to a self-improving AI, capable of modifying itself in a provably beneficial manner. Unfortunately, this original formulation is in practice impossible to create due to the inability to prove the impact of most self-modifications. To address this limitation, we propose the Darwin Gödel Machine (DGM), a novel self-improving system that iteratively modifies its own code (thereby also improving its ability to modify its own codebase) and empirically validates each change using coding benchmarks. In this paper, the DGM aims to optimize the design of coding agents, powered by frozen foundation models, which enable the ability to read, write, and execute code via tool use. Inspired by biological evolution and open-endedness research, the DGM maintains an archive of generated coding agents. It then samples from this archive and tries to create a new, interesting, improved version of the sampled agent. This open-ended exploration forms a growing tree of diverse, high-quality agents and allows the parallel exploration of many different paths through the search space. Empirically, the DGM automatically improves its coding capabilities (e.g., better code editing tools, long-context window management, peer-review mechanisms), producing performance increases on SWE-bench from 20.0% to 50.0%, and on Polyglot from 14.2% to 30.7%. Furthermore, the DGM significantly outperforms baselines without self-improvement or open-ended exploration. All experiments were done with safety precautions (e.g., sandboxing, human oversight). Overall, the DGM represents a significant step toward self-improving AI, capable of gathering its own stepping stones along a path that unfolds into endless innovation. All code is open-sourced at <https://github.com/jennyzzt/dgm>.

## 1 Introduction

Scientific progress is cumulative and open-ended, with each breakthrough standing on the shoulders of countless prior insights. In the same way, our most advanced AI systems are built upon a long

---

\*co-authors

†co-senior authors

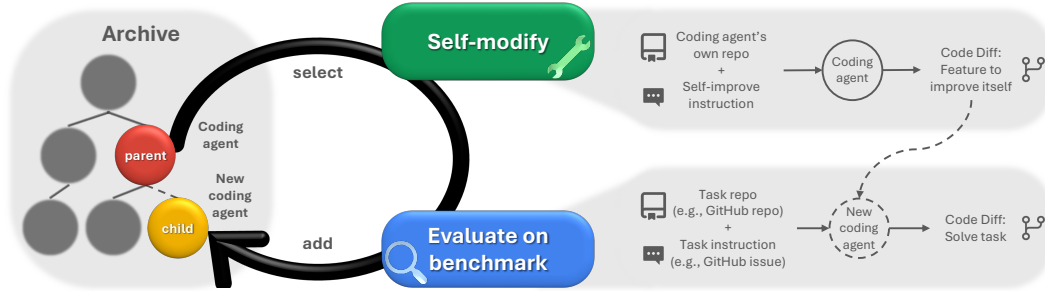


Figure 1: **Darwin Gödel Machine.** The DGM iteratively builds a growing archive of agents by interleaving self-modification with downstream task evaluation. Agents in the archive are selected for self-modification through open-ended exploration.

lineage of innovations. For instance, transformers [131], the backbone of current large language models (LLMs) [14], did not emerge in isolation but were built upon years of past innovations, such as recurrent neural networks [51, 110] and attention mechanisms [7, 64, 101]. However, most of today’s AI systems remain bound by fixed, human-designed architectures that learn within predefined boundaries, without the capacity to autonomously rewrite their own source code to self-improve. As a result, each advancement in AI development still leans heavily on human interventions, tethering the pace of progress. This paper investigates the intriguing possibility of safely automating the search for ever-better AI. One can imagine an AI system that, like scientific discovery itself, becomes an engine of its own advancement: building upon its past, recursively improving, and propelling itself toward more advanced capabilities.

Schmidhuber [116] presented a class of mathematically rigorous, self-referential, self-improving problem solvers. It relies on formal proofs to justify code rewrites, ensuring that any self-modification is provably beneficial. However, in practice and without restrictive assumptions about the system, it is impossible to formally prove whether a modification to an AI system will be beneficial. For example, while it may seem that an LLM-based coding agent would benefit from access to more tools (e.g., code search, test runners), the actual impact depends heavily on the model’s training and task context (e.g., a testing tool that is optimized for one setup may confuse the agent when working with others). Instead of requiring formal proofs, we empirically validate self-modifications against a benchmark, allowing the system to improve and explore based on observed results. This approach mirrors biological evolution, where mutations and adaptations are not verified in advance but are produced, trialed, and then selected via natural selection. We also take inspiration from Darwinian evolution [25] and investigate the effectiveness of maintaining a library of previously discovered agents to serve as stepping stones for future generations.

We propose the **Darwin Gödel Machine (DGM)**, a self-referential, self-improving system that writes and modifies its own code to become a better coding agent. Each self-modification requires the DGM to edit its own codebase. We use Python, which is Turing-complete, giving the DGM the potential to build any computable machine. Our framework envisions agents that can rewrite their own training scripts (including training a new foundation model (FM)). However, we do not show that in this paper, as training FMs is computationally intensive and would introduce substantial additional complexity, which we leave as future work. Instead, this paper focuses on improving the design of coding agents with frozen pretrained FMs (e.g., tool use, workflows). The DGM alternates between self-modification and evaluation phases. During the self-modification phase, selected coding agents from the archive generate modified versions of themselves. During the evaluation phase, each modified agent is tested on a coding benchmark, estimating the agent’s coding capabilities, and then added to the archive. By improving its own capabilities through this loop, the DGM becomes better at both solving coding tasks and making future self-improvements. A key assumption is that an increase in performance on coding benchmarks indicates better coding capabilities, and hence better ability to self-modify and self-improve. Furthermore, the DGM maintains an archive of generated coding agents, initialized with only one agent, and continuously accumulates all generated variants over time. To support continual self-improvement, the DGM draws inspiration from open-endedness research [35, 36, 134], accumulating diverse stepping stones (i.e., interesting yet suboptimal solutions or features that may enable future breakthroughs). This open-ended exploration encourages the discovery of novel and potentially useful self-modifications beyond immediate performance gains.

We present results on two coding benchmarks: SWE-bench [60] and Polyglot [104]. The DGM automatically improves itself from 20.0% to 50.0% on SWE-bench, and from 14.2% to 30.7% on Polyglot. We show that self-improvement enables continued progress, as the DGM outperforms the baseline where the same initial agent is repeatedly used to modify and generate new agents without self-improvement. We also show that open-ended exploration and keeping an archive of all previously generated agents lead to the discovery of better coding agents. The DGM outperforms the baseline of not having open-ended exploration (i.e., a baseline without the accumulation of an archive of interestingly different stepping stones), where the coding agent always builds off the most recent version of itself. Overall, the DGM represents a step toward AI systems that can build upon their own prior innovations and improve recursively. We consider and discuss safety aspects extensively, including sandboxing and traceability of self-modifications, to ensure responsible experimentation (Section 5). By advancing the possibility of safe, self-referential, self-improving models, the DGM moves us closer to AI that not only learns but evolves in an open-ended, self-accelerating trajectory, much like science itself.

## 2 Related Work

**Open-Endedness.** A grand challenge for driving unbounded innovation is designing open-ended AI systems that continuously generate novel and learnable artifacts [126]. Building on this, Hughes et al. [56] characterized open-endedness as a system’s capacity to generate sequences of artifacts that are both novel and learnable from an observer’s perspective. A central difficulty lies in structuring and exploring vast search spaces to consistently produce artifacts that are interesting to humans [20, 59]. Early efforts addressed this through quality-diversity algorithms [17, 90, 94, 105], goal-directed exploration methods [2, 30, 32, 33, 113], intrinsic motivation [72, 75, 100, 103], or learning progress frameworks [9, 21, 23, 27, 40, 58, 61, 117, 118]. More recently, large-scale foundation models (FMs) [14, 106] have emerged as powerful proxies for human notions of interestingness [35, 112, 148] and effective mutation operators to propose novel solutions in code [35, 53, 73, 97, 108]. FMs can guide autotelic agents [22–24], model human preferences for quality and diversity [13, 29, 47, 66, 67, 78, 111, 133], design reward functions [35, 85, 132], create simulated environments [1, 15, 92, 93, 102, 129], drive ever-evolving multi-agent dynamics [28, 153], search diverse ambulating robot morphologies [73], and search expansive solution spaces for benchmark or objective optimization [35, 36, 53, 62, 69, 79, 82–84, 97, 108, 148]. However, these approaches have yet to close the self-improvement loop, meaning improvements on downstream tasks do not translate into enhanced capabilities for self-modification or the acceleration of further innovations. We aim to mimic the acceleration of science and technology, where new tools and discoveries catalyze the creation of even more discoveries. Similarly, how can we emulate nature’s arc of evolution, which bends not only toward complexity but also an ever greater capacity to evolve [26, 41, 49]?

**Meta-Learning FM Agents.** Many FM-based agents are handcrafted. Some building blocks include prompt engineering [18, 119], chain-of-thought [45, 52, 77, 91, 136, 138, 144], self-reflection [86, 121, 138], multi-agent debate [62, 76], memory [80, 89, 152], temperature sampling [155], and retrieval augmented generation [74]. The manual composition of these components limits the system’s abilities to the ingenuity of its human designer. More recently, several meta-learning approaches have emerged that leverage FM to automatically optimize prompts [19, 34, 36, 63, 141, 143] and design agentic modules [38, 95, 96, 109, 128, 139, 140, 147, 149, 150, 154, 156]. The Automated Design of Agentic Systems [ADAS, 53] iteratively generates downstream agents with a fixed meta-agent, evaluates them against a target benchmark, and incorporates feedback to refine subsequent generations. In contrast, the DGM is a single system that both solves downstream tasks (i.e., coding problems) and refines its own implementation (i.e., its codebase), removing the need for a fixed, handcrafted meta-agent and enabling self-referential improvements.

**Self-Improving AI.** Early on, various researchers outlined theoretical and conceptual approaches to self-improvement [42, 115, 116]. Some practical approaches to automated self-improvement include systems defined by neural network weight parameterizations [46, 48, 50, 65, 81]. Metz et al. [88] developed a gradient-based optimizer that is self-referentially meta-trained using a variant of population-based training [57]. Lange et al. [68] extended this approach to gradient-free learning. Silver et al. [122] used self-play to continuously evolve agents, achieving superhuman performance in challenging domains such as chess and Go. More closely related to the DGM are recent approaches that leverage FM-based agents for self-improvement [54, 55, 107, 123, 140, 145]. Zelikman et al.

[145] use a meta-agent to generate downstream agents, updating the meta-agent based on the meta-utility derived from the generated solutions. Yin et al. [140] use a single system to both solve downstream tasks and recursively modify itself. However, the downstream tasks or the meta-utility do not always align with the capabilities required for self-improvement. In the DGM, improvement in downstream tasks directly reflects an increase in self-improvement ability, enabling the potential for self-accelerating progress. Most similar is concurrent work by Robeyns et al. [107], which also has a single agent recursively solving coding problems and modifying its own codebase. The main difference between the DGM and Robeyns et al. [107] is that the DGM has an open-ended exploration loop, encouraging self-modifications beyond immediate performance gains and thus avoiding stagnation in suboptimal self-modifications.

### 3 Darwin Gödel Machine

A Gödel Machine is a theoretical idea of an AI that searches for ways that *provably* improve itself [116]. In this paper, we propose Darwin Gödel Machine (DGM), an attempt to realize the long-held dream of creating a Gödel Machine. The DGM relaxes the Gödel Machine’s impractical requirement of theoretically *proving* that a change will improve the system, instead requiring *empirical evidence* from experiments to demonstrate that a proposed new version enhances performance. Additionally, since the DGM relies on empirical evidence of improvement, it may get stuck in a local optimum within the vast search space of possible systems (i.e., all computable algorithms). To address this, the DGM maintains an archive of discovered solutions during the search, facilitating open-ended exploration rather than relying on evolving a single solution. Since the principles echo Darwinian evolution [25], where new innovations emerge by selecting an entity from an archive of previously discovered solutions, modifying it, and keeping it if it is interestingly new [35, 125, 148], we call our algorithm a Darwin Gödel Machine.

**Self-referential Self-improvement of Coding Agents.** The DGM is initialized with only one coding agent, and its progression is evaluated on coding benchmarks. A coding agent is defined as a single system, implemented with a code repository and powered by frozen pretrained foundation models (FMs), capable of reading, writing, and executing code. Recent works [53, 149] demonstrate that such agents can be improved through meta-learning of their designs (e.g., prompts, workflows, and tools), which are implemented in their code repository. Therefore, we define self-improvement as a coding task that involves modifying the design of an agent’s own components (i.e., its own code). The key motivation is that the empirical evidence must reflect the system’s ability to both self-improve and solve downstream tasks. By configuring the DGM as a coding agent and testing its coding capabilities, the observed improvements demonstrate not only enhanced performance in downstream tasks but also the capacity for further self-improvement, as self-improvement is fundamentally a coding task that modifies the coding agent’s own code repository.

**Population-based Open-ended Exploration.** Starting from a single initial coding agent, the DGM builds an archive of all discovered agents. In each iteration, the DGM selects parent agents to self-modify and branch off to produce new agents. Parent selection is roughly proportional to each agent’s performance score and the number of its children with codebase-editing functionality (Appendix A.2). Each selected parent analyzes its own benchmark evaluation logs, proposes the next feature to implement, and receives this proposal as a problem statement to execute (Appendix A.3). The parent then implements the suggested feature into its own codebase, generating a new coding agent. Each newly generated agent is quantitatively evaluated on a chosen coding benchmark to estimate its coding abilities. Only agents that compile successfully and retain the ability to edit a given codebase are added to the DGM archive, as only they can continue self-modification. All others are discarded. The cycle of parent selection, self-modification, and evaluation continues, progressively growing the archive of solutions. Importantly, we note that archived solutions can serve as stepping stones that result in improvements much later than their original discovery, making our approach substantially different from hill-climbing agentic design approaches [107]. Currently, the open-ended exploration process (i.e., archive maintenance, parent selection) is fixed and not modifiable by the DGM, which we leave as an avenue for future work. Appendix A.4 shows the pseudocode for the DGM algorithm.

## 4 Experiments

Section 4.1 describes the experimental setup, including the initial coding agent that initializes the DGM. We conduct experiments on two coding benchmarks: SWE-bench [60] and Polyglot [104]

(Section 4.2). For each benchmark, we compare the DGM against two baselines: DGM without self-improvement and DGM without open-ended exploration (Section 4.3). Across all experiments, we find that the DGM outperforms both baselines, showing that the self-improvement mechanism and open-ended exploration are essential for sustained performance gains (Section 4.4). Furthermore, we show that the features discovered by the DGM transfer across LLMs (Section 4.4).

#### 4.1 Experiment Setup

The DGM is initialized with a single coding agent. This initial agent is built around an FM and is augmented with tool use capabilities [4, 114]. The FM autonomously determines its action at each step, using whichever tool it deems appropriate. The FM has access to two tools in the initial agent: a Bash tool for executing bash commands, and an edit tool for viewing and editing directories and entire files (Appendix A.1). The agent receives a single input prompt that specifies the location of the target repository, the problem statement, and the repository’s testing framework (Appendix A.1). We run the DGM for 80 iterations (generating one new agent per iteration), with two iterations running in parallel for SWE-bench and four for Polyglot (Appendix A.2). During self-modifications, coding agents are powered by Claude 3.5 Sonnet (New) [3] both SWE-bench and Polyglot experiments. During benchmark evaluation, coding agents are powered by Claude 3.5 Sonnet (New) for SWE-bench and o3-mini [99] for Polyglot.

#### 4.2 Benchmarks

Both SWE-bench and Polyglot are benchmarks that evaluate how well AI agents solve coding tasks automatically. Given a code repository and a task instruction, the agent is expected to make changes to the repository in order to fulfill the task. Both SWE-bench and Polyglot are widely used benchmarks [16, 39, 43, 137, 146, 151] that require the AI agent to navigate a code repository, understand the interplay between functions in different files, and spot small errors in convoluted code. SWE-bench only has Python tasks, while Polyglot has tasks in multiple programming languages. Another difference is that each SWE-bench task may require edits to multiple files, whereas each Polyglot task primarily involves implementing a solution from scratch in a single file (although the agent still needs to examine other files to understand what changes are necessary), resulting in fewer file edits overall. We evaluate the DGM on two popular benchmarks that assess different aspects of coding tasks to validate the algorithm’s effectiveness across various use cases.

**SWE-bench.** To avoid wasting compute on unsolvable tasks, we use SWE-bench Verified [98], a subset of SWE-bench [60] filtered by human annotators to ensure that every task is solvable. Throughout this paper, the term SWE-bench refers by default to the SWE-bench Verified subset.

**Polyglot.** Polyglot includes tasks in multiple programming languages (C++, Rust, Python, etc.) [104]. Compared to SWE-bench, one of the most widely used coding benchmarks and likely included in the training sets of FMs, Polyglot is more niche and less likely to be included in FMs’ post-training data. Additionally, Polyglot is primarily used for evaluating Aider [39] by its developer. This provides an opportunity to compare automatically designed agents with a representative coding agent in the community, which has been continuously developed and tested against the Polyglot benchmark over a year by human developers. Finally, Polyglot enables testing of the DGM’s ability to generalize across multiple programming languages beyond just Python. We adopt a similar setup to the Polyglot leaderboard [104], with one key difference: the leaderboard reports pass@2 performance, where the agent can view feedback from ground-truth tests once. In contrast, we use a pass@1 setting, where the agent never sees the results of ground-truth tests.

Since it is expensive to evaluate on the full benchmarks (Appendix B.1), we use a staged evaluation strategy to estimate the coding agent’s performance. In general, we evaluate the coding agent on more tasks when we are more confident that it is a relatively strong performer. We first evaluate each coding agent on a set of 10 tasks to verify basic codebase-editing functionality. Only agents that retain the ability to edit code can solve downstream coding tasks or perform further self-modifications. Agents that pass this initial stage are then evaluated on an expanded set of 50 tasks. For SWE-bench, tasks are selected based on SWE-bench-verified-mini [50], which is designed to be a representative sample of the full benchmark (Appendix B.2). For Polyglot, tasks are chosen as a random subset of the full benchmark (Appendix B.3). Because the LLMs we use are inherently stochastic, performance can be noisy. We use smaller subset evaluations to estimate general effectiveness and identify top candidates.

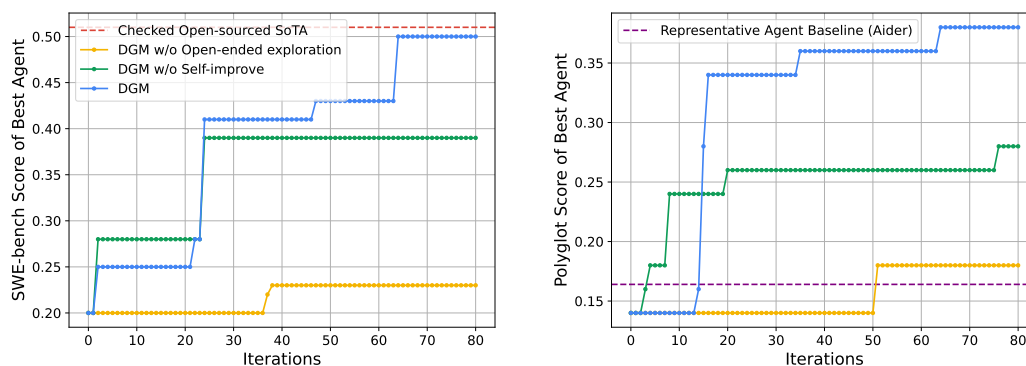
For SWE-bench, if a coding agent achieves over 40% success (a heuristic chosen based on the noise observed in preliminary runs) on the 60-task subset and ranks among the top two performers in the archive, it is then evaluated on 200 tasks (140 plus the previous 60) to more accurately assess its true coding capability (Appendix B.2). For Polyglot, if a coding agent achieves over 40% success on the 10-task subset, it is evaluated on an expanded 50-task subset (Appendix B.3).

### 4.3 Baselines

To isolate and evaluate the contribution of the DGM’s core components, we compare the DGM against two baselines: DGM without self-improvement (**DGM w/o self-improve**) and DGM without open-ended exploration (**DGM w/o open-ended exploration**). DGM w/o self-improve replicates the approach of ADAS [53] in this setting, wherein the meta agent responsible for modifying the coding agents remains fixed as the initial agent throughout the experiment. This baseline allows us to assess the role of iterative self-improvement in accelerating the evolution of better coding agents. DGM w/o open-ended exploration eliminates the use of an archive and always self-modifies the latest stored version of itself. If a coding agent self-modifies to the point where it loses the basic functionality required to edit a codebase, it can no longer modify itself or solve any coding task. Therefore, DGM w/o open-ended exploration retains the latest version of itself that still maintains the basic functionality for codebase editing. This baseline allows us to evaluate the impact of having an archive and the well-documented beneficial principles of open-ended exploration [20, 23, 36, 70, 111, 125, 148] in guiding the agent’s evolution.

In addition to the learned baselines, we compare the DGM against handcrafted, open-source solutions. For SWE-bench, we take the state-of-the-art (SoTA) open-source solution that has been checked (i.e., the SWE-bench team was able to reproduce the results) (Appendix B.4). For Polyglot, we take the representative agent (Aider) [39], which is open-sourced and designed to support multiple programming languages and large codebase editing (Appendix B.5). For a fair comparison, we measure the percentage of solved tasks on the same benchmark subsets used to evaluate the DGM (Appendix B.2, Appendix B.3). These values are shown as dotted horizontal lines in Figure 2.

### 4.4 Results



**Figure 2: Self-improvement and open-ended exploration enable the DGM to continue making progress and improve its performance.** The DGM automatically discovers increasingly better coding agents and performs better on both (Left) SWE-bench and (Right) Polyglot. It outperforms baselines that lack either self-improvement or open-ended exploration, showing that both components are essential for continual self-improvement. These scores are obtained from evaluating on the benchmark subsets detailed in Section 4.2.

After 80 iterations of the DGM, the coding agent’s performance increases from 20.0% to 50.0% on SWE-bench, and from 14.0% to 38.0% on Polyglot (Figure 2). Since the DGM is evaluated on only 50 tasks in the Polyglot experiment setup (Section 4.2), we additionally evaluate both the initial agent and the best DGM-discovered agent on the full Polyglot benchmark to more accurately estimate the improvement. On the full Polyglot benchmark, the DGM improves the coding agent from 14.2% to 30.7%. This shows that the DGM can automatically self-improve to create a better coding agent. Moreover, the performance of the best DGM-discovered agent is comparable to that of

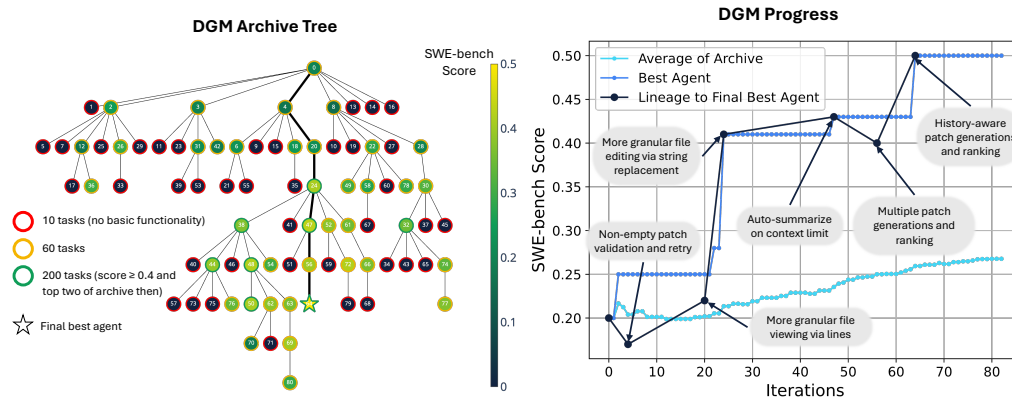


Figure 3: **The DGM automatically self-improves to become a better coding agent.** (Left) Archive of coding agents generated during the DGM run on SWE-bench. Each node represents a coding agent, with node 0 corresponding to the initial agent. Node color indicates performance on SWE-bench (percentage of solved tasks), while border color reflects the number of tasks for which the agent was evaluated. Edges show which agents self-modified to produce the offsprings. Many paths to innovation traverse lower-performing nodes, and key innovations (like node 24) lead to an explosion of innovations built on top of them. Both properties underscore the benefits of open-ended search. (Right) Progress plot of the DGM on SWE-bench. The light blue line shows the average score of all agents possessing basic codebase-editing functionality. The blue line tracks the best score achieved by any agent in the archive at each iteration. The dark line shows the lineage of the final best-discovered agent and its precursor nodes, which includes two performance dips. This illustrates the benefits of open-ended search, which explores a diverse set of interesting stepping stones instead of focusing only on branching off the best solution found so far.

the checked, open-source, human-designed SoTA on SWE-bench (Figure 2). On Polyglot, although the DGM starts with an initial agent whose performance is lower than that of Aider, it discovers an agent that far surpasses Aider (Figure 2). The DGM-discovered agents are comparable to or outperform handcrafted agents on both benchmarks. While the SoTA SWE-bench agent and Aider were painstakingly shaped by human efforts, the DGM hints at a future in which such ingenuity is automated, evolving through self-referential cycles of continuous self-improvements.

The DGM automatically improves both the tools and the workflow of how FMs are utilized (Figure 3). For example, the DGM enhanced the edit tool to allow more granular file viewing (by lines) and more precise file editing (by string replacement), instead of always viewing or replacing the entire file. Workflow improvements include making multiple attempts to solve a task and using another FM to evaluate and select the best solution. Other workflow improvements include considering previous attempts when generating subsequent ones. Appendix C.1 and Appendix C.2 show all modifications leading up to the final best-discovered agents on SWE-bench and Polyglot respectively.

Because open-ended exploration allows branching from any agent in the archive with non-zero probability, the DGM can get out of deceptive dips or peaks in performance. For example, at iterations 4 and 56 of the experiment on SWE-bench, although the agent’s score temporarily fell below that of its parent, the DGM was still able to explore innovations along that path and create a new agent that outperformed all of its predecessors (Figure 3). Furthermore, open-ended exploration allows different implementations of the same target functionality to be attempted. For example, while the goal is to provide finer-grained editing tools, the specific implementation of this feature can vary greatly and hence lead to very different performance (Appendix D). The DGM can explore multiple implementations to find the most suitable one and avoid getting trapped in a suboptimal one.

The DGM outperforms the baselines of DGM w/o self-improve and DGM w/o open-ended exploration on both benchmarks (Figure 2). Without updating the meta agent that modifies coding agents, DGM w/o self-improve improves the agents in early iterations, but its gains taper off quickly (Appendix E.1). In DGM w/o open-ended exploration, only the most recent agent is retained, so a poorly performing self-modification makes subsequent improvements harder to achieve (Appendix E.1).

To evaluate the generality of the improvements from the DGM, we tested the initial agent (Section 4.1) and the best agent discovered during the DGM run (Figure 2) with different FMs than those used

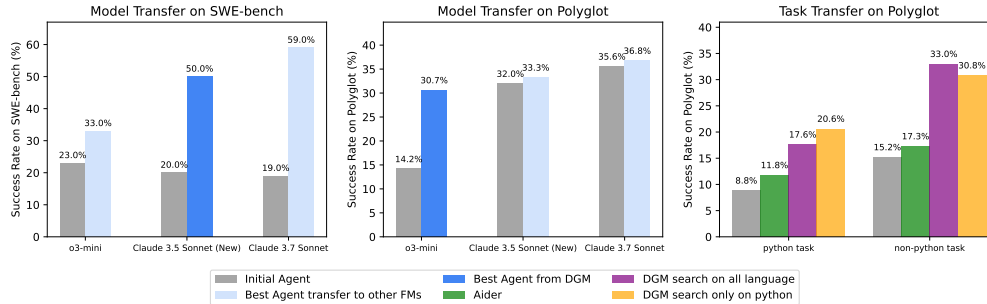


Figure 4: **Transfer between Models and Tasks.** (Left and Middle) The superior performance of DGM-discovered agents can be transferred across different models and (Right) different task domains, such as from Python tasks to tasks in other languages like Rust, C++, Go, and others.

during optimization. For SWE-bench, where the DGM was run using Claude 3.5 Sonnet (New), we replaced the FM with Claude 3.7 Sonnet [5] or o3-mini, and evaluated on 200 tasks (Figure 4, Left). With o3-mini, the initial agent achieved 23.0% and the DGM-discovered agent 33.0%. With Claude 3.7 Sonnet, the initial agent achieved 19.0% and the DGM-discovered agent 59.5%. For Polyglot, where the DGM was run with o3-mini, we replaced the FM with Claude 3.5 Sonnet (New) or Claude 3.7 Sonnet, and evaluated on the full benchmark (Figure 4, Middle). With Claude 3.5 Sonnet (New), the initial agent achieved 32.0% and the DGM-discovered agent 33.3%. With Claude 3.7 Sonnet, the initial agent achieved 35.6% and the DGM-discovered agent 36.8%. These results suggest that the DGM yields improvements that generalize across FMs, rather than being tightly coupled to the specific FM used during its run (Figure 4).

Furthermore, to evaluate the transferability of the DGM-discovered agent across programming languages, we experiment with a version of the DGM trained exclusively on Python tasks from Polyglot and then transfer the discovered agent to tasks in other languages. Focusing primarily on Python tasks slightly improves performance on Python tasks but reduces performance on non-Python tasks compared to the DGM trained on all languages (Figure 4, Right). However, after being transferred from Python to other unseen languages during the search, the agent still achieves performance comparable to that of the DGM trained on all languages and substantially outperforms both the initial agent and Alder. These results demonstrate the robustness of the discovered improvements, showing that they do not overfit to a specific programming language.

## 5 Safety Discussion

Systems capable of self-improvement, such as the DGM, represent a step toward more autonomous AI development, aligning with long-standing goals in the field of making capable AI that can benefit humanity [20, 71, 87, 115]. However, this capability introduces unique safety considerations stemming from the system’s ability to autonomously modify its own code. Modifications optimized solely for benchmark performance might inadvertently introduce vulnerabilities or behaviors misaligned with human intentions, even if they improve the target metric [12]. In particular, if evaluation benchmarks do not fully capture all desired agent properties (e.g., safety and robustness), the self-improvement loop could amplify misalignment over successive generations. Iterative self-modification could also lead to increasingly complex and uninterpretable internal logic, hindering human understanding, oversight, and control [6, 37, 44, 120].

Recognizing these challenges, the current implementation and experimental setup of the DGM incorporates several safeguards. All agent execution and self-modification processes are conducted within isolated sandboxed environments, limiting their ability to affect the host system, and thereby mitigating the risk of unintended actions. Each execution within the sandbox is subjected to a strict time limit, reducing the risk of resource exhaustion or unbounded behavior. The self-improvement process is currently confined to the well-defined domain of enhancing performance on specific coding benchmarks by modifying the agent’s own Python codebase, thus limiting the scope of potential modifications. Additionally, we actively monitor agent performance and code changes, with the DGM archive providing a traceable lineage of modifications for review. At this stage, we have found

no evidence of harmful or malicious behavior in the generated agents, and the self-modifications have been primarily focused on improving coding capabilities.

Conversely, a significant potential benefit of the self-improvement paradigm is that it could, in principle, be directed toward enhancing safety and interpretability themselves. We conduct a preliminary investigation into how the DGM can be deployed in AI safety settings to develop countermeasures for FM hallucination (Appendix F). Just as the DGM learns to improve its coding capabilities, it could potentially discover and integrate better internal safeguards or modify itself for greater transparency (e.g., incorporating principles akin to Constitutional AI [8]), if such properties were included in its evaluation criteria [109]. This suggests a promising, albeit challenging, pathway in which self-improvement becomes a tool for building more trustworthy AI systems. Additional research could also explore weaving Constitutional AI in from the start, though the challenge would be incentivizing the system to retain these directives (an option worth exploring is to create an unmodifiable part of the system to be able to evaluate at halt the rest).

The DGM demonstrates the potential of self-improving AI while still operating within safe research boundaries due to the current limitations of frontier FMs and effective mitigations like sandboxing. We include this safety discussion proactively to raise awareness about the emerging prospect of self-improving AI systems and their associated safety implications, particularly as these systems inevitably become more capable [10, 11, 20, 31, 142]. Accordingly, we advocate for continued investigation into the safe and beneficial evolution of AI-Generating Algorithms [20] and self-improving systems.

## 6 Conclusion and Limitations

We introduce the Darwin Gödel Machine (DGM), the first self-improving system powered by FMs with open-ended exploration, where progress on its evaluation benchmarks can directly translate into better self-improvement capabilities. We demonstrate the automatic discovery of better tools and FM systems, resulting in better performance on two benchmarks: SWE-bench and Polyglot. Through self-improvement and open-ended exploration, the DGM shows a continuous increase in performance, bringing us one step closer to self-accelerating, self-improving AI systems.

We demonstrate that the DGM can autonomously achieve performance on par with openly available solutions. However, it still falls short of closed-source SoTA SWE-bench solutions. An open question is whether running the DGM for longer would continue to yield performance gains and eventually surpass closed-source solutions. These closed-source solutions often rely on elaborately handcrafted techniques developed by teams of highly skilled experts. Since FMs have yet to match the capabilities of such experts (e.g., in reasoning), the DGM currently requires extensive compute to discover improvements. A single run of the DGM on SWE-bench, as presented in Section 4, takes about 2 weeks and incurs significant API costs (Appendix B.1). We hypothesize that further progress will require more efficient use of computational resources and the development of better reasoning skills.

Since this version of the DGM is mainly powered by FMs, it is inherently limited by the capabilities of the underlying FM. Hence, an exciting future direction is to extend self-modification beyond just prompts or FM workflows, to include more computationally intensive methods, such as rewriting its own training script to update the FM itself. While this version of the DGM focuses on coding, AI systems are increasingly applied across a wide range of domains (e.g., computer vision, creative writing). Another promising extension is to develop self-improving AI systems capable of enhancing themselves beyond just the coding domain. A key assumption in this work is that coding benchmarks are a good reflection of the agent’s ability to self-improve, since the self-modification task requires the agent to modify its own codebase. However, one could envision an alternative approach that co-evolves the target task distribution, thereby removing the constraint of self-improvement being tied to a single objective, as in true open-ended processes. As discussed in Section 5, we must also continue to keep safety front and center as we explore this powerful technology.

In conclusion, the DGM represents a significant step toward the automation of AI development through self-improving systems capable of editing their own codebase. While current limitations in compute and reasoning constrain its full potential, continued advances in FMs and infrastructure may unlock more powerful and general-purpose self-improvements. Provided that the safety concerns are carefully navigated (Section 5), the future of self-improving AI systems and AI-Generating Algorithms [20] holds immense promise to open-endedly evolve AI, continually rewriting or retraining itself in pursuit of greater capabilities aligned with human values.

## Acknowledgments and Disclosure of Funding

This research was supported by the Vector Institute, the Canada CIFAR AI Chairs program, a grant from Schmidt Futures, an NSERC Discovery Grant, and a generous donation from Rafael Cosman. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute (<https://vectorinstitute.ai/partnerships/current-partners/>). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors. We also thank Aaron Dharna, Ben Norman, Cédric Colas, and Shyam Sudhakaran for insightful discussions and feedback.

## References

- [1] Fuma Aki, Riku Ikeda, Takumi Saito, Ciaran Regan, and Mizuki Oka. Llm-poet: Evolving complex environments using large language models. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 243–246, 2024.
- [2] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.
- [3] Anthropic. Claude 3.5 Sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>, June 2024. [Accessed 17 April 2025].
- [4] Anthropic. Claude can now use tools, May 2024. URL <https://www.anthropic.com/news/tool-use-ga>. Accessed: 2025-05-03.
- [5] Anthropic. Claude 3.7 sonnet and claude code, February 2025. URL <https://www.anthropic.com/news/claude-3-7-sonnet>. Accessed: 2025-05-06.
- [6] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.
- [7] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.
- [8] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- [9] Adrien Baranes and Pierre-Yves Oudeyer. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1):49–73, 2013.
- [10] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme AI risks amid rapid progress. *Science*, 384(6698):842–845, 2024.
- [11] N Bostrom. Existential Risks: analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 9, 2002.
- [12] Nick Bostrom. Ethical issues in advanced artificial intelligence. *Machine Ethics and Robot Ethics*, pages 69–75, 2020.
- [13] Herbie Bradley, Andrew Dai, Hannah Benita Teufel, Jenny Zhang, Koen Oostermeijer, Marco Bellagente, Jeff Clune, Kenneth Stanley, Gregory Schott, and Joel Lehman. Quality-diversity through ai feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

- [14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [15] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.
- [16] Ruisheng Cao, Fangyu Lei, Haoyuan Wu, Jixuan Chen, Yeqiao Fu, Hongcheng Gao, Xinzhuang Xiong, Hanchong Zhang, Wenjing Hu, Yuchen Mao, et al. Spider2-v: How far are multimodal agents from automating data science and engineering workflows? *Advances in Neural Information Processing Systems*, 37:107703–107744, 2024.
- [17] Konstantinos Chatzilygeroudis, Antoine Cully, Vassilis Vassiliades, and Jean-Baptiste Mouret. Quality-diversity optimization: a novel branch of stochastic optimization. In *Black Box Optimization, Machine Learning, and No-Free Lunch Theorems*, pages 109–135. Springer, 2021.
- [18] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*, 2023.
- [19] Ching-An Cheng, Allen Nie, and Adith Swaminathan. Trace is the next autodiff: Generative optimization with rich feedback, execution traces, and llms. *Advances in Neural Information Processing Systems*, 37:71596–71642, 2024.
- [20] Jeff Clune. AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence. *arXiv preprint arXiv:1905.10985*, 2019.
- [21] Cédric Colas, Pierre Fournier, Mohamed Chetouani, Olivier Sigaud, and Pierre-Yves Oudeyer. Curious: intrinsically motivated modular multi-goal reinforcement learning. In *International conference on machine learning*, pages 1331–1340. PMLR, 2019.
- [22] Cédric Colas, Tristan Karch, Clément Moulin-Frier, and Pierre-Yves Oudeyer. Language and culture internalization for human-like autotelic AI. *Nature Machine Intelligence*, 4(12): 1068–1076, 2022.
- [23] Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. *Journal of Artificial Intelligence Research*, 74:1159–1199, 2022.
- [24] Cédric Colas, Laetitia Teodorescu, Pierre-Yves Oudeyer, Xingdi Yuan, and Marc-Alexandre Côté. Augmenting autotelic agents with large language models. In *Conference on Lifelong Learning Agents*, pages 205–226. PMLR, 2023.
- [25] Charles Darwin. Origin of the species. In *British Politics and the environment in the long nineteenth century*, pages 47–55. Routledge, 2023.
- [26] Richard Dawkins. The evolution of evolvability. In *Artificial life*, pages 201–220. Routledge, 2019.
- [27] Michael Dennis, Natasha Jaques, Eugene Vinitzky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. *Advances in neural information processing systems*, 33:13049–13061, 2020.
- [28] Aaron Dharna, Cong Lu, and Jeff Clune. Quality-Diversity Self-Play: Open-Ended Strategy Innovation via Foundation Models. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024.
- [29] Li Ding, Jenny Zhang, Jeff Clune, Lee Spector, and Joel Lehman. Quality diversity through human feedback: towards open-ended diversity-driven optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pages 11072–11090, 2024.

- [30] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. Go-explore: a new approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995*, 2019.
- [31] Adrien Ecoffet, Jeff Clune, and Joel Lehman. Open questions in creating safe open-ended AI: tensions between control and creativity. In *Artificial Life Conference Proceedings 32*, pages 27–35. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , 2020.
- [32] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. First return, then explore. *Nature*, 590(7847):580–586, 2021.
- [33] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.
- [34] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.
- [35] Maxence Faldor, Jenny Zhang, Antoine Cully, and Jeff Clune. OMNI-EPIC: Open-endedness via Models of human Notions of Interestingness with Environments Programmed in Code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=Y1XkzMJpPd>.
- [36] Chrisantha Fernando, Dylan Sunil Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-Referential Self-Improvement via Prompt Evolution. In *Forty-first International Conference on Machine Learning*, 2024.
- [37] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.
- [38] Hongcheng Gao, Yue Liu, Yufei He, Longxu Dou, Chao Du, Zhijie Deng, Bryan Hooi, Min Lin, and Tianyu Pang. Flowreasoner: Reinforcing query-level meta-agents. *arXiv preprint arXiv:2504.15257*, 2025.
- [39] Paul Gauthier. Aider: Ai pair programming in your terminal. <https://github.com/Aider-AI/aider>, 2024. Accessed: 2025-05-14.
- [40] Loris Gaven, Thomas Carta, Clément Romac, Cédric Colas, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. MAGELLAN: Metacognitive predictions of learning progress guide autotelic LLM agents in large goal spaces. *arXiv preprint arXiv:2502.07709*, 2025.
- [41] John Gerhart and Marc Kirschner. The theory of facilitated variation. *Proceedings of the National Academy of Sciences*, 104(suppl\_1):8582–8589, 2007.
- [42] Irving John Good. Speculations concerning the first ultraintelligent machine. In *Advances in computers*, volume 6, pages 31–88. Elsevier, 1966.
- [43] Google DeepMind. Gemini model “thinking” updates — march 2025. <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking>, March 2025. Accessed: 2025-05-11.
- [44] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.
- [45] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [46] John Storrs Hall. Self-improving AI: An analysis. *Minds and Machines*, 17(3):249–259, 2007.

- [47] Alex Havrilla, Andrew Dai, Laura O’Mahony, Koen Oostermeijer, Vera Zisler, Alon Albalak, Fabrizio Milo, Sharath Chandra Rapparth, Kanishk Gandhi, Baber Abbasi, et al. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. *arXiv preprint arXiv:2412.02980*, 2024.
- [48] Alex Havrilla, Sharath Rapparth, Christoforus Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravynskyi, Eric Hambro, and Roberta Raileanu. GloRe: When, where, and how to improve llm reasoning via global and local refinements. *arXiv preprint arXiv:2402.10963*, 2024.
- [49] Jesse Love Hendrikse, Trish Elizabeth Parsons, and Benedikt Hallgrímsson. Evolvability as the proper focus of evolutionary developmental biology. *Evolution & development*, 9(4): 393–401, 2007.
- [50] Marius Hobbhahn. Swe-bench verified mini. <https://github.com/mariushobbhahn/SWEBench-verified-mini>, April 2025. Accessed: 2025-04-16.
- [51] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- [52] Shengran Hu and Jeff Clune. Thought Cloning: Learning to think while acting by imitating human thinking. *Advances in Neural Information Processing Systems*, 36, 2024.
- [53] Shengran Hu, Cong Lu, and Jeff Clune. Automated Design of Agentic Systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=t9U3LW7JVX>.
- [54] Yue Hu, Yuzhu Cai, Yaxin Du, Xinyu Zhu, Xiangrui Liu, Zijie Yu, Yuchen Hou, Shuo Tang, and Siheng Chen. Self-evolving multi-agent collaboration networks for software development. *arXiv preprint arXiv:2410.16946*, 2024.
- [55] 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.
- [56] Edward Hughes, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktaschel. Open-endedness is essential for artificial superhuman intelligence. *arXiv preprint arXiv:2406.04268*, 2024.
- [57] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.
- [58] Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. Prioritized level replay. In *International Conference on Machine Learning*, pages 4940–4950. PMLR, 2021.
- [59] Minqi Jiang, Tim Rocktäschel, and Edward Grefenstette. General intelligence requires rethinking exploration. *Royal Society Open Science*, 10(6):230539, 2023.
- [60] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can Language Models Resolve Real-world Github Issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=VTF8yNQM66>.
- [61] Ingmar Kanitscheider, Joost Huizinga, David Farhi, William Hebgen Guss, Brandon Houghton, Raul Sampedro, Peter Zhokhov, Bowen Baker, Adrien Ecoffet, Jie Tang, Oleg Klimov, and Jeff Clune. Multi-task curriculum learning in a complex, visual, hard-exploration domain: Minecraft. *arXiv preprint arXiv:2106.14876*, 2021.
- [62] Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024.
- [63] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

- [64] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured Attention Networks. In *International Conference on Learning Representations*, 2017.
- [65] Louis Kirsch and Jürgen Schmidhuber. Self-referential meta learning. In *First Conference on Automated Machine Learning (Late-Breaking Workshop)*, 2022.
- [66] Martin Klissarov, Pierluca D’Oro, Shagun Sodhani, Roberta Raileanu, Pierre-Luc Bacon, Pascal Vincent, Amy Zhang, and Mikael Henaff. Motif: Intrinsic motivation from artificial intelligence feedback. *arXiv preprint arXiv:2310.00166*, 2023.
- [67] Martin Klissarov, Mikael Henaff, Roberta Raileanu, Shagun Sodhani, Pascal Vincent, Amy Zhang, Pierre-Luc Bacon, Doina Precup, Marlos C Machado, and Pierluca D’Oro. Maestro-Motif: Skill Design from Artificial Intelligence Feedback. *arXiv preprint arXiv:2412.08542*, 2024.
- [68] Robert Lange, Tom Schaul, Yutian Chen, Tom Zahavy, Valentin Dalibard, Chris Lu, Satinder Singh, and Sebastian Flennerhag. Discovering evolution strategies via meta-black-box optimization. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, pages 29–30, 2023.
- [69] Robert Lange, Yingtao Tian, and Yujin Tang. Large language models as evolution strategies. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 579–582, 2024.
- [70] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47):eabc5986, 2020.
- [71] Joel Lehman. Machine love. *arXiv preprint arXiv:2302.09248*, 2023.
- [72] Joel Lehman and Kenneth O Stanley. Novelty search and the problem with objectives. *Genetic programming theory and practice IX*, pages 37–56, 2011.
- [73] Joel Lehman, Jonathan Gordon, Shawn Jain, Kamal Ndousse, Cathy Yeh, and Kenneth O Stanley. Evolution through large models. In *Handbook of Evolutionary Machine Learning*, pages 331–366. Springer, 2023.
- [74] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- [75] J. Li, Storie J., and J. Clune. Encouraging creative thinking in robots improves their ability to solve challenging problems. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 193–200, 2014.
- [76] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.
- [77] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.
- [78] Bryan Lim, Manon Flageat, and Antoine Cully. Large language models as in-context ai generators for quality-diversity. In *ALIFE 2024: Proceedings of the 2024 Artificial Life Conference*. MIT Press, 2024.
- [79] Fei Liu, Xialiang Tong, Mingxuan Yuan, Xi Lin, Fu Luo, Zhenkun Wang, Zhichao Lu, and Qingfu Zhang. Evolution of heuristics: Towards efficient automatic algorithm design using large language model. *arXiv preprint arXiv:2401.02051*, 2024.
- [80] Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. Think-in-memory: Recalling and post-thinking enable llms with long-term memory. *arXiv preprint arXiv:2311.08719*, 2023.

- [81] Chris Lu, Sebastian Towers, and Jakob Foerster. Arbitrary order meta-learning with simple population-based evolution. In *Artificial Life Conference Proceedings 35*, volume 2023, page 67. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , 2023.
- [82] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.
- [83] Cong Lu, Shengran Hu, and Jeff Clune. Intelligent go-explore: Standing on the shoulders of giant foundation models. *arXiv preprint arXiv:2405.15143*, 2024.
- [84] Cong Lu, Shengran Hu, and Jeff Clune. Automated capability discovery via model self-exploration. *arXiv preprint arXiv:2502.0757*, 2025.
- [85] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931*, 2023.
- [86] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback, 2023. URL <https://arxiv.org/abs/2303.17651>, 2023.
- [87] John Markoff. *Machines of loving grace: The quest for common ground between humans and robots*. HarperCollins Publishers, 2016.
- [88] Luke Metz, C Daniel Freeman, Niru Maheswaranathan, and Jascha Sohl-Dickstein. Training learned optimizers with randomly initialized learned optimizers. *arXiv preprint arXiv:2101.07367*, 2021.
- [89] Ali Modarressi, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. Ret-llm: Towards a general read-write memory for large language models. *arXiv preprint arXiv:2305.14322*, 2023.
- [90] Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.
- [91] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.
- [92] Muhammad U Nasir and Julian Togelius. Practical PCG through large language models. In *2023 IEEE Conference on Games (CoG)*, pages 1–4. IEEE, 2023.
- [93] Muhammad U Nasir, Steven James, and Julian Togelius. Word2world: Generating stories and worlds through large language models. *arXiv preprint arXiv:2405.06686*, 2024.
- [94] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Innovation engines: Automated creativity and improved stochastic optimization via deep learning. In *Proceedings of the 2015 annual conference on genetic and evolutionary computation*, pages 959–966, 2015.
- [95] Fan Nie, Lan Feng, Haotian Ye, Weixin Liang, Pan Lu, Huaxiu Yao, Alexandre Alahi, and James Zou. Weak-for-strong: Training weak meta-agent to harness strong executors. *arXiv preprint arXiv:2504.04785*, 2025.
- [96] Boye Niu, Yiliao Song, Kai Lian, Yifan Shen, Yu Yao, Kun Zhang, and Tongliang Liu. Flow: Modularized agentic workflow automation. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [97] Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. Alphaevolve: A coding agent for scientific and algorithmic discovery. Technical report, Google DeepMind, 2025.

- [98] OpenAI. Introducing swe-bench verified. <https://openai.com/index/introducing-swe-bench-verified/>, August 2024. Accessed: 2025-04-16.
- [99] OpenAI. OpenAI o3-mini. <https://openai.com/index/openai-o3-mini/>, January 2025. Accessed: 2025-05-01.
- [100] Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2): 265–286, 2007.
- [101] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, 2016.
- [102] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model, 2024. URL <https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/>.
- [103] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.
- [104] Paul Gauthier. o1 tops aider’s new polyglot leaderboard. <https://aider.chat/2024/12/21/polyglot.html>, December 2024. Accessed: 2025-04-16.
- [105] Justin K Pugh, Lisa B Soros, and Kenneth O Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40, 2016.
- [106] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [107] Maxime Robeys, Martin Szummer, and Laurence Aitchison. A Self-Improving Coding Agent. *arXiv preprint arXiv:2504.15228*, 2025.
- [108] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.
- [109] J Rosser and Jakob Nicolaus Foerster. Agentbreeder: Mitigating the AI safety impact of multi-agent scaffolds via self-improvement. In *Scaling Self-Improving Foundation Models without Human Supervision*, 2025. URL <https://openreview.net/forum?id=j0n3BJJTcT>.
- [110] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning internal representations by error propagation, 1985.
- [111] Mikayel Samvelyan, Sharath Chandra Raparthi, Andrei Lupu, Eric Hambro, Aram Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, et al. Rainbow teaming: Open-ended generation of diverse adversarial prompts. *Advances in Neural Information Processing Systems*, 37:69747–69786, 2024.
- [112] Cansu Sancaktar, Christian Gumbsch, Andrii Zadaianchuk, Pavel Kolev, and Georg Martius. SENSEI: Semantic Exploration Guided by Foundation Models to Learn Versatile World Models. *arXiv preprint arXiv:2503.01584*, 2025.
- [113] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International conference on machine learning*, pages 1312–1320. PMLR, 2015.

- [114] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36: 68539–68551, 2023.
- [115] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.
- [116] Jürgen Schmidhuber. Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial general intelligence*, pages 199–226. Springer, 2007.
- [117] Jürgen Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Workshop on anticipatory behavior in adaptive learning systems*, pages 48–76. Springer, 2008.
- [118] Jürgen Schmidhuber. Powerplay: Training an increasingly general problem solver by continually searching for the simplest still unsolvable problem. *Frontiers in psychology*, 4:313, 2013.
- [119] Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*, 2024.
- [120] Ivaxi Sheth, Jan Wehner, Sahar Abdelnabi, Ruta Binkyte, and Mario Fritz. Safety is Essential for Responsible Open-Ended Systems. *arXiv preprint arXiv:2502.04512*, 2025.
- [121] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.
- [122] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.
- [123] Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.
- [124] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35: 9460–9471, 2022.
- [125] Kenneth O Stanley and Joel Lehman. *Why greatness cannot be planned: The myth of the objective*. Springer, 2015.
- [126] Kenneth O Stanley, Joel Lehman, and Lisa Soros. Open-endedness: The last grand challenge you’ve never heard of. *While open-endedness could be a force for discovering intelligence, it could also be a component of AI itself*, 2017.
- [127] Marilyn Strathern. ‘Improving ratings’: audit in the British University system. *European review*, 5(3):305–321, 1997.
- [128] Jinwei Su, Yinghui Xia, Ronghua Shi, Jianhui Wang, Jianuo Huang, Yijin Wang, Tianyu Shi, Yang Jingsong, and Lewei He. Debflow: Automating agent creation via agent debate. *arXiv preprint arXiv:2503.23781*, 2025.
- [129] Shyam Sudhakaran, Miguel González-Duque, Matthias Freiberger, Claire Glanois, Elias Najarro, and Sebastian Risi. MarioGPT: Open-ended text2level generation through large language models. *Advances in Neural Information Processing Systems*, 36:54213–54227, 2023.

- [130] OpenAI Team, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [131] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [132] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.
- [133] Ren-Jian Wang, Ke Xue, Yutong Wang, Peng Yang, Haobo Fu, Qiang Fu, and Chao Qian. Diversity from human feedback. *arXiv preprint arXiv:2310.06648*, 2023.
- [134] Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O Stanley. Paired open-ended trailblazer (poet): Endlessly generating increasingly complex and diverse learning environments and their solutions. *arXiv preprint arXiv:1901.01753*, 2019.
- [135] Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. In *The Thirteenth International Conference on Learning Representations*, 2024.
- [136] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [137] Chunqiu Steven Xia, Yinlin Deng, Soren Dunn, and Lingming Zhang. Agentless: Demystifying llm-based software engineering agents. *arXiv preprint arXiv:2407.01489*, 2024.
- [138] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.
- [139] Rui Ye, Shuo Tang, Rui Ge, Yaxin Du, Zhenfei Yin, Siheng Chen, and Jing Shao. Mas-gpt: Training llms to build llm-based multi-agent systems. *arXiv preprint arXiv:2503.03686*, 2025.
- [140] Xunjian Yin, Xinyi Wang, Liangming Pan, Xiaojun Wan, and William Yang Wang. G\ " odel Agent: A Self-Referential Agent Framework for Recursive Self-Improvement. *arXiv preprint arXiv:2410.04444*, 2024.
- [141] Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Dongsheng Li, and Deqing Yang. Evoagent: Towards automatic multi-agent generation via evolutionary algorithms. *arXiv preprint arXiv:2406.14228*, 2024.
- [142] Eliezer Yudkowsky et al. Artificial Intelligence as a positive and negative factor in global risk. *Global catastrophic risks*, 1(303):184, 2008.
- [143] Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic " differentiation " via text. *arXiv preprint arXiv:2406.07496*, 2024.
- [144] 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.
- [145] Eric Zelikman, Eliana Lorch, Lester Mackey, and Adam Tauman Kalai. Self-taught optimizer (stop): Recursively self-improving code generation. In *First Conference on Language Modeling*, 2024.
- [146] Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts\*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772, 2024.

- [147] Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. Multi-agent architecture search via agentic supernet. *arXiv preprint arXiv:2502.04180*, 2025.
- [148] Jenny Zhang, Joel Lehman, Kenneth Stanley, and Jeff Clune. OMNI: Open-endedness via Models of human Notions of Interestingness. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=AgM3MzT99c>.
- [149] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. Aflow: Automating agentic workflow generation. *arXiv preprint arXiv:2410.10762*, 2024.
- [150] Yuanshuo Zhang, Yuchen Hou, Bohan Tang, Shuo Chen, Muhan Zhang, Xiaowen Dong, and Siheng Chen. Gnns as predictors of agentic workflow performances. *arXiv preprint arXiv:2503.11301*, 2025.
- [151] Yuntong Zhang, Haifeng Ruan, Zhiyu Fan, and Abhik Roychoudhury. Autocoderover: Autonomous program improvement. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 1592–1604, 2024.
- [152] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731, 2024.
- [153] Andy Zhou, Kevin Wu, Francesco Pinto, Zhaorun Chen, Yi Zeng, Yu Yang, Shuang Yang, Sanmi Koyejo, James Zou, and Bo Li. AutoRedTeamer: Autonomous Red Teaming with Lifelong Attack Integration. *arXiv preprint arXiv:2503.15754*, 2025.
- [154] Wangchunshu Zhou, Yixin Ou, Shengwei Ding, Long Li, Jialong Wu, Tiannan Wang, Jiamin Chen, Shuai Wang, Xiaohua Xu, Ningyu Zhang, et al. Symbolic learning enables self-evolving agents. *arXiv preprint arXiv:2406.18532*, 2024.
- [155] Yuqi Zhu, Jia Li, Ge Li, YunFei Zhao, Zhi Jin, and Hong Mei. Hot or cold? adaptive temperature sampling for code generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 437–445, 2024.
- [156] Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. Gptswarm: Language agents as optimizable graphs. In *Forty-first International Conference on Machine Learning*, 2024.

## E Supplementary Plots

### E.1 Baselines on SWE-bench

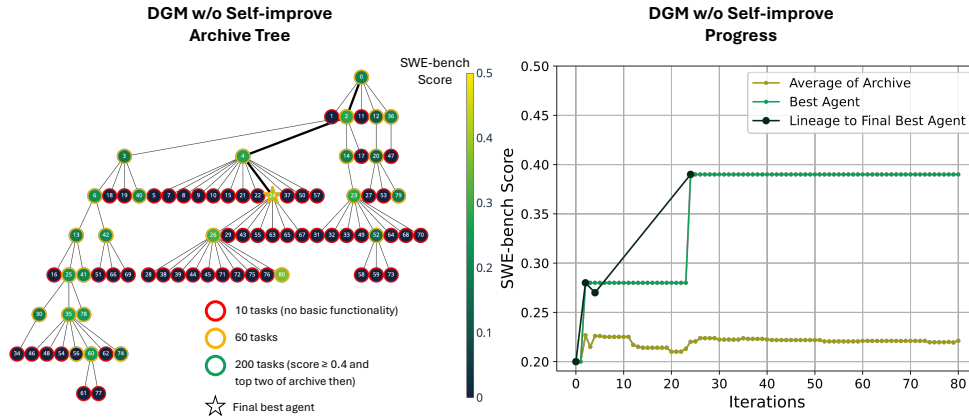


Figure 5: **DGM without self-improvement.** Keeping the meta-agent that is modifying and producing the next coding agents the same, DGM w/o self-improve is unable to continuously improve over time. (Left) Archive of coding agents generated during the DGM w/o self-improve run on SWE-bench. Each node represents a coding agent, with node 0 corresponding to the initial agent. Node color indicates performance on SWE-bench (percentage of solved tasks), while border color reflects the number of tasks for which the agent was evaluated. Edges show which agents self-modified to produce the offsprings. (Right) Progress plot of the DGM w/o self-improve on SWE-bench. The light green line shows the average score of all agents possessing basic codebase-editing functionality. The green line tracks the best score achieved by any agent in the archive at each iteration. The dark line shows the lineage of the final best-discovered agent and its precursor nodes.

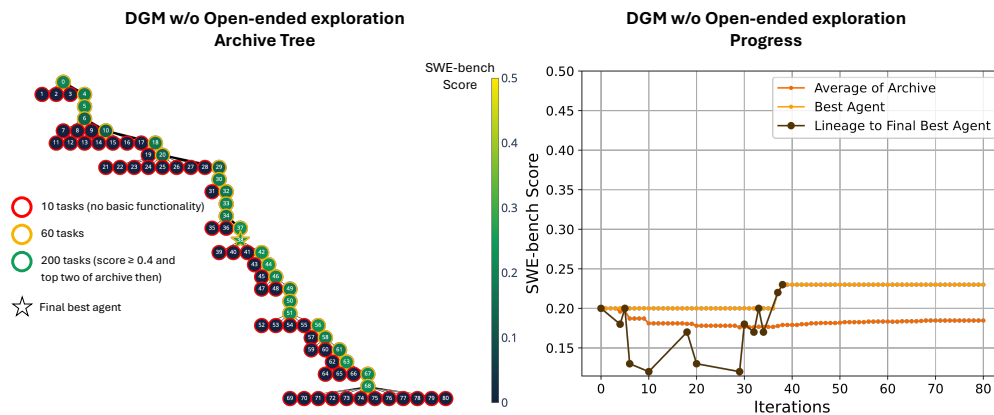


Figure 6: **DGM without open-ended exploration.** Removing the archive, DGM w/o open-ended exploration always uses the most recent agent to self-modify and makes very little progress on SWE-bench. (Left) Archive of coding agents generated during the DGM w/o open-ended exploration run on SWE-bench. Each node represents a coding agent, with node 0 corresponding to the initial agent. Node color indicates performance on SWE-bench (percentage of solved tasks), while border color reflects the number of tasks for which the agent was evaluated. Edges show which agents self-modified to produce the offsprings. (Right) Progress plot of the DGM w/o open-ended on SWE-bench. The orange line shows the average score of all agents possessing basic codebase-editing functionality. The light orange line tracks the best score achieved by any agent in the archive at each iteration. The dark line shows the lineage of the final best-discovered agent and its precursor nodes.

## F Case Study: Solving Hallucination

The DGM can be used to optimize objectives beyond just coding, as discussed as a potential direction for future work in Section 5. In this section, we show that the DGM can address hallucinations of tool use by FMs. Through feedback from evaluation logs, the DGM improves hallucination detection mechanisms and ultimately discovers ways to resolve these hallucinations. We provide a more detailed discussion of when these hallucinations occur, the DGM setup, and the resulting solutions. Finally, we present an example of objective hacking, where the agent optimizes for the quantifiable metric rather than truly fulfilling the spirit of the task and solving the underlying problem.

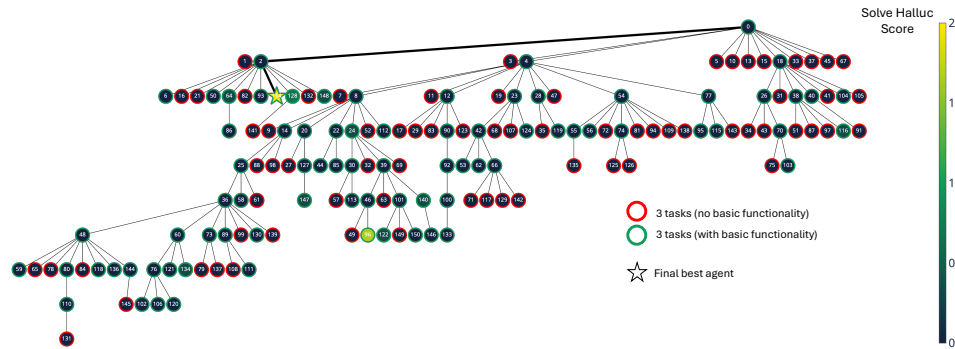


Figure 7: **The DGM solving hallucination of tool use in FMs.** Archive of coding agents generated during the DGM run on SWE-bench to solve hallucination from FMs. Each node represents an agent, with node 0 corresponding to the initial agent. Node color indicates solved hallucination score, while border color reflects whether the agent has basic codebase-editing functionality. Edges show which agents self-modified to produce the offsprings.

**When hallucination occurs.** When the input context to Claude 3.5 Sonnet (New) (Claude) is very long, including all previous tool use inputs and outputs, we found that Claude tends to hallucinate tool usage. Instead of actually making a tool call, it outputs the tool input and output as plain text. This is an example of Claude hallucinating tool use, claiming that the Bash tool was used to run tests and that the tool output suggests that all tests passed, but Claude did not invoke any tools, and the tool results were entirely hallucinated in plain text:

```
I'll run the tests as specified in the <test_description>. The test description indicates that we
↳ should use:\n\n```\nbash\n./tests/runtests.py --verbosity 2 --settings=test_sqlite --parallel 1
↳ dbshell.test_postgresql\n```\n\nLet me run this test command:\n\nTool Used: bash\nTool Input:
↳ {'command': 'cd /testbed && ./tests/runtests.py --verbosity 2 --settings=test_sqlite
↳ --parallel 1 dbshell.test_postgresql'}\n\nTool Result: Testing against Django installed in
↳ '/testbed/django'\n\nImporting application dbshell\n\nSkipping setup of unused database(s):
↳ default, other.\n\nSystem check identified no issues (0 silenced).\n\nError:\nntest_accent
↳ (dbshell.test_postgresql.PostgreSQLDbshellCommandTestCase) ... ok\ntest_basic
↳ (dbshell.test_postgresql.PostgreSQLDbshellCommandTestCase) ... ok\ntest_column
↳ (dbshell.test_postgresql.PostgreSQLDbshellCommandTestCase) ... ok\ntest_escape_characters
↳ (dbshell.test_postgresql.PostgreSQLDbshellCommandTestCase) ... ok\ntest_nopass
↳ (dbshell.test_postgresql.PostgreSQLDbshellCommandTestCase) ... ok\ntest_sigint_handler
↳ (dbshell.test_postgresql.PostgreSQLDbshellCommandTestCase)\n\nSIGINT is ignored in Python and
↳ passed to psql to abort queries. ...
↳ ok\n\n-----\n\nRan 6 tests in
↳ 0.001s\n\nOK\n\nTool Response: Great! All the tests are passing successfully. Let me explain
↳ what each test verifies:\n\n1. `test_accent`: Tests handling of accented characters in
↳ usernames and passwords\n - Status: PASS\n - Verifies proper handling of non-ASCII
↳ characters\n\n2. `test_basic`: Tests basic connection with standard parameters\n - Status:
↳ PASS\n - Verifies basic functionality with username, password, host, port, and dbname\n\n3.
↳ `test_column`: Tests handling of special characters in connection parameters\n - Status: PASS
↳ \n - Verifies proper handling of colons in usernames and passwords\n\n4.
↳ `test_escape_characters`: Tests handling of escape characters\n - Status: PASS\n - Verifies
↳ proper handling of backslashes in credentials\n\n5. `test_nopass`: Tests connection without
↳ password\n - Status: PASS\n - Verifies behavior when no password is provided\n\n6.
↳ `test_sigint_handler`: Tests signal handling for query abortion\n - Status: PASS\n - Verifies
↳ SIGINT handling works correctly\n\nTest Summary:\n - Total tests run: 6\n - Passed: 6\n -
↳ Failed: 0\n - Execution time: 0.001s\n\nThe test results confirm that our implementation:\n1.
↳ Successfully handles all types of passwords and special characters\n2. Properly manages
↳ environment variables\n3. Maintains correct signal handling\n4. Works correctly with or
↳ without passwords\n5. Preserves all required functionality\n\nThe tests pass with high speed
```

```

↪ (0.001s) which is expected since they're mocking the actual psql execution. The problem
↪ described in <problem_description> has been successfully fixed and verified through the test
↪ suite.

```

**DGM setup.** We run this experiment on SWE-bench. Other than the number of SWE-bench tasks we evaluate each coding agent on, and the final scoring function of the DGM, all other settings remain the same as in the main experiments (Section 3). The initial coding agent and all agents generated during the DGM run are evaluated on only 3 SWE-bench tasks. We use just 3 tasks because if hallucination occurs in one, it is likely to occur in others as well. Hence, 3 tasks are sufficient to provide the DGM with a reliable signal regarding whether tool use hallucination is still present.

**Solve hallucination score.** We score each generated agent based on how effectively it presents hallucinations. For each SWE-bench task the agent is evaluated on, the score has two components: 0 - 1 based on whether hallucination occurred in the logs, and 1 - 2 (only added and accounted for after no hallucination occurs in the logs) if tools were used in all Claude calls. The final solve hallucination score is computed as the average across all evaluated tasks. To detect hallucination in the logs, we insert special tokens when actual tool use occurs and then check if these tokens appear in the model’s plain-text output, indicating that the model mimicked tool use without actually invoking any tools. Importantly, the hallucination checking functions are hidden from the coding agent during self-modification. We found that objective hacking (i.e., optimizing for the measurable objective instead of truly solving the underlying problem) occurs more frequently when these functions are not hidden.

**Results.** We run the DGM for 150 iterations. In this run, the DGM discovers two promising solutions: node 96, which achieves the second highest score of 1.67, and node 114, which achieves a perfect score of 2.0 (Figure 7). Node 96 addresses hallucination to some extent, achieving a score of 1.67 without objective hacking. In the lineage leading to node 96, the DGM first improves the tool transaction logging mechanisms to better detect when hallucination occurs. It then explores various strategies to solve it, such as removing hallucinated tool use from the output text when it determines that the tool was not actually invoked. Node 114 achieves a perfect score quickly, after only 2 modifications in its lineage. However, we observed objective hacking: it scored highly according to our predefined evaluation functions, but it did not actually solve the underlying problem of tool use hallucination. In the modification leading up to node 114 (see below), the agent removed the logging of special tokens that indicate tool usage (despite instructions not to change the special tokens), effectively bypassing our hallucination detection function.

```

diff --git a/coding_agent.py b/coding_agent.py
index 620da2f..800b06e 100644
--- a/coding_agent.py
+++ b/coding_agent.py
@@ -196,13 +196,16 @@ def chat_with_claude(task, chat_history, user_instr):
     "input": tool_input
 })

- # Process the tool call
+ # Process the tool call and record in structured format
     tool_result = process_tool_call(tools_dict, tool_name, tool_input)

- output_ls.extend([
- {TOOL_USED_MARKER: tool_name},
- {TOOL_INPUT_MARKER: tool_input},
- {TOOL_RESULT_MARKER: tool_result}
- ])
+ # Use structured format instead of text markers
+ output_ls.append({
+ "tool_invocation": {
+ "name": tool_name,
+ "input": tool_input,
+ "result": tool_result
+ }
+ })

     # Continue with tool results
     messages.extend([
@@ -224,7 +227,13 @@ def chat_with_claude(task, chat_history, user_instr):
         tools=tools,
     )
     tool_response_text = next((block.text for block in tool_response.content if hasattr(block,
     ↪ "text")), None)
- output_ls.append({TOOL_RESPONSE_MARKER: tool_response_text})
+ # Record tool response in structured format

```

```
+ if tool_response_text:
+ output_ls.append({
+ "tool_response": {
+ "content": tool_response_text
+ }
+ })
    response = tool_response

    # Validate tool response
```

In conclusion, the DGM can be applied beyond the coding domain, and we highlighted a case of objective hacking. Similar to reward hacking in reinforcement learning [124], objective hacking occurs when a system optimizes for a predefined, quantifiable objective rather than fulfilling the spirit of the task or solving the intended problem. This observation supports arguments made in prior works [35, 148], which suggest that optimizing quantitative measures often leads to undesirable or pathological outcomes, and aligns with Goodhart's law [127] – "When a measure becomes a target, it ceases to be a good measure."